İstatistikçiler Dergisi: İstatistik & Aktüerya

Journal of Statisticians: Statistics and Actuarial Sciences

IDIA 10, 2017, 2, 76-85

Geliş/Received:13.07.2017, Kabul/Accepted: 06.12.2017

www.istatistikciler.org

Araştırma Makalesi / Research Article

Rastlantısal Olmayan Kayıp Veri Varlığında Seçim Modelleri ile bir Duyarlılık Analizi Uygulaması

Oya Kalaycıoğlu

Abant İzzet Baysal Üniversitesi
Ekonometri Bölümü
14300, Bolu, Türkiye
oyakalaycioglu@ibu.edu.tr
0000-0003-2183-7080

Öz

Bu çalışmadaki amaç, bağımlı değişkende rastlantısal olmayan kayıp veriler olduğunda, nasıl bir istatistiksel modelleme stratejisi izlenebileceğini açıklamak ve regresyon parametrelerinin farklı kayıp veri mekanizmaları varsayımlarına ne kadar duyarlı olabileceğini göstermektir. Bu amaç doğrultusunda, bir hanehalkı araştırmasından elde edilen veriler kullanılmış ve eğitim seviyesinin gelir düzeyi üzerindeki etkisi, doğrusal regresyon modeli kullanılarak, farklı kayıp veri mekanizmaları varsayımları altında ölçülmüştür. Analiz için Bayesci tahmin yöntemleri kullanılarak seçim modelleri yardımı ile, regresyon modeli ve kayıp veri modeli bileşik olarak modellenmiştir. Kayıp veri modelinin parametreleri değiştirilerek duyarlılık analizi yapılmış ve farklı kayıp veri mekanizmaları altında tahmin edilen regresyon katsayılarında ciddi farklılıklar görülmüştür.

Anahtar sözcükler: Kayıp veri, rastlantısal olmayan kayıp veri, seçim modelleri, Bayesci analiz, duyarlılık analizi

Abstract

An Application of Sensitivity Analysis in the Presence of Non-random Missing Data Using Selection Models

The aim of this research is to explain a statistical modelling strategy in the presence of non-random missing data, and thereby to indicate how sensitive would the regression parameter estimates be under different assumptions of missing data mechanisms. For this purpose, a data set from a household survey was used, and the effect of education on individuals' income levels has been assessed under different assumptions of missing data mechanisms by using a linear regression model. The modelling framework that was used comprised both the regression model and a missing data model using Bayesian estimation techniques jointly. Sensitivity analysis was carried out each time by changing the parameters of missing data model. It has been found out that under different assumptions of missing data mechanisms the parameter estimates of the regression model altered significantly.

Keywords*: Missing data, non-random missing data, selection models, Bayesian analysis, sensitivity analysis.*

1. Giriş

Hanehalkı araştırmalarında hanenin veya bireyin gelir düzeyi oldukça önemli bir demografik göstergedir. Ancak, böyle bir araştırmada, araştırmaya katılan bireylerin gelir düzeylerini açığa çıkaran soru veya sorulara cevap vermemeleri sıklıkla karşılaşılan bir durumdur. Nüfus araştırmalarında, bireylerin gelir düzeyinin %20 ile 50% seviyelerinde kayıp olarak ölçülebildiği ortaya konulmuştur [7, 12]. Bireylerin gelir seviyesi ile ilgili sorular, doğaları gereği hassastır ve katılımcılar böylesine soruları özel buldukları için, bu bilginin üçüncü şahıslarla paylaşılmasının güvenlik açısından riskli olduğunu düşünüp, yanıt

vermeyebilirler [13]. Mandal ve Stasny'e [10] göre nüfus veya hanehalkı araştırmalarında, onlar yüz yüze yapılan anketlerle gerçekleştirildiğinde, katılımcıların anketöre gelir düzeyini söyleme konusunda isteksiz davranmaları ve bunu riskli bulmalarının nedeni, gelir düzeylerinin çok düşük veya çok yüksek olması ile ilişkilidir. Bu nedenle, hanehalkı araştırmalarında gelir değişkenin çoğu zaman rastlantısal olmayan kayıp veri mekanizmasına sahip olduğu söylenebilir. Ancak istatistiksel analizler gerçekleştirilirken, bu kayıp veri mekanizması çoğunlukla göz ardı edilmektedir.

