Türkiye'de Eğitim Araştırmalarında Kayıp Veri Sorunu [1]

Ergül DEMİR[*]
Ankara Üniversitesi

Burcu PARLAK[**]
MEB

Özet

Bu çalışmanın amacı; Türkiye'de eğitim araştırmalarında kayıp veri sorunun ne düzeyde dikkate alındığının ve kayıp veri sorununa yönelik olarak hangi yöntemlerin daha yaygın bir şekilde kullanıldığının belirlenmesidir. Nitel bir araştırma olarak tasarlanan bu araştırmada, belgesel tarama yöntemi kullanılmıştır. Türkiye'de, araştırma kapsamında belirlenen ölçütleri karşılayan dört eğitim dergisinde, 2009-2011 yılları arasında yayımlanan toplam 708 makale, üç araştırmacı tarafından eş zamanlı olarak incelenmiştir. Ulaşılan bulgular, özellikle kayıp veri sürecinin raporlanmasında ciddi eksiklikler bulunduğunu göstermektedir. İncelenen araştırmaların 405'i, istatistiksel analiz süreçleri içermekle birlikte bu araştırmaların ancak 31'inde kayıp veri sorunu bulunduğu açıkça belirlenebilmiştir. Söz konusu 31 araştırmadan sadece 7 tanesinde kayıp veri yöntemi kullanıldığı, fakat bu kullanımın çok da bilinçli olarak yapılmadığı görülmüştür. Anlaşılan odur ki; Türkiye'de eğitim araştırmalarında kayıp veri sorununa nerdeyse tamamen ilgisiz kalınmaktadır.

Anahtar sözcükler: kayıp veri, liste yoluyla silme, en çok olabilirlik, çoklu veri atama

Abstract

In this study, it is aimed to determine how the educational researchers have dealt with the missing value problem and which missing value methods have been commonly used in these studies in Turkey. This study is designed as a qualitative research. Four educational journals and 708 articles published in these journals between 2009 and 2011 are considered for this aim. These journals meet the criteria determined in the context of the study. Articles are examined with documental analysis method by the three researchers synchronously. Findings are indicated that research reports have serious deficiencies especially missing data process. 405 researches include in the statistical analysis process. But only 31 researches have missing value problem clearly. There are only 7 researches from 31 using with some missing data methods. Consequently, it is cleared that missing data problem is generally ignored in educational researches in Turkey.

Keywords: missing data, listwise deletion, maximum likelihood, multiple imputation

İstatistiksel analizlerde kayıp veri sorunu, önemli bir tartışma alanıdır. Bu tartışmalar, esas itibariyle standart istatistiksel yöntemlerin doğasından kaynaklanmaktadır. Standart istatistiksel yöntemler, dikdörtgensel veri setlerinin analizine yönelik olarak geliştirilmiştir. Bir matris şeklinde hazırlanan bu veri setlerinde satırlar gözlemleri, sütunlar ise değişkenleri temsil etmektedir. Bir değişkene yönelik bir gözlemin bulunmaması durumunda, bu gözlemi temsil eden hücre boş kalır ve kayıp veri oluşur. Sonuçta tipik bir veri setinin kayıp veri içermesi, veri setinde yer alan bazı

[1] Bu makalenin hazırlanmasında, sayın Prof. Dr. Ezel TAVŞANCIL tarafından verilen Nitel Araştırma ve Veri Analizi doktora dersi kapsamında yapılan çalışmalardan yararlanılmıştır.

Bu makale, III. Ulusal Eğitimde ve Psikolojide Ölçme ve Değerlendirme Kongresi (18-19 Eylül 2012)'nde sözlü bildiri olarak sunulmuştur.

[*] Öğretim Görevlisi, Ankara Üniversitesi Ölçme ve Değerlendirme ABD, erguldemir@gmail.com
[**] Öğretmen, MEB Yenilik ve Eğitim Teknolojileri Genel Müdürlüğü, burcuprlk@yahoo.com

değişkenlere ilişkin bilgilerin ya da gözlemlerin bulunmaması anlamına gelmektedir (Little ve Rubin, 1987).

Çoğu yirminci yüzyılın başlarında geliştirilmiş olan standart istatistiksel yöntemler, eksiksiz veri setleri dikkate alınarak yapılandırılmıştır. Tam bilgi gerektiren bu yöntemler, genellikle kayıp verilere yönelik herhangi bir çözüm mekanizması içermemektedir. Dolayısıyla ilk önemli çalışmaların yapılmaya başlandığı 1970'lerin sonlarına kadar kayıp verilerin, istatistiksel analizlerde bir sorun olarak görülmediği söylenebilir. Afifi, Elashof, Hartley, Hocking, Orchard, Woodbury, Rubin, Dempster, Laird, ve Heckman gibi araştırmacılar tarafından 1970'lerde yürütülen çalışmalar, kayıp veriler üzerine ilk önemli örnekler olarak gösterilmektedir. Bununla birlikte asıl kırılmanın, Little ve Rubin tarafından hazırlanan 'Analysis with Missing Data' ve yine Rubin tarafından hazırlanan 'Multiple Imputation for Nonresponse in Surveys' adlı kitapların 1987 yılında yayımlanmasıyla yaşandığı belirtilmektedir. Bu önemli çalışmalar, istatistiksel analizlerde kayıp verilerin yol açabileceği sorunlara dikkat çekmenin yanı sıra bu sorunlarla başa çıkabilmeye yönelik temel yaklaşım ve yöntemleri de içermektedir (Little ve Rubin, 1987; Allison, 2002; Graham, 2009).

İstatistiksel analizlerde kayıp veriler, önemli sorunlara yol açabilmektedir. Öncelikle ve üzerinde sıklıkla durulan bir sorun olarak kayıp veriler, istatistiksel kestirimlerde olası bir yanlılık kaynağıdır. Örneğin yanıtlayanlar ile yanıtlamayanlar arasındaki, sıklıkla sistematik olan farklılıklar, olası bir yanlılık kaynağıdır. Bu tür bir örnekte yanıtlayanlardan oluşan örneklem, 'seçkisizlik (randomization)' özelliğini kaybetmiştir. Örneklemin evreni temsil edebilirlik düzeyi düşüktür. Araştırma sonuçları, ancak yanıtlayanlarla sınırlıdır. Verilen kararların yanlılık içermesi olasıdır. Dahası, yanıtlamayanların ve yanıtlamama gerekçelerinin genellikle bilinmiyor olması, bu tür bir yanlılığın giderilmesini zorlaştırmaktadır. İkinci bir sorun, kayıp verilerin, bilgi eksikliğine ve buna bağlı olarak istatistiksel analizin gücünün azalmasına yol açmasıdır. Bazı değişkenlerin analiz dışı bırakılması, t testi gibi istatistiksel testlerde serbestlik derecesinin hata miktarını artırmaktadır. Bu artış istatistiksel gücün azalmasına ve standart hatanın artmasına yol açmaktadır. Üçüncü bir sorun, yaygın olarak kullanılan standart istatistiksel yöntemlerin, kayıp veriler içeren veri setlerinde kullanımının zor olmasıdır. Örneğin kayıp veriler, faktör analizinde dengesizliğe yol açmaktadır. Özel istatistik yazılımlarının kullanımını gerektiren çok değişkenli istatistiksel yöntemler de ancak eksiksiz veri setlerinde uygulanabilir durumdadır. Diğer bir sorun, değerlendirilebilir kaynakların, kayıp veriler nedeniyle boşa harcanmasıdır. Boylamsal çalışmalar, geniş ölçekli testler, tamsayım ve benzeri çalışmalarda verilerin toplanması, önemli bir zaman ve maliyet içermektedir. Bu tür araştırmalarda araştırmacının yüksek yanıtlanma oranları ve yanıtlayıcılara yönelik tam bir profil elde edebilmesi, ekstra çaba, zaman ve maliyet anlamına gelmektedir (Rubin, 1987; Allison, 2002; Peng, Harwell, Liou ve Ehman, 2007).

Kayıp Verilerin İhmal Edilebilirliği ve Kayıp Veri Mekanizması

Kayıp verilerin yol açabileceği sorunlar, araştırmacıları eksiksiz veri setleri üzerinde çalışmaya yönlendirmektedir. Diğer taraftan eksiksiz veri seti elde etmenin zorluğu da açıktır. Dahası bazı araştırmalarda, araştırmanın kapsamı gereği, eksiksiz veri seti elde edilebilmesi mümkün olmamaktadır. Bu durumda araştırmacılar genellikle, kayıp verilerin 'ihmal edilebilir (ignorable)' olduğuna yönelik bir kanıt sağlama ve verileri analiz dışı bırakma eğilimi göstermektedir. Gerçekten de kayıp verilerin analiz dışı bırakılabilmesi, öncelikle kayıp verilerin ihmal edilebilir olmasını gerektirmektedir. İhmal edilebilirlik, kayıp değerlerle, seçkisiz olarak belirlenen gözlenen verilerin örtüşeceği varsayımını destekler. Bu durumda kestirim sürecinin bir parçası olarak kayıp veri mekanizmasının modellenmesine gerek olmadığı söylenebilir (Rubin, 1976; Allison, 2002).

Kayıp verilerin ihmal edilebilirliği, kayıp verinin niteliği, dolayısıyla kayıp verinin oluşmasına yol açan süreçlerle yakından ilgilidir. Kayıp veri mekanizması, uygun analiz yöntemlerinin belirlenmesinde ve sonuçların yorumlanmasında anahtar bir rol oynamaktadır. Bazı durumlarda kayıp veri mekanizması, istatistikçinin kontrolü altındadır. Örneğin örneklem üzerinde yürütülen bir araştırmada, kayıp veriye yol açan mekanizma basitçe örnekleme sürecinden kaynaklanıyor olabilir. Örnekleme sürecinin olasılıklı örneklemeye dayalı olarak yapılmış olması durumunda, kayıp veri mekanizmasının kontrol altında ve ihmal edilebilir olduğu söylenebilir. Örnekleme çatısı eksikse ya da bazı birimler yanıt vermemişse, gözlenen verilerin oluşmasına yol açan mekanizma iyi bir şekilde

anlaşılamaz. Bu durumda veri analizi, kayıp veri mekanizmasına yönelik temel varsayımların sağlanmasına bağlı olarak yürütülebilir (Little ve Rubin, 1987).

Rubin (1976, sy. 582), kayıp verilerin oluşmasına yol açan süreçlere yönelik üç durum tanımlamıştır. Bu tanımlamalarda iki parametreye atıf yapılmaktadır. θ, veriye yönelik parametreyi, Φ ise kayıp veri sürecine yönelik parametreyi temsil etmektedir. Örneğin Φ, veri setinde yer alan kayıp veri göstergesinin koşullu dağılımına yönelik bir parametre olabilir. Buna göre;

1. Kayıp veri, 'tesadüfî kayıp (missing at random-MAR)' olabilir. Bu durumda Φ parametresinin olası her bir değeri için, kayıp ve gözlenen veriler verildiğinde kayıp verilerin gözlenen örüntüsünün koşullu olasılığı, kayıp verilerin olası tüm değerleri için aynıdır.
2. Gözlenen veri, 'tesadüfî gözlenen (observed at random-OAR)' olabilir. Bu durumda kayıp verilerin ve Φ parametresinin olası her bir değeri için, kayıp ve gözlenen veriler verildiğinde kayıp verilerin gözlenen örüntüsünün koşullu olasılığı, gözlenen verilerin olası tüm değerleri için aynıdır.
3. Φ parametresi, θ parametresinden farklı olabilir. Bu durumda Φ ve θ arasında, parametre uzayının sınırlılıkları ya da olasılık dağılımları aracılığıyla oluşan hiçbir bağ yoktur.

Allison (2002), Rubin'in oldukça teknik sınıflamasını, daha anlaşılabilir şekilde açıklamaktadır. Buna göre kayıp veri mekanizması temelde iki varsayıma dayalıdır: (i) *Tamamıyla tesadüfî kayıp (missing completely at radom-MCAR)* ve (ii) *tesadüfî kayıp (missing at random-MAR)*. MCAR varsayımı, bir değişkendeki kayıp verilerin olasılığının, bu değişkenin kendi değeriyle ya da veri setindeki diğer herhangi bir değişkenin değeriyle ilişkili olmadığını ifade etmektedir. Bu varsayımın veri setindeki tüm değişkenler için anlamlı olması durumunda, kayıp veriler dışında kalan veri seti, orijinal gözlemler kümesinin basit seçkisiz örneklemi olarak kabul edilebilir. MAR, MCAR'a göre daha zayıf bir varsayımdır. MAR varsayımı, bir değişkendeki kayıp verilerin olasılığının, analizdeki diğer değişkenler kontrol altına alındığında, bu değişkenin kendi değeri ile ilişkisiz olduğunu ifade etmektedir. X, daima gözlenen değişken ve Y, kayıp veri içeren değişken olmak üzere MAR varsayımı, daha formel bir gösterimle ifade edilebilir:

$$P\,(Y_{kayip}/Y, X) = P\,(Y_{kayip}/X)$$

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

Kayıp veri mekanizmasının 'ihmal edilebilir' olup olmadığı kararı MCAR ve MAR varsayımlarına bağlı olarak verilebilmektedir. MCAR varsayımının karşılanması durumunda ya da MAR varsayımının karşılanması ve kayıp veri sürecine yönelik parametrelerin kestirilen parametrelerle ilişkisiz olması durumunda, kayıp veri mekanizması ihmal edilebilirdir. MAR varsayımının karşılanmaması durumunda ise kayıp veri mekanizması ihmal edilebilir değildir. Kayıp veri mekanizmasının ihmal edilebilir olması, kayıp verilerin analiz dışı bırakılabilmesi ile araştırmacının işini oldukça kolaylaştıracaktır. Fakat aksi durumda kayıp veri mekanizmasının modellenmesi gerekir. İhmal edilemez veriler içeren veri seti, genellikle hangi modellerin daha uygun olacağı konusunda yeterli bilgi sağlayamaz. Araştırma sonuçlarının seçilen modele oldukça duyarlı olacağı da açıktır. Dolayısıyla ihmal edilemez kayıp verilerle etkili bir kestirim yapılabilmesi için, kayıp veri sürecinin doğasına yönelik oldukça iyi ön bilgilere ihtiyaç vardır (Allison, 2002).

Geleneksel Kayıp Veri Yöntemleri

İstatistik yazılımlarında yaygın bir şekilde kullanılan geleneksel kayıp veri yöntemleri, genel olarak kayıp verilerin analiz dışı bırakılması ya da kayıp veriler yerine basit veri atama yöntemlerini içermektedir. Bu yöntemlerin tamamına yakını ancak MCAR varsayımı altında kullanılabilirdir. Diğer taraftan bu yöntemlerin özellikle kayıp verilerin sınırlı miktarda olduğu durumlar dışında kullanılmaması önerilmektedir (Little ve Rubin, 1987; Allison, 2009).