Anket tekniğiyle toplanan veri setleri analiz edilirken, çoğu zaman her soruya yanıt vermemiş olan bireyler analizden çıkartılarak analiz gerçekleştirilir. Ancak bu yöntem, istatistiksel olarak iki probleme yol açmaktadır. Birincisi, eğer her soruya cevap vermeyen katılımcılar sayıca çok ise, bu durum istatistiksel analizde güç kaybına neden olacaktır. İkincisi ve daha önemlisi, oluşan kayıp veriler rastlantısal değilse, istatistiksel analizin yanlı sonuç vermesi sözkonusudur.

Kayıp veri içeren veri setlerinde, bu verileri tahmin edebilmek için hangi yöntemin kullanılacağı, kayıp verilerin oluşma nedenini açıklayan kayıp veri mekanizmasına, başka bir deyişle kayıp verilerin rastlantısal dağılıp dağılmadığına bağlıdır. Rubin [15] ve Rubin ve Little [17], kayıp veri mekanizmalarını üç ana kategori altında sınıflandırmıştır. Bu sınıflandırma, kayıp veri olup olmama olasılığının veri setinde gözlemlenen veya gözlemlenmeyen değerlerle ilişkisine dayanmaktadır.

Kesitsel bir hanehalkı çalışmasında her *i* bireyi için bağımlı değişken değeri Y_i olarak belirlensin. Bu durumda, bu bağımlı değişkenin gözlemlenen ve kayıp değerlerini göstermek üzere, $\boldsymbol{Y}$ vektörü $\boldsymbol{Y}^{gözlenen}$ ve $\boldsymbol{Y}^{kayıp}$ şeklinde iki parçaya ayrılabilir. Yine aynı çalışmada kayıpsız olarak gözlenen *p* tane bağımsız değişken, $\boldsymbol{X} = (\boldsymbol{X_1}, \boldsymbol{X_2}, \ldots, \boldsymbol{X_k})$ ($k=1,\ldots, p$) matrisi ile tanımlanmış olsun. Bu koşullarda, her bir *i* bireyi için bağımlı değişken Y_i'ye ait kayıp veri indeks matrisi $\boldsymbol{R}$, şu şekilde tanımlanabilir:

$$R_i = \begin{cases} 1, \text{ eğer } Y_i \text{ kayıp ise} \\ 0, \text{ eğer } Y_i \text{ gözlemlendi ise} \end{cases}$$

Bu gösterim altında kayıp veri mekanizmaları, aşağıda belirtilen üç kategori altında toplanabilir:

1. Tamamen rastlantısal kayıp veri mekanizması (TRKV): Kayıp veri olma olasılığı, veri setinde kayıp veri içeren bağımlı değişkenin gözlenen veya kayıp hiçbir değeri ile ilişkili değildir. Sembolik olarak ifade edersek,

$$f\left(\boldsymbol{R}\middle|\boldsymbol{Y}^{gözlenen}, \boldsymbol{Y}^{kayıp}\right) = f(\boldsymbol{R}).$$

   Bu varsayım altında, değişkenlerde kayıp verilerin oluşma olasılığı ile bu değişkenlerin gözlenme olasılığı birbirine eşittir. Uygulamada bu, sağlanabilmesi en zor varsayımdır [14]. Kayıp gözlemler içeren veri seti, kayıpsız olarak gözlenebilecek orijinal veri setinin basit rastgele bir örneklemidir ve bu nedenle kayıp veri içeren gözlemler çıkarılarak gerçekleştirilen bir istatistiksel analiz, doğru sonuç verir.