Kayıp veriler için kullanılan en bilindik çözüm, herhangi bir değişkene yönelik kayıp verinin analizden çıkarılmasıdır. Böylece kayıp veri içermeyen, eksiksiz bir veri seti elde edilir ve artık

bilindik istatistiksel analizlerin herhangi biri kolaylıkla uygulanabilir. Kayıp verilere yönelik bu yaygın yaklaşım *dizin yoluyla silme (listwise deletion-LD)*, *hücre yoluyla silme (casewise deletion-CD)* ya da *tam hücre analizi (complete case analysis-CCA)* olarak tanımlanmaktadır. Kullanımı oldukça kolay olan LD, ilk bakışta pratik ve basit bir çözüm gibi durmaktadır. Bununla birlikte LD, temelde bazı verilerin ihmal edilmesine yönelik bir yaklaşımdır ve veri kaybı söz konusudur. Veri kaybı, standart hatanın artmasına, güven aralığının genişlemesine ve hipotez testinin gücünün azalmasına yol açmaktadır. Dezavantajlarına rağmen LD'nin MCAR ve MAR varsayımlarına karşı oldukça kararlı ve sağlam olduğu ifade edilmektedir. Bu durum özellikle regresyon analizlerinde önemli bir üstünlük sağlamaktadır. LD yönteminin bir diğer olumlu özelliği, gerçek standart hata kestirimindeki titizliğidir. Bu ve benzeri olumlu yanları nedeniyle LD, geleneksel kayıp veri yöntemleri arasında 'dürüst' bir yöntem olarak değerlendirilmektedir (Little ve Rubin, 1987; Allison, 2002, 2009).

LD, özel bir değişkenin değerlerinin, diğer bir değişendeki kayıp veriye bağlı olarak analiz dışı bırakılmasından dolayı, ortalamaların ve marjinal dağılım fonksiyonlarının kestirimini gerektiren tek değişkenli analizlerde kullanışlı görülmemektedir. Bu durumda *eşleştirme yoluyla silme (pairwise deletion-PD)* yöntemi, alternatif bir yöntem olarak ortaya çıkmaktadır. *Erişilebilir hücre analizi (avaliable case analysis-ACA)* olarak da bilinen PD, özellikle doğrusal regresyon analizi, faktör analizi ve birçok yapısal eşitlik modelinde kullanılabilmektedir. PD yöntemi, kestirim sürecinde, erişilebilir olan tüm verilerin kullanımına dayanmaktadır. Kestirim sürecinde daha fazla veriyi dikkate almasına bağlı olarak PD'nin, LD'ye göre daha iyi bir yöntem olduğu söylenebilir. Bununla birlikte PD'nin en önemli sınırlılığı, örneklem büyüklüğüne fazlasıyla duyarlı olmasıdır. Kayıp veri örüntüsüne bağlı olarak erişilebilir örneklem büyüklüğünün yeterli düzeyde olmaması, kovaryans matrisinin pozitif tanımlı olmasını engellemekte, bu durumda kovaryans ve korelasyon matrisleri hesaplanamayabilmektedir (Little ve Rubin, 1987; Allison, 2002).

Geleneksel kayıp veri yöntemlerinden bir diğeri *dummy değişken düzeltmesi (dummy variable adjustment-DVA)* ya da *kayıp gösterge yöntemi (missing indicator method-MIM)* olarak bilinmektedir. Cohen ve Cohen'in 1985 yılındaki çalışmalarına ithaf edilen DVA basitçe, kayıp veri içeren yordayıcı değişkenlerin tamamında, verinin kayıp veri olup olmamasına bağlı olarak bir dummy değişkeni oluşturulması ve bu değişkenin de bir yordayıcı olarak analize dâhil edilmesi fikrine dayanmaktadır. Buna göre bir X değişkenine yönelik D dummy değişkeninde, X değişkenindeki kayıp veriler 1, diğer veriler 0 olarak kodlanır. Ayrıca bir X* değişkeni tanımlanır;

$$X^* = \begin{cases} X & \text{veri kayıp veri değil ise} \\ c & \text{veri kayıp veri ise} \end{cases}$$

X* değişkeninde c, genellikle erişilen veri değerlerinin ortalaması kullanılmakla birlikte, herhangi bir katsayı olabilir. c katsayısı sadece D değişkeni ile ilişkilidir. X*, herhangi bir c katsayısı için değişmez kalır. Dolayısıyla c katsayısının seçiminde herhangi bir sınırlılık söz konusu değildir. Yapılan tanımlamalar ile bir Y bağımlı değişkeni üzerinde X* ve D değişkenlerini içeren model üzerinde regresyon analizi süreci yürütülür. Benzer şekilde tanımlanacak diğer bağımsız değişkenlerin de modele dâhil edilmesi mümkündür. DVA yönteminin en önemli üstünlüğü, kayıp verilere yönelik erişilebilir tüm bilgileri kullanmasıdır. Bununla birlikte DVA'nın, MCAR varsayımı karşılansa bile regresyon katsayılarının kestiriminde yanlılığa yol açtığı da ifade edilmektedir (Allison, 2002, 2009).

Kayıp verilerin silinmesi ya da ihmal edilmesi yaklaşımının alternatifi, kayıp veri atamasıdır. Geleneksel kayıp veri yöntemleri içerisinde b*asit atama (simple imputation-SI),* her bir kayıp veriye kabul edilebilir bir tahmini değerin atanması ve veri setinde kayıp veri yokmuş gibi analizlere devam edilmesine dayalı bir takım yöntemleri ifade etmektedir. En basit ve bilindik veri atama yöntemi, *ortalama atama (mean substitution-MS)* yöntemidir. Bu yöntemde kayıp veri içeren bir değişkende, kayıp veriler yerine mevcut verilerin ortalaması atanmaktadır. Ancak bu yöntemin varyans ve kovaryans kestiriminde yanlılık ürettiği, artık iyi bilinen bir durumdur. Kayıp veriler yerine veri atanmasında çoklu regresyon kullanımının daha iyi bir yöntem olduğu belirtilmektedir. *Koşullu ortalama atama (conditional mean imputation-CMI)* ya da *Buck yöntemi* olarak bilinen bu yöntemde, regresyon modelinde yer alan bağımsız değişkenlerden biri kayıp veriler içermektedir. Bu değişken, diğer bağımsız değişkenlerle birlikte regresyon analizine tabi tutulur. Kestirilen denklem kullanılarak

kayıp veriler yerine atama yapılır. Bu yöntem, kayıp veri içeren bağımsız değişken sayısı arttıkça daha zorlaşmaktadır. Veri atama sadece bağımsız değişkenlere bağlı olmak üzere, MCAR varsayımı karşılanıyorsa, en küçük kareler katsayıları tutarlı ve kararlı olmakta ve özellikle geniş örneklemlerde yansız kestirimler elde edilebilmektedir. Diğer taraftan neredeyse tüm veri atama yöntemlerinin temel bir problemi olarak, bazı parametrelerin kestiriminde yanlılık oluşabilmektedir. Veri atamasından sonra kayıp veri yokmuş gibi analizlere devam edilmesi, standart hatanın daha düşük, test istatistiklerinin ise daha yüksek kestirimine yol açmaktadır. Basit veri atama yöntemleri, veri atama sürecinde kayıp verilere yönelik belirsizlik olması durumunda, herhangi bir düzeltme sağlayamamaktadır. Bu nedenle geleneksel veri atama yöntemlerinin 'dürüst olmayan' yöntemler olduğu belirtilmektedir (Little ve Rubin, 1987; Allison, 2002, 2009).