2. Rastlantısal kayıp veri mekanizması (RKV): Kayıp veri içeren değişkenlerdeki kayıp veri oluşma olasılığı sadece gözlenen değerlerle ilişkilidir, ancak kayıp veri içeren değişkenlerden bağımsızdır.

$$f\left(\boldsymbol{R}\middle|\boldsymbol{Y}^{gözlenen}, \boldsymbol{Y}^{kayıp}\right) = f\left(\boldsymbol{R}\middle|\boldsymbol{Y}^{gözlenen}\right).$$

   Bu varsayım altında, bağımlı değişkendeki kayıp veri olasılığı, bağımsız değişkenlerdeki gözlenen veya kayıp verilerle de ilişkili olabilir. Şöyle ki,

$$f\left(\boldsymbol{R}\middle|\boldsymbol{Y}^{gözlenen}, \boldsymbol{Y}^{kayıp}\right) = f\left(\boldsymbol{R}\middle|\boldsymbol{Y}^{gözlenen}, \boldsymbol{X}\right).$$

Bu varsayım altında kayıp veri eğer sadece bağımlı değişkende oluştuysa, kayıp veri içeren gözlemleri çıkararak gerçekleştirilen bir istatistiksel analiz, tamamen rastlantısal kayıp veri mekanizmasında olduğu gibi, doğru sonuç verir.

3. Rastlantısal olmayan kayıp veri mekanizması (ROKV): Bağımlı değişkendeki kayıp veri olasılığı, bu değişkenin kendisindeki kayıp veriler ile ($\boldsymbol{Y}^{kayıp}$), ilişkilidir. Bu mekanizma altında, kayıp verilerin neden gerçekleştiğine dair bir varsayım, bileşik modeller kullanılarak mutlaka istatistiksel analize dâhil edilmelidir. Ancak verilerin neden kayıp olduğuna dair bir ön bilgi olmadan doğrulanamayan bu varsayımın istatistiksel analize dâhil edilmesi, diğer yöntemlere göre daha karmaşık istatistiksel modellerle mümkündür. Uygulamadaki bu zorluk nedeni ile rastlantısal olmayan kayıp veriler varlığında istatistiksel modelleme, literatürde geniş bir uygulama alanı bulamamıştır.

Başka bir tanımlamada, rastlantısal kayıp veri mekanizmaları (1. ve 2.) ihmal edilebilir, rastlantısal olmayan kayıp veri mekanizması (3.) ise ihmal edilemeyen kayıp veri mekanizması olarak belirlenmiştir [17]. Bu tanımda ihmal edilip edilemeyen, kayıp verilerin kendisi değil, kayıp veri mekanizmasıdır. Kayıp veri mekanizması ihmal edilemez ise, uygulanacak istatistiksel analiz modeli ile birlikte kayıp veri mekanizmasının da modellenmesi gerekmektedir.

Bilimsel çalışmalarda kayıp veriler varlığında yapılan analizler, sıklıkla bu verilerin ihmal edilebilir olduğunu varsayan (tam durum analizi, "hot deck" veri atama, çoklu atama, vb. gibi) basit yöntemlere dayanmaktadır. Oysa ki kayıp veriler, rastlantısal olmayan kayıp veri mekanizmasında olduğu gibi, bağımlı değişkenin gözlemlenmeyen değerleri ile ilişkiliyse, bu yöntemleri kullanmak regresyon modelinin parametre değerlerinin yanlı tahmin edilmesine neden olabilir [1]. Rastlantısal olmayan kayıp veri mekanizmasında en büyük zorluk, kesin bir kanıt olmadan sadece gözlenen değerlere bakarak bir bilinmeyen hakkında doğru varsayımı yapabilmektir. Örneğin, bir ankete katılan bireylerin niçin gelir düzeyini cevaplamadığını bilmediğimiz için, kayıp veri mekanizmasını da kesin doğru olarak tahmin etmemiz mümkün değildir. Bu nedenle kayıp verilerin rastlantısal dağıldığına dair kesin bir önbilgi yoksa, uygulanan istatistiksel model parametrelerinin ne kadar değiştiği, farklı kayıp veri mekanizmaları varsayımları altında yapılabilecek bir duyarlılık analizi yardımıyla karşılaştırılmalıdır [11, 18, 19]. Bunun için kayıp veri mekanizması ile analiz modeli bileşik olarak modellenir. Bu bileşik modelleme için literatürde en sık kullanılan yöntemlerden biri seçim modelleri'dir [6].