Kayıp Veri Sorununa Yeni Yaklaşımlar

Geleneksel yöntemlerin eksiklikleri ve dezavantajları, kayıp verilere yönelik yeni yaklaşımlar üzerinde çalışılması ihtiyacını beraberinde getirmiştir. Alternatif olarak geliştirilen pek çok yöntem arasında, 1990 başlarında teorik alt yapısı şekillendirilen ve 1990 sonlarında uygulama boyutuyla olgunlaştırılan *en çok olabilirlik (maximum likelihood-ML)* ve *çoklu veri atama (multiple imputation-MI)* yaklaşımlarının öne çıktığı ve giderek daha yaygın bir şekilde kullanıldığı ifade edilmektedir (Allison, 2002).

ML, bir yöntem olmaktan çok parametre kestiriminde olasılık temelli bir yaklaşımdır. ML yaklaşımının temel ilkesi, kestirim için gözlemlerin olasılığını maksimum yapacak verilerin seçilmesidir. ML, kayıp verilere yönelik oldukça iyi bir yöntem olarak görülmektedir. ML yönteminin özellikle MAR varsayımının sağlandığı fakat MCAR varsayımının sağlanmadığı durumlarda daha iyi kestirimler ürettiği belirtilmektedir. ML yaklaşımının kayıp verilerde uygulanabilmesi için ilgili tüm değişkenlerin 'ortak dağılımına (joint distribution)' ya da 'marginal dağılımına (marginal distribution)' yönelik bir modele ve olabilirliği maksimize edecek bir sayısal yönteme ihtiyaç duyulmaktadır. Örneğin tüm değişkenlerin kategorik olması durumunda, 'log-linear' ya da 'multinominal' modeller kullanılabilir. Diğer taraftan tüm değişkenlerin sürekli olması durumunda, tipik olarak 'çok değişkenli normallik' varsayımı altında bir modelin kullanılması gerekir. Bu tür bir modelde, ML yaklaşımı temelinde geliştirilmiş *beklenti-maksimizasyon algoritması (expectation-maximization algorithm-EM)* ya da *doğrudan maksimum olabilirlik (direct maximum likelihood-DMI)* gibi yöntemlerin kullanılması mümkündür (Allison, 2002, 2009).

EM, bazı verilerin kayıp olması durumunda ML kestirimleri elde edilmesinde kullanılan, iki aşamalı ve 'ötelemeli (iterative)' bir yöntemdir. 'Beklenti (expectation-E)' aşamasında, beklentinin hesaplanmasına yönelik parametre kestirimlerinin mevcut değerleri kullanılarak, kayıp veri içeren değişkenlerden, beklenen logaritmik olabilirlik değerleri elde edilir. 'Maksimizasyon (maximization-M)' aşamasında ise bir önceki aşamada elde edilen logaritmik olabilirlik değerleri, parametre kestirimlerine yönelik yeni değerler elde edilecek şekilde maksimize edilir. EM yönteminde bu iki aşama ötelemeli bir süreç içerisinde benzer kestirimler elde edilene kadar defalarca tekrar edilir. EM algoritmalarının temel çıktısı, ortalama, varyans ve kovaryansların ML kestirimleridir. Kestirilen bu parametre değerleri, aynı zamanda kestirim sürecinin bir parçası olarak kullanılmaktadır. Diğer parametrelerin kestirim sürecine dâhil edilmesi durumunda, yanlı kestirimler elde edileceği belirtilmektedir. EM algoritmalarının en önemli dezavantajı ise standart hata kestirimi üretmemesidir (Dempster, Laird ve Rubin, 1977; Allison, 2002, 2009).

EM yönteminin bir alternatifi olarak *doğrudan en çok olabilirlik (direct maximum likelihood-DMI)* yöntemi, standart hata kestirimine imkân sağlamaktadır. DMI yönteminde temel girdi, kovaryans matrisi yerine ham verilerdir. Bu nedenle bu yöntem *ham en çok olabilirlik (raw ML-RML)* ya da *tam bilgi en çok olabilirlik (full information ML-FIML)* yöntemi olarak da bilinmektedir. DMI yönteminde, ilgilenilen doğrusal model özelleştirilmiştir. Bu modelde ortalamalar ve kovaryans matrisi, modeldeki parametrelerin bir fonksiyonu olarak ifade edilir. DMI yönteminde olabilirlik fonksiyonu, modeldeki parametrelere göre doğrudan maksimize edilir. Standart hataların, bilgi matrisinin negatif tersinin hesaplanması gibi geleneksel ML yöntemleri ile belirlenmesi mümkündür. Bununla birlikte çok değişkenli normal dağılım altındaki bir modelin gerektirdiği temel varsayımlar DMI yönteminde de aynen geçerlidir (Allison, 2009).

ML yaklaşımının alternatifi olarak *çoklu veri atama (multiple ımputation-MI),* her bir kayıp veri yerine, olasılıkların dağılımını yansıtan, kabul edilebilir iki ya da daha fazla verinin atanmasını öngören bir yaklaşımdır (Rubin, 1987). MI yaklaşımı temelinde geliştirilmiş 'tanımlanmış atama (deterministic imputation)', 'seçkisiz atama (random imputation)' yöntemleri ya da *Markov Chain Monte Carlo (MCMC)* yöntemi gibi daha özel yöntemler bulunmaktadır (Allison, 2009). MI yöntemlerinde kayıp veri yerine m sayıda veri ataması yapılmakta ve m sayıda tamamlanmış veri seti elde edilmektedir. Her bir veri seti, atanan veriler gerçek verilermiş gibi kabul edilerek, standart eksiksiz veri süreçlerine göre analiz edilmektedir. MI yöntemlerinde 2 ile 10 arası tamamlanmış veri setinin kullanımının mümkün olduğu belirtilmektedir (Rubin, 1987). ML yöntemlerinde olduğu gibi MI yöntemlerinin kayıp verilerde kullanılabilmesi için MAR varsayımının karşılanması ya da kayıp veri mekanizmasına yönelik doğru bir modelin kurulması gereklidir (Allison, 2009). Karmaşık yapıların karmaşık modeller gerektireceği de açıktır. MI yöntemleri, ML yöntemlerine göre model seçiminde daha az duyarlıdır. Çünkü basitçe, MI yöntemlerinde model, sadece kayıp verilere atama yapılmasında kullanılmakta, ML yöntemlerinde olduğu gibi diğer parametrelerin kestiriminde kullanılmamaktadır (Allison, 2002). ML kestirimlerinde olduğu gibi MI yöntemleri için de en kullanışlı model, çok değişkenli normal modeldir. Bazı değişkenler normal dağılım varsayımını karşılamasa da çok değişkenli modelde MI yöntemlerinin oldukça başarılı olduğu ifade edilmektedir (Schaffer, 1997).