Her ne kadar kayıp verilerin varlığında, duyarlılık analizi yoluyla farklı kayıp veri mekanizmaları altında analiz sonuçlarının ne kadar değiştiğinin gözlemlenmesi mutlaka uygulanması gereken bir basamak ise de, gerek bu konuda yol gösterecek kaynak azlığından ve gerekse uygulama zorluğundan ötürü bu basamak, çoğu çalışmada atlanmaktadır [1]. Ancak gözardı edilen ve atlanan bu basamak, yanlış istatistiksel sonuçlara yol açabilmektedir.

Bu çalışmanın amacı, rastlantısal olmayan kayıp veriler varlığında bir duyarlılık analizi uygulaması sunmak ve böylelikle araştırmacılara yol göstermektir. Sözkonusu amaç doğrultusunda kayıp verilere sahip tek bağımlı değişken içeren doğrusal regresyon modeli parametreleri, farklı kayıp veri mekanizmaları altında duyarlılık analizi kullanılarak incelenecek ve regresyon modeli parametrelerinin bu farklı mekanizmalar altında ne ölçüde farklılaşabileceği ortaya konulacaktır. Kullanılan veri seti, 300 haneyle gerçekleştirilen pilot hanehalkı anket çalışmasından elde edilmiştir. Yüzyüze görüşmelerle toplanan bu veri setinde, bağımlı değişken olarak kullanılacak gelir düzeyi değişkenine, katılımcıların önemli bir bölümü cevap vermemişlerdir. Bu katılımcıların gelir sorusunu neden cevaplamadığına dair bir bilgi bulunmamaktadır. Bu alandaki diğer çalışmalarda da [10, 13] belirtildiği üzere, gelir düzeyi değişkeni ile bir analiz gerçekleştirirken bu değişkenin rastlantısal olmayan kayıp verilere sahip olabileceği göz önünde bulundurulmalıdır.

2. Rastlantısal Olmayan Kayıp Veri Mekanizması ile İstatistiksel Analiz

2.1 Seçim Modelleri

Seçim modelleri kayıp verinin, veri setindeki değişkenlerle ilişkisine odaklanmaktadır. Bu modeller verinin neden kayıp olduğuna veya diğer bir deyişle gözlenen değerlerin nasıl seçildiğine dayandığı için, bu şekilde adlandırılmıştır [11]. Bağımlı değişkende kayıp veriler var ise ve bağımsız değişkenler kayıpsız olarak gözlemlendiğinde, seçim modelleri kayıp veri indeksi $\boldsymbol{R}$ ve bağımlı değişken $\boldsymbol{Y}$'nin bileşik olarak modellenmesinden oluşur. Bu bileşik model, $\boldsymbol{Y}$'nin marjinal dağılımı ve verilen $\boldsymbol{Y}$ için $\boldsymbol{R}$'in koşullu dağılımı yardımıyla aşağıdaki şekilde tanımlanır:

$$f(\boldsymbol{Y},\boldsymbol{R}|\beta,\delta) = \int f(\boldsymbol{Y},\boldsymbol{R}|\beta,\delta)\, d\boldsymbol{Y}^{\boldsymbol{kayıp}} = \int f(\boldsymbol{Y}|\beta) f(\boldsymbol{R}|\boldsymbol{Y},\delta)\, d\boldsymbol{Y}^{\boldsymbol{kayıp}} \quad (1)$$

Yukarıdaki eşitlikte, $f(\boldsymbol{Y}|\beta)$ analiz modelini belirtmektedir. Bu analiz modeli, veri setinin yapısına göre herhangibir regresyon modeli olarak belirlenebilir. β, bu analiz modeline ait parametre tahminleridir. Örneğin, sürekli bağımlı değişkenler için, aşağıda verilen doğrusal regresyon modeli şu şekildedir:

$$Y_i = \beta_0 + \sum_{j=1}^{k} \beta_j X_j + \varepsilon_i, \quad \varepsilon_i \sim N(0,\sigma^2) \quad (2)$$

$f(\boldsymbol{R}|\boldsymbol{Y},\delta)$ ise kayıp veri modelidir. δ, bu kayıp veri modeline ait parametreleri ifade etmektedir. Diggle ve Kenward [2] kayıp veri modelini lojistik regresyon modeli olarak tanımlamışlardır ve bu model kayıp verilerinin oluşma olasılığının bağımlı değişken $\boldsymbol{Y}$'nin değerleri ile olan ilişkisini şu şekilde irdeler:

$$logit\big(P(R_i = 1)\big) = \delta_0 + \delta_1 Y_i \quad (3)$$

Rastlantısal olmayan kayıp veri mekanizmasında, istatistiksel olarak en büyük problem, yukarıdaki lojistik regresyon modelinde δ_1 ile gösterilen kayıp veri parametresinin, gözlemlenen $\boldsymbol{Y}$ değerleri kullanarak tahmin edilmesidir. Çünkü, $R_i = 1$ olduğu durumlarda Y_i'nin değeri bilinmemektedir. Bu problemi çözmek, ancak Model 1 ve 2'yi bileşik olarak modellemek ile mümkün olur.

2.2 Bayesci Yaklaşım ile Parametre Tahmini ve Bileşik Modelleme

Bayesci yaklaşımda regresyon modelinin parametreleri, olasılık dağılımı olan birer rastlantı değişkenidir. Bu nedenle bu parametreler için birer önsel beklenti dağılımı tanımlanır. Bu önsel beklenti dağılımlarından gelen önbilgiler ve veri setinde gözlenen değerler birleştirilerek analiz, Markov Zinciri Monte Carlo (MZMC) yaklaşımı ile birden fazla (çoğu zaman binlerce, onbinlerce kez) tekrarlanarak, her bir parametre için sonsal beklenti dağılımları hesaplanır [3]. Analiz sonunda regresyon parametrelerine ait nokta tahminleri bu sonsal beklenti dağılımlarının ortalaması alınarak hesaplanır. Aralık tahminleri ise, yine sonsal beklenti dağılımlarını kullanarak hesaplanır. Örneğin, bir regresyon katsayısının %95 güvenilir aralığı, bu parametrenin sonsal dağılımından elde edilen ortalamaların %95'ini kapsayan aralık olarak tanımlanır. Bayesci tahmin yöntemlerinin günümüzdeki uygulamalarının pek yaygın olmamasındaki ana nedenlerden biri, frekansçı yöntemlerde olduğu gibi parametrelerin istatistiksel anlamlılık düzeyini gösteren p-değeri türünde net bir çıktı vermemesidir. Ancak, regresyon katsayıları için güvenilir aralıklarının 0 değerini içerip içermemesine göre, Bayesci yöntemler ile de, bu katsayıların istatistiksel olarak anlamlı olup olmadığına dair olasılığa dayalı çıkarsamalar yapmak mümkündür.

Bayesci tahmin yöntemi ile Bölüm 2.1'de açıklanan 2 ve 3 numaralı modelleri bileşik olarak modellemek için, her iki modelin parametrelerine ait önsel beklenti dağılımları tanımlanır ve MZMC yaklaşımı kullanılır. Her bir MZMC tekrarında, analiz modeli ile kayıp veri modeli eş zamanlı olarak modellenir ve iki model arasındaki bağ, δ_1 yardımıyla kurulur. Kayıp veri parametresi δ_1 için belirlenen önsel beklentiler, kayıp veri mekanizması hakkında bilgi vermektedir. Böylece, kayıp olan Y_i değerleri δ_1 için varsayılan önbilgiler ışığında tahmin edilebilir. Belirlenen δ_1 değeri için, kayıp olan Y_i değerlerine her bir MZMC tekrarında farklı bir değer atanacaktır. Bu nedenle her bir MZMC tekrarında β_j'ler için de farklı

değerler elde edilir. MZMC öngörülen sayı kadar tekrar edildiğinde, her bir tekrardan elde edilen β_j değerlerinin ortalaması ise, β_j'lerin nokta tahminleri olarak kullanılır.