Yeni yaklaşımlar temelinde ortaya konulan kayıp veri yöntemleri, geleneksel yöntemlere göre oldukça avantajlıdır. Bununla birlikte varsayımların sağlanmasındaki zorluk ve karmaşık hesaplama süreçleri bu yeni yöntemlerin kullanımında önemli sınırlılıklar oluşturmaktadır. Hâlihazırda ML yaklaşımı ancak doğrusal ve logaritmik doğrusal modellere uygulanabilir durumdadır. Örneğin lojistik regresyon, Poisson regresyon ya da Cox regresyon gibi çok değişkenli modellerde ML yaklaşımının uygulanabilirliğine yönelik teorik bir alt yapı henüz kurulamamıştır. ML yöntemleri ile karşılaştırıldığında MI yöntemlerinin iki önemli üstünlüğü bulunmaktadır. İlk olarak MI yöntemleri her çeşit veri ve modele uygulanabilmektedir. İkinci olarak MI yöntemlerinde, LES ve AMOS gibi özel yazılımlar yerine geleneksel istatistik yazılımlarının kullanılabilmektedir. Fakat gerek ML gerekse MI yöntemlerinde, istatistik yazılımlarının kullanımındaki en önemli sorun, veri atamasının genellikle seçkisiz olarak yapılmasından dolayı, her bir yazılımın farklı sonuçlar üretmesidir (Allison, 2002, 2009)

Kayıp veri sorununa yönelik olarak geliştirilmiş pek çok çözüm yönteminden bazılarının, diğerlerine göre daha iyi olduğu söylenebilir. Fakat bu yöntemlerden hiç birisi gerçek anlamda 'iyi' olarak tanımlanamaz. Gerçek çözüm, kayıp veri olmaması ya da kayıp veri miktarının ihmal edilebilir düzeyde olmasıdır. Bu nedenle bir araştırmada kayıp veri miktarının en alt düzeye indirgenmesi çabası önemlidir. Özensiz ve dikkatsiz bir şekilde yürütülen bir araştırmada herhangi bir istatistiksel düzeltmenin bir anlamı da yoktur (Allison, 2002).

Literatürde kayıp veri sorununa yönelik olarak geliştirilmiş yöntemlerin karşılaştırıldığı birçok araştırmaya rastlamak mümkündür (örneğin Rubin, 1976; Dempster ve diğerleri, 1977; Brown, 1983; Acock, 2005; Enders, 2006; Graham, 2009; Çokluk ve Kayrı, 2011). Bu araştırmaların büyük çoğunluğu, oldukça teknik çalışmalardır. Genellikle kayıp veri yöntemleri, özet olarak tartışılmakta ve örnek bir uygulama üzerinden karşılaştırılmaktadır. Örnek uygulamaların genellikle türetilmiş veriler üzerinden yapıldığı görülmektedir. Beklenen bir sonuç olarak ML ve MI yaklaşımları temelinde geliştirilen yöntemlerin, kayıp veri sorunu ile başa çıkmada daha işlevsel ve uygun olduğu belirtilmektedir. Kayıp veri yöntemlerinin ne düzeyde kullanıldığının sorgulandığı araştırmaların ise çok daha seyrek olduğu söylenebilir (örneğin Peng ve diğerleri, 2007).

Kayıp veri sorununa yönelik yaklaşımların ancak 1970 sonrasında tartışılmaya başlanması ve kayıp veri yöntemlerinin ancak 1990 sonrasında istatistiksel yazılımlara yansıtılması, bu yöntemlerin kullanımının henüz istenen düzeyde yaygınlaşmamış olmasını bir ölçüde açıklamaktadır. Bununla birlikte elde edilen sonuçların yansızlığı ve kestirimlerin doğruluğu açısından, özellikle eğitim araştırmalarında kayıp veri sorununun ne düzeyde dikkate alındığı, hangi yaklaşım ve yöntemlerin kullanıldığı önemli bir sorgulama alanıdır. Bir araştırmada kayıp verilerin dikkate alınması ve kestirim sürecinde doğru kayıp veri yöntemlerinin kullanılması, araştırmacının önemli sorumluluklarından biri

olmanın yanı sıra araştırmanın titiz bir şekilde yürütüldüğünün de bir göstergesidir. Bu kapsamda özellikle istatistiksel analiz süreçleri içeren araştırmalarda, kayıp veri mekanizmasını ortaya koyan temel koşulların belirlenmiş olması, bu koşullara bağlı olarak uygun kayıp veri yönteminin seçilmiş olması, seçilen yönteme bağlı olarak veri seti üzerinde gerekli düzeltmelerin yapılmış olması ve ileri analizlerin bu düzeltmelerden sonra sürdürülmüş olması beklenir.

Amaç

Bu çalışmada Türkiye'de, istatistiksel analiz süreçleri içeren eğitim araştırmalarında, kayıp verilerin ne düzeyde dikkate alındığının betimlenmesi amaçlanmaktadır. Bu amaçla, belirlenen eğitim araştırmaları makale örneklemi, betimsel bir yöntemle incelenmiştir. Aşağıdaki sorulara yanıt aranmıştır:

1. Türkiye'de eğitim araştırmalarında kayıp veri sorunu ne düzeyde dikkate alınmaktadır?
2. Türkiye'de eğitim araştırmalarında kayıp veri sorununa yönelik olarak en yaygın şekilde kullanılan yöntemler hangileridir?

Belirlenen amaç doğrultusunda her bir makaleye yönelik olarak aşağıdaki sorulara yanıt aranmıştır:

1. Araştırma, istatistiksel analiz süreçleri içermekte midir? İçeriyor ise;
   a. Hangi istatistiklerin kestirimi yapılmıştır?
   b. Hangi istatistiksel analizler kullanılmıştır?
   c. Hangi istatistik yazılımları kullanılmıştır?
2. Analizlerde kullanılan veri seti kayıp veri içermekte midir? İçeriyor ise;
   a. Kayıp veri mekanizması incelenmiş midir?
   b. Kayıp veri mekanizmasına yönelik hangi temel varsayımlar dikkate alınmıştır?
3. Uygun bir kayıp veri yöntemi kullanılmış mıdır? Kullanılmış ise hangi yöntem tercih edilmiştir?

Önem

Kayıp veri yöntemlerine yönelik araştırmaların literatürde belli bir birikime ulaştığı görülmektedir. Bununla birlikte eğitim araştırmalarında kayıp veri sorunu, üzerinde daha az çalışılmış bir alandır. Türkiye özelinde ise konu ancak birkaç araştırmacının ilgisini çekmiş görünmektedir. Bu bağlamda bu araştırmanın özellikle Türkiye'de ilgili literatüre katkı sağlayacağı ve kayıp veri sorununa dikkat çekeceği düşünülmektedir.

Açıktır ki analiz süreçlerinde kayıp verilerin dikkate alındığı araştırmalarda, doğru sonuçlara ulaşabilme olasılığı daha yüksektir. Bu bağlamda bu araştırmanın, Türkiye'de eğitim araştırmalarının ne düzeyde doğru sonuçlar üretebildiğine yönelik bir fikir verebileceği öngörülmektedir. Diğer taraftan kayıp veri yöntemlerinin uygun ve etkin bir şekilde kullanılabilmesi, öncelikle araştırmacıların bilgi ve yeterlik düzeylerine bağlıdır. Dolayısıyla bu araştırmanın, Türkiye'de eğitim alanında, araştırmacıların, kayıp veri yöntemlerine yönelik bilgi ve yeterlik düzeyleri hakkında bir fikir vereceği de düşünülmektedir.