Kayıp veri problemlerinde Bayesci tahmin yöntemleri kullanarak çıkarsama yapmak ve bileşik model kurmak, en çok olabilirlik yöntemine göre uygulama açısından daha kolaydır. Ayrıca Bayesci yöntemler ile yapılan parametre tahminleri, kayıp veri modelinin yanlış tanımlanmasından kaynaklanacak hatalara karşı daha az hassastır [9]. Ancak Bayesci tahmin yöntemlerinde birden çok model kullanıldığı takdirde, bu modellerin birbirlerine nazaran yakınsamaları mutlaka Gelman ve Rubin [4] veya Geweke [5] yakınsaklık testleri ile test edilmelidir.

2.3 Duyarlılık Analizi

Yukarıda anlatılan bileşik modelde, δ_1 hakkında herhangibir önsel bilgi olmadan, kayıp veri mekanizması hakkında bir varsayımda bulunulamaz. Bu nedenle duyarlılık analizi uygulanarak, kayıp verilerin rastlantısal kayıp veri mekanizmasından ne kadar uzakta olduğu gözlemlenmelidir. Duyarlılık analizinin ana fikri, makûl rastlantısal olmayan kayıp veri mekanizması varsayımları altında analiz modelini uygulamak ve analiz modeli parametrelerinin bu farklı varsayımlar altında ne kadar değiştiğini gözlemlemektir. Öyle ki, bileşik modelde kayıp veri ile Y_i arasındaki ilişki, yani kayıp veri mekanizmasını tanımlayan kayıp veri parametresi δ_1 değeri için farklı önsel beklenti dağılımları tanımlanarak veya onlara farklı sabit değerler verilerek, β_j değerlerinin ne kadar değiştiğine bakılır. Model 3'te pozitif δ_1 değerleri, Y_i'nin değeri arttıkça kayıp olma olasılığının arttığını, negatif δ_1 değerleri ise, Y_i'in değeri arttıkça kayıp olma olasılığının azaldığını belirtir. δ_1'in 0 olması halinde, gözlenmeyen Y_i değerlerinin rastlantısal olarak kayıp olduğu ve kayıp veri oluşma olasılığının Y_i'nin aldığı değerlerden bağımsız olduğu anlaşılır.

3. Uygulama

3.1 Veri seti

Bu araştırmada kullanılan veri seti, Ankara kent merkezindeki hanelerin ve bu hanelerde yaşayan bireylerin sosyo-ekonomik göstergeleri üzerine gerçekleştirilen bir hanehalkı araştırmasının pilot çalışmasından elde edilmiştir. Bu pilot çalışma kapsamında, yüzyüze görüşmelerle yapılan anketlerle 300 haneden veri toplanması öngörülmüştür. Veri toplama aşamasında, o anda hanede bulunan 18-69 yaş arası bireylerden rastgele seçilen biriyle görüşülmüş ve görüşülen kişiden, hem kendisi hem de diğer aile fertleri hakkında bilgi edinilmiştir.

Bu makalede sunulan istatistiksel çalışmanın amaçları doğrultusunda, bu görüşülen kişilerden herhangi bir kaynaktan gelir elde eden 232 birey, veri setine dâhil edilmiştir. Bu kapsamda, sözkonusu bireylerin aylık gelir seviyesi bağımlı değişken; eğitim yılı, yaşı ve cinsiyeti de bağımsız değişkenler olarak belirlenmiştir. Ancak gelir seviyesi değerleri, 232 bireyden 67'si için elde edilememiştir; bu da yaklaşık %29 oranında bir kayıp veri olmasına neden olmuştur. Kayıp veriler sadece bağımlı değişkende ortaya çıkarken, bağımsız değişkenler kayıpsız olarak gözlemlenmiştir. Kullanılan veri seti Çizelge 1.'de özetlenmiştir.

Çizelge 1. Bağımlı ve Bağımsız Değişkenler.