Yöntem

Nitel bir araştırma olarak tasarlanan bu araştırmada, belgesel tarama yöntemi kullanılmıştır. Türkiye'de, araştırma kapsamında belirlenen ölçütleri karşılayan eğitim dergilerinde, 2009-2011 yılları arasında yayımlanan makaleler, belirlenen değişkenler çerçevesinde incelenmiş, elde edilen veriler betimsel analiz yöntemi ile analiz edilmiştir.

İncelenecek Makalelerin Belirlenmesi

Makalelerin belirlenmesinde, amaçlı örnekleme yöntemlerinden ölçüte dayalı örnekleme kullanılmıştır. Öncelikle incelenecek makalelerin yayımlandığı eğitim dergileri seçilmiştir. Bu dergilerin belirlenmesinde aşağıdaki ölçütler dikkate alınmıştır:

1. Hakemli dergi olması.

2. Uluslararası indekslerden en az birine kayıtlı olması.
3. Beş yıldan uzun bir süredir yayımlanıyor olması.
4. Yılda en az iki sayı olarak yayımlanması.
5. İnternet üzerinde açık erişim hizmeti veriyor olması.

Bu ölçütleri karşılayan dergilerin, eğitim alanında ciddi araştırmaların yayımlanmasına olanak sağladığı, erişilebilirliklerinin yüksek olduğu, kayıp veri yöntemleri hakkında yeterlik sahibi hakemler istihdam edebildiği ve henüz çok yaygın olarak kullanılmayan kayıp veri yöntemlerinin teşvik edilmesine yönelik daha uygun bir alt yapıya sahip olduğu varsayılmaktadır. Bu ölçütler doğrultusunda dört dergi belirlenmiştir[2]: (i) *Ankara Üniversitesi Eğitim Bilimleri Fakültesi Dergisi-AÜEBFD,* (ii) *Eğitim ve Bilim-EB,* (iii) *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi-HÜEFD* ve (iv) *The Turkish Online Journal of Educational Technology-TOJET.*

Belirlenen dergilerde 2009-2011 yılları arasında yayımlanan toplam makale sayısı 708'dir. Bu makaleler içerisinde, uygun bir kayıp veri yönteminin kullanılmasını gerektiren araştırmalara yönelik olanların belirlenmesinde, iki temel ölçüt dikkate alınmıştır:

1. İstatistiksel analiz süreçleri içerme.
2. Analizlerde kullanılan veri setinin kayıp veri içermesi.

Veri analizi aşaması istatistiksel analiz içermeyen araştırmalarda, herhangi bir kayıp veri yönteminin kullanımının gerekli olmadığı açıktır. İstatistiksel analiz süreçleri içermekle birlikte, analizlerde kullanılan veri setinin eksiksiz veri seti olması da herhangi bir kayıp veri yönteminin kullanımını gerektirmeyen bir durumdur.

Makalelerin İncelenmesi

Belirlenen dergilerde 2009-2011 yılları arasında yayımlanan makaleler, araştırmanın amacı ve araştırma soruları çerçevesinde, üç araştırmacı tarafından incelenmiştir. Araştırma soruları çerçevesinde yapılan incelemelerden elde edilen sonuçlar, bu amaçla hazırlanan bir form üzerinde kaydedilmiştir. İnceleme sonuçları arasında tutarsızlık olduğu durumlarda, araştırmacılar bir araya gelerek söz konusu makaleyi tekrar incelemiş ve uzlaşmayla ortak bir sonuca varmıştır.

İncelemelerden elde edilen veriler, betimsel analiz yöntemi ile analiz edilmiştir. Bu kapsamda analiz sürecinde, standart sınıflama ve sayma prosedürleri kullanılmış, sıklık dağılımları üzerinden betimlemeler yapılmıştır. Verilerin analizi, araştırmacılar tarafından eş zamanlı olarak yapılmıştır. Takip edilen süreç ve ulaşılan bulgular karşılaştırılmıştır. Her üç araştırmacının da benzer süreçleri takip ettiği ve benzer bulgulara ulaştığı görülmüştür.

Geçerlik ve Güvenirlik Çalışmaları

Nitel araştırma yaklaşımının benimsendiği bu araştırmada geçerlik ve güvenirlik, nitel araştırmaların doğasına uygun olarak 'inandırıcılık', 'aktarılabilirlik', 'tutarlılık' ve 'teyit edilebilirlik' özellikleri kapsamında değerlendirilmiştir. Araştırmanın bu özellikler açısından yeterliliğini sağlamaya yönelik bazı tedbirler alınmıştır.

Ayrıntılı bir raporlama yapılmasına özen gösterilmiştir. Özellikle araştırmanın dayandığı kavramsal çerçeve, araştırmada incelenen makalelerin seçim süreci ve bu süreçte dikkate alınan ölçütler, olabildiğince sade, anlaşılabilir ve ayrıntılı şekilde verilmeye çalışılmıştır.

Araştırmanın amacına uygun bir örneklemin belirlenmesinde olabildiğince titiz davranılmıştır. İncelemeye konu olan eğitim dergilerinin ve araştırma makalelerinin seçiminde, amaçlı örnekleme yöntemlerinden biri olan ölçüte dayalı örnekleme kullanılmıştır. Bu amaçla dergilerin belirlenmesinde ve makalelerin incelenmesinde dikkate alınan ölçütler, gerekçeleriyle birlikte ayrıntılı bir şekilde açıklanmıştır.

Araştırma örneklemini oluşturan makaleler, üç araştırmacı tarafından eşzamanlı olarak incelenmiştir. Bu incelemede önceden üzerinde uzlaşılmış bir inceleme formu kullanılmıştır. Ayrıca

[2]Bkz. http://dergiler.ankara.edu.tr; http://egitimvebilim.ted.org.tr; http://www.efdergi.hacettepe.edu.tr; http://www.tojet.net

elde edilen verilerin analizi de araştırmacılar tarafından yine eşzamanlı olarak yürütülmüştür. İnceleme ya da analiz sonuçlarında ortaya çıkan tutarsızlıklar, uzlaşmayla çözülmüştür.

Yapılan analizler sonucunda elde edilen bulguların, öncelikle betimsel düzeyde ve ham olarak verilmesine dikkat edilmiştir. Bulguların yorumlarla karışmamasına özen gösterilmiştir. Elde edilen bulguların araştırmanın amacı ve araştırma soruları ile uyumu da araştırmacılar tarafından dikkatli bir şekilde sorgulanmıştır.

Bulgular

Bu çalışmada AÜEBFD, EB, HÜEFD ve TOJET'de 2009-2011 yılları arasında yayımlanan toplam 708 makale incelenmiştir. Bu makalelerin yayımlandıkları dergilere ve yayım yılına göre dağılımları Tablo 1'de verilmektedir.

Tablo 1. İncelenen Makalelerin Dağılımı

Dergiler	Yayım Yılı			Toplam
	2009	2010	2011	
AÜEBFD	35	17	18	70
EB	53	58	95	206
HÜEFD	49	60	80	189
TOJET	32	81	130	243
Toplam	169	216	323	708

İncelenen 708 makaleden 110'u, Türkiye örneklemi üzerinde yürütülmüş olmayan araştırmalara yöneliktir. Bu araştırmaların bir kısmı, yabancı araştırmacılar tarafından yürütülmüş, derleme ve kuramsal düzeyde çalışmalar ya da nitel araştırmalardır. Diğer bir kısmı ise nicel araştırma olmakla birlikte, Türkiye dışında örneklemler üzerinde yürütülmüştür.

Türkiye evreninde yürütülmüş olduğu belirlenen 598 makaleden 193'ü de, benzer şekilde herhangi bir istatistiksel analiz süreci içermemektedir. Bu araştırmalar genellikle nitel araştırma yaklaşımına göre yürütülmüş ya da ilgilenilen konunun literatüre dayalı olarak özetlendiği derleme türü çalışmalardır. Nitel araştırma yaklaşımının doğası gereği bu araştırmalarda, en fazla yüzde ve frekans dağılımları dikkate alınmıştır. İstatistiksel analiz yöntemleri yerine genellikle standart sınıflama ve sayma prosedürleri kullanılmıştır. Nicel araştırma yaklaşımının benimsendiği, fakat verilere yönelik sadece yüzde ve frekans dağılımlarının dikkate alındığı birkaç araştırma ise istatistiksel analiz süreçleri içermeyen araştırmalar arasında değerlendirilmiştir.

İncelenen 708 araştırmadan 405'inin Türkiye evreninde yürütüldüğü ve istatistiksel analiz süreçleri içerdiği belirlenmiştir. Bu 405 araştırmanın yayımlandığı dergi ve yayım yılına göre dağılımları Tablo 2'de gösterilmektedir.

Tablo 2. Türkiye Örnekleminde Yürütülen ve İstatistiksel Analiz Süreci İçeren Araştırma Makalelerinin Dağılımı

Dergiler	Yayım Yılı			Toplam
	2009	2010	2011	
AÜEBFD	19	11	10	40
EB	35	35	66	136
HÜEFD	34	48	58	140
TOJET	15	28	46	89
Toplam	103	122	180	405

Türkiye evreninde yürütülen ve istatistiksel analiz süreci içeren 405 araştırmadan 213'ünde, analizlerin yürütüldüğü veri setinin herhangi bir kayıp veri içermediği, analizlerin eksiksiz veri seti üzerinde gerçekleştirilmiş olduğu belirlenmiştir. Eksiksiz veri seti üzerinde çalışıldığı genellikle raporlanmadığından bu tespit, analiz çıktıları ve özellikle bu çıktılara yönelik tabloların incelenmesine bağlı olarak yapılmıştır.

Türkiye evreninde yürütülen ve istatistiksel analiz süreci içeren 405 araştırmadan 161'inde analizlerin yürütüldüğü veri setinin kayıp veri içerip içermediğine yönelik herhangi bir kanıta rastlanmamıştır. Bu araştırmaların tamamında standart hata, ortalama, varyans, kovaryans, korelasyon, t ve F istatistiği, güvenirlik ve geçerlik katsayıları, faktör yükleri, regresyon denklemi katsayıları gibi istatistiklerin bir kaçının kestiriminin yapıldığı görülmüştür. Kestirimlerde sıklıkla ki-kare testi, t testi, varyans analizi ve faktör analizi gibi istatistiksel analiz yöntemleri kullanılmıştır. Daha seyrek olmakla birlikte madde analizi, regresyon analizi ve yapısal eşitlik modeli analizlerine de rastlanmıştır. Ayrıca söz konusu bu 161 araştırmanın 69'unda kullanılan istatistik yazılımına yönelik bilgi verilmediği görülmüştür. Geriye kalan 92 araştırmadan 77'sinde ise SPSS yazılımının, sadece birkaç araştırmada ise LISREL, STATA, AMOS, BILOG, Mplus ve EQS yazılımlarının kullanıldığı raporlanmıştır. Bu araştırmalarda kayıp veri mekanizmasının incelenmediği ve herhangi bir kayıp veri yönteminin kullanılmadığı da belirlenmiştir. Kestirilen istatistikler, kullanılan analizler ve istatistik yazılımları dikkate alındığında bu araştırmalarda kayıp veri sorunu bulunması olasıdır. Fakat raporlama eksikleri nedeniyle bu 161 araştırmada, kayıp veri sorununa yönelik bir inceleme yapılması mümkün olmamıştır.

Türkiye evreninde yürütülen ve istatistiksel analiz süreci içeren 405 araştırmadan ancak 31'inde, analizlerin yürütüldüğü veri setinin kayıp veri içerdiği bilgisine kesin olarak ulaşılmıştır. Bu araştırmaların ancak birkaç tanesinde, veri setinde kayıp veri bulunduğu raporlanmıştır. Diğerlerinde ise veri setinin kayıp veri içerdiği bulgusuna, analiz çıktılarının ve özellikle analiz çıktı tablolarının incelenmesiyle ulaşılabilmiştir. Bu 31 araştırmanın, yayımlandığı dergilere ve yayım yılına göre dağılımı, Tablo 3'te gösterilmektedir.

Tablo 3. Kayıp Veri Seti İçeren Araştırma Makalelerinin Dağılımı

Dergiler	Yayım Yılı			Toplam
	2009	2010	2011	
AÜEBFD	2	3	0	5
EB	0	1	6	7
HÜEFD	0	11	2	13
TOJET	1	3	2	6
Toplam	3	18	10	31

Analizlerin yürütüldüğü veri setinin kayıp veriler içerdiği kesin olarak belirlenen 31 araştırmanın tamamında standart hata, ortalama, varyans, kovaryans, korelasyon, t ve F istatistiği, güvenirlik ve geçerlik katsayıları, faktör yükleri, regresyon denklemi katsayıları gibi istatistiklerin bir kaçının kestiriminin yapıldığı görülmüştür. Bu kestirimlerde sıklıkla ki-kare testi, t testi, tek yönlü ya da çok değişkenli varyans analizi, açımlayıcı ve doğrulayıcı faktör analizleri gibi istatistiksel analiz yöntemleri kullanıldığı belirlenmiştir. Daha seyrek olmakla birlikte kümeleme analizi, regresyon analizi ve yapısal eşitlik modeli analizleri gibi daha üst düzey analiz yöntemleri de kullanılmıştır. Bununla birlikte söz konusu 31 araştırmanın 14'ünde, kullanılan istatistik yazılımı hakkında bilgi verilmemiştir. Geriye kalan araştırmaların 11'inde SPSS, 1'inde Microsoft Excell, 4'ünde SPSS ve LISREL, 1'inde ise SPSS ve EZDIF yazılımları kullanılmıştır.

Kayıp veri sorunu bulunduğu kesinleşen 31 araştırmadan sadece 4'ünde, kayıp veri örüntüsünün incelendiği görülmüştür. Fakat sadece 1 araştırmada, bu inceleme, temel kayıp veri varsayımları dikkate alınarak yapılmış ve kayıp veri setinin MCAR varsayımını karşıladığı belirlenmiştir. Diğer araştırmalarda ise kayıp veriler, ya kullanılan istatistiksel yazılımın bir özelliği olarak dikkate alınmış

ya da ihmal edilebilir görülerek analiz dışı bırakılmıştır. Dolayısıyla kayıp verilerin seçkisiz olarak oluşup oluşmadığı, değişkenlerdeki kayıp veriler arasında sistematik bir ilişki bulunup bulunmadığı, kayıp veri mekanizmasının ihmal edilebilir olup olmadığı, teknik olarak sorgulanmamıştır.