Değişkenler	**Açıklama**	**Tür**	**Ortalama (Standard Sapma)**	**Aralık**	**Kayıp veri %'si**
Bağımlı Değişken					
`Gelir`	Bireyin kazandığı aylık gelir (TL)	sürekli	1661.0 (1454.6)	250 - 12500	28.9
Bağımsız Değişkenler					
`Eğitim yılı`	Bireyin katıldığı toplam formal eğitim yılı	sürekli	10.2 (5.5)	0-20	0
`Cinsiyet`	Kadın(temel kategori)/ Erkek	ikili	Kadın (%53.3) Erkek(%46.7)	0/1	0
`Yaş`	Bireyin yaşı	sürekli	44.8 (19.2)	19-86	0

3.2 İstatistiksel Analiz

Gelir düzeyi üzerinde bireyin eğitimi, yaşı, cinsiyeti gibi faktörlerin etkisini ölçmek için analiz modeli, doğrusal regresyon modeli olarak belirlenmiştir. Bağımlı değişkenin normallik varsayımını sağlamak ve β katsayılarını anlamlı kılmak için gelir değişkeni 100'e bölünerek logaritması alınmıştır. Bayesci yaklaşımda, bağımlı değişkenin, ortalaması μ ve varyansı σ^2 olan bir normal dağılımdan gelen, sürekli bir rastgele değişken olduğu varsayılmaktadır. Bu durumda,

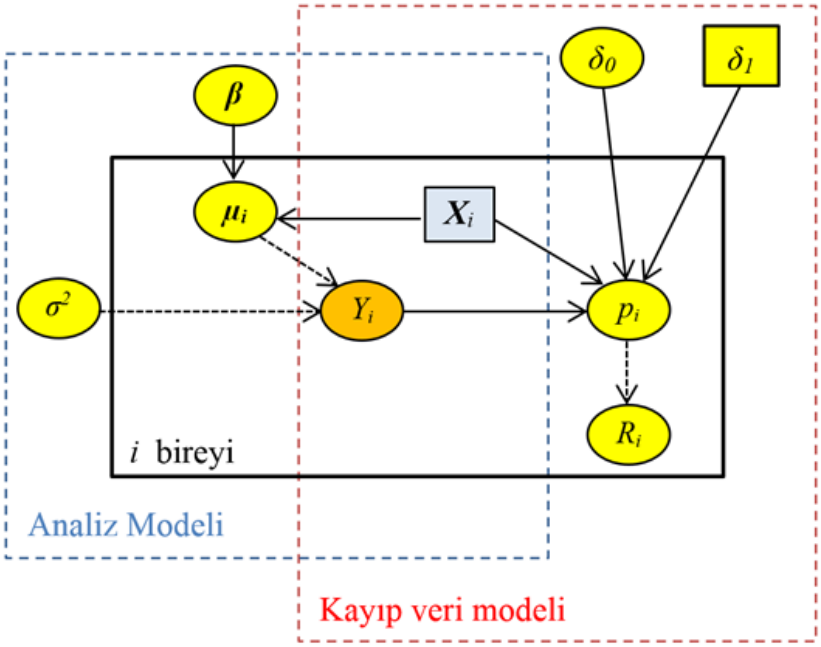
$$\log(gelir_i/100) \sim N(\mu_i, \sigma^2) \tag{4}$$

$$\mu_i = \beta_0 + \beta_1 eğitimyılı_i + \beta_2 erkek_i + \beta_3 yaş_i$$

Kayıp veri mekanizmasını gösteren kayıp veri modeli ise, lojistik regresyon kullanılarak modellenmiştir. Bu lojistik regresyon modelinde kayıp veri indeksinin (R_i), olasılığı p_i olan bir Bernoulli dağılımına sahip olduğu tanımlanmıştır. O halde,

$$R_i \sim Bernoulli(p_i) \tag{5}$$

$$logit(p_i) = \delta_0 + \delta_1 \log(gelir_i/100)\,.$$



Şekil 1. Analiz Modeli ve Kayıp Veri Modeline Ait Parametrelerin ve Değişkenlerin Yönlü Çevrimsiz Grafik ile Görselleştirilmesi.

Analiz modeli ve kayıp veri modelinin bileşik olarak Bayesci yaklaşımla modellenmesi, Şekil 1'de Bayes ağları ile görselleştirilmiştir. Bu kapsamda, düğümler ve oklar aracılığıyla bileşik modelde kullanılan değişkenler arasındaki bağımlılık ilişkilerini gösteren yönlü çevrimsiz grafik kullanılmıştır. Bu grafikte, görsel basitlik için

$Y_i = \log(gelir_i/100)$, $\boldsymbol{X_i} = (eğitimyılı_i, yaş_i, cinsiyet_i)$ ve $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)$ şeklinde

tanımlanmıştır.