Kayıp veri sorunu bulunduğu kesinleşen 31 araştırmanın ancak 7'sinde kayıp veri sorununa yönelik bir yöntem kullanılmıştır. Bu araştırmaların 3'ünde LD, 1'inde MS, 1'inde EM, 1'inde MI ve 1'inde MCMC yöntemi kullanılmıştır. LD yönteminin kullanıldığı 2 ve MI yönteminin kullanıldığı 1 araştırmada, bu yönteminin kullanılma gerekçesine yönelik bir açıklama yapılmadığı gibi kayıp veri sorunundan da bahsedilmemektedir. Analiz sürecinde bu yöntemlerin kullanıldığı, analiz çıktılarının verildiği tabloların altındaki dipnottan anlaşılmıştır. Diğer 3 araştırmada ise kayıp veriler, varsayımlar dikkate alınmaksızın incelenmiş ve genellikle tercih edilen istatistik yazılımına bağlı olarak uygun bulunan LD, MS ve EM yöntemleri kullanılmıştır. Sadece 1 araştırmada bilinçli bir şekilde kayıp veri mekanizmasının incelendiği ve MCAR varsayımı altında uygun bir yöntem olarak MCMC yönteminin kullanıldığı görülmüştür.

Sonuç

İncelenen makalelerle temsil edilen eğitim araştırmaları içerisinde, nitel yaklaşımla yürütülen ya da derleme şeklinde hazırlanan araştırmalar bulunmakla birlikte, ağırlık oluşturan kısmın istatistiksel analiz süreçleri içeren araştırmalar olduğu görülmektedir. Bu durum eğitim araştırmalarında daha çok istatistiksel analizler gerektiren veri setleriyle çalışıldığını göstermek için yeterli değildir. Belki ancak bu çalışmada incelenen makalelerin yer aldığı dergilerde, bir takım istatistiksel kestirimler yapmayı gerektiren araştırmaların daha sık yayımlandığı söylenebilir.

İncelenen makaleler kapsamında, istatistiksel analiz süreçleri içeren araştırmalarda genellikle tam veri seti üzerinde çalışıldığı görülmektedir. Bir araştırmada eksiksiz bir veri seti elde etmek, maliyetli ve ekstra çaba gerektiren zor bir iştir. Bununla birlikte tam veri seti ile çalışmak, olası yanlılık ve hatalı kestirim risklerinin en aza indirilmesi açısından, araştırmacı için önemli avantajlar sağlamaktadır.

İncelenen makalelerle temsil edilen eğitim araştırmalarında, kullanılan veri setinin özelliklerinin genellikle iyi bir şekilde raporlanmadığı görülmektedir. Kullanılan veri setinin kayıp veri içerip içermediği, ancak analiz çıktıları ve bulgular incelenerek anlaşılabilmektedir. Bu tür bir incelemeye rağmen, kayıp veri bulunup bulunmadığına yönelik kanıt elde edilemeyen araştırmaların sayısı, oldukça fazladır.

İstatistiksel analiz süreçleri içerdiği belirlenen araştırmalarda sıklıkla, standart hata, ortalama, varyans, kovaryans, korelasyon, t ve F istatistikleri, geçerlik ve güvenirlik katsayıları, faktör yükleri, regresyon katsayıları gibi istatistiklerin birkaçının kestirimi yapıldığı görülmüştür. Kayıp verilere yönelik herhangi bir düzeltme yapılmaksızın gerçekleştirilen bu tür kestirimlerin, kayıp verilerin miktarı ve niteliğine bağlı olarak yanlılık ya da kestirim hataları içerebileceği açıktır.

İstatistiksel analiz süreçleri içerdiği belirlenen araştırmalarda genellikle, kullanılan istatistik yazılımı belirtilmemektedir. Kullanılan istatistik yazılımının belirtildiği az sayıda araştırmanın neredeyse tamamında ise SPSS yazılımının kullanıldığı görülmektedir. SPSS, 'missing data analysis' menüsü altında ancak bazı geleneksel kayıp veri yöntemlerinin kullanımına olanak sağlamaktadır. MI, FIML, MCMC gibi daha üst düzey kayıp veri yöntemleri için, bu yöntemlere yönelik özel yazılımların kullanılması gerekmektedir. Dolayısıyla araştırmacıların bu tür özel yazılımlara yönelik bilgi ve yeterlik düzeylerinin de yüksek olmadığı düşünülmektedir.

İncelenen araştırmalar içerisinde kayıp veri sorunu bulunduğu kesinleşen az sayıda araştırma tespit edilebilmiştir. Bu araştırmaların büyük çoğunluğunda ise kayıp veri mekanizmasının dikkate alınmadığı ve herhangi bir kayıp veri yönteminin kullanılmadığı görülmüştür. Birkaç tane de olsa kayıp veri mekanizmasının incelendiği ve bir kayıp veri yönteminin kullanıldığı araştırmaya rastlanmıştır. Bu araştırmalarda ise kullanılan kayıp veri yönteminin, genellikle tercih edilen istatistik yazılımının bir özelliği olduğu görülmektedir.

Gelinen noktada görülmektedir ki; kayıp veri yaklaşımları açısından Türkiye'de eğitim araştırmalarında ve araştırmaların raporlanmasında önemli eksiklikler bulunmaktadır. Kayıp veri sorununun, nerdeyse tamamen ilgisiz kalınan bir alan olduğu görülmektedir. Üst düzey istatistiksel analiz süreçlerinin kullanıldığı ve kayıp veri içerdiği açıkça belli olan araştırmalarda bile kayıp veri mekanizmasına yönelik yeterli bir inceleme yapılmamakta ve uygun bir kayıp veri yöntemi kullanılmamaktadır. Esas sorun olarak, metodoloji ve kayıp veri yöntemleri konusunda yeterlik düzeyinin beklenenin altında olması, kendisini açıkça göstermektedir.

Kaynaklar

Allison, P.D. (2009). Missing Data. Ed. Roger E. Millsao ve Alberto Maydeu-Olivares, *Quantitative Methods in Psychology,* sy.72-89. London: SAGE Publication.

Allison, P.D. (2002). *Missing Data.* California: Sage Publication, Inc.

Acock, A.A. (2005). Working with Missing Values. *Journal of Marriage and Family,* s.65, sy.1012-1028.

Brown, C.H. (1983). Asymptotic Comparision of Missing Data Procedures for Estimating Factor Loadings. *Biometrika,* s.48, n.2, sy.269-291.

Çokluk, Ö. ve Kayrı, M. (2011). The Effects of Methods of Imputation for Missing Values on the Validity and Reliability of Scales. *Educational Sciences: Theory&Practice,* s.11, sy.303-309.

Dempster, A.P., Laird, N.M. ve Rubin, D.B. (1977). Maximum Likelihood Estimation from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society,* seri B, s.39, sy.1-38.

Enders, C.K. (2006). A Primer on the Use of Modern Missing-Data Methods in Psychomatic Medicine Research. *Psychomatic Medicine,* s.68, sy.427-436.

Graham, J.W. (2009). Missing Data Analysis: Making It Work in the Real World. *Annual Review of Psychology,* s.60, 4, sy.549-576.

Little, R.J.A ve Rubin, D.B. (1987). *Statistical Analysis with Missing Data, 2nd ed.* New York: John Wiley & Sons, Inc.

Peng, C.Y.J., Harwell,M., Liou, S.M. ve Ehman L.H. (2007). Advances in Missing Data Methods and Implication for Educational Research. Ed. Sholomo S. Sawilowski, *Real Data Analysis,* sy.31-77. USA: IAP-Information Age Publishing.

Rubin, D.B. (1976). Inference and Missing Data. *Biometrika,* s.63, 3, sy.581-592.

Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys.* New York: John Wiley & Sons, Inc.

Scahffer, J.L. (1997). *Analysis of Incomplete Multivariate Data.* London: Chapman&Hall