Bileşik modelde, türlerine göre üç farklı düğüm vardır:

1. *Sabit düğümler* ($\boldsymbol{X}_i$, δ_1), her bir MZMC tekrarı için aynı sabit değeri alan düğümlerdir. Herhangi bir önsel beklenti dağılımı yoktur. Şekil 1'de kare içinde gösterilmiştir.

2. *Stokastik düğümler* ($Y_i, \sigma^2, \boldsymbol{\beta}, \delta_0, R_i$), önsel beklenti dağılımları MZMC başlamadan tanımlanan düğümlerdir. Değeri bilinmeyen model parametreleri ve kayıp veriler stokastik düğüm tipindedir. Şekil 1'de yuvarlak ile gösterilmiştir. Analizde bağımlı değişken olan Y_i'nin, ortalaması μ_i, varyansı σ^2 olan bir normal dağılımdan geldiği varsayılmıştır. Bu nedenle Y_i stokastik bir düğümdür ve Şekil 1'de μ_i ve σ^2 ile arasındaki ilişki kesikli oklar ile gösterilmiştir.

3. *Deterministik düğümler* (μ_i, p_i), stokastik düğümlerin birer mantıksal fonksiyonudur. Başka bir deyişle bileşik modelde kullanılan Model 4 ve 5'in bağımlı değişkenleridir. Şekil 1'de yuvarlak içinde gösterilmiştir. Ancak bu düğümleri modellemek için kullanılan değişkenler ve parametreler ile aralarındaki ilişki, kesiksiz oklar ile gösterilmiştir.

Yukarıdaki yönlü çevrimsiz grafikte yer alan her bir düğümün türü, bu düğümlere ait önsel beklenti dağılımları veya değerleri, Çizelge 2'de özetlenmiştir.

Çizelge 2. Şekil 1'de Kullanılan Düğümlerin Detaylı Açıklaması

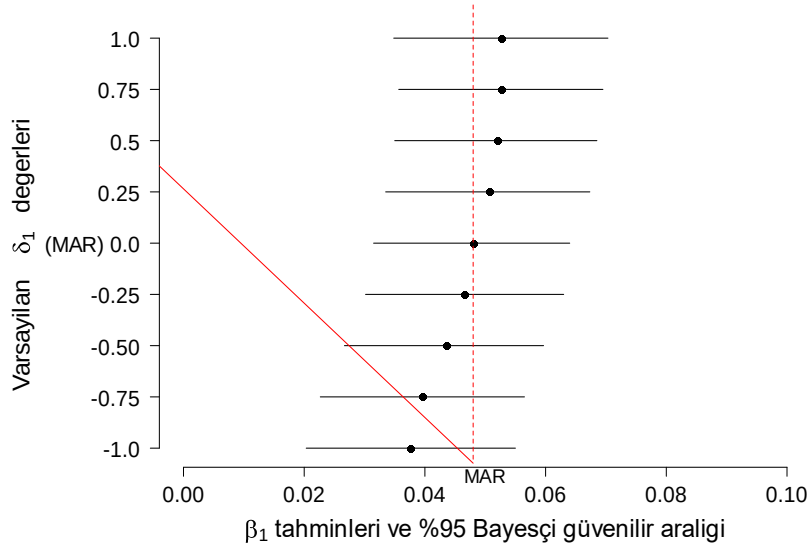
Düğüm	Tür	Dağılım[1]	Değer[2]
Y_i	Stokastik	$N(\mu_i, \sigma^2)$	
μ_i	Deterministik	-	$\beta_0 + \beta_1 eğitimyılı_i + \beta_2 erkek_i + \beta_3 yaş_i$
σ^2	Stokastik	$1/Gamma(0.01,0.01)$	
$\boldsymbol{X}_i$	Sabit	-	-
$\boldsymbol{\beta}$	Stokastik	$N(0, 0.1^4)$	
δ_0	Stokastik	$N(0, 0.1^4)$	
δ_1	Sabit	-	Duyarlılık analizinde farklı sabit değerler alır.
p_i	Deterministik	-	$\delta_0 + \delta_1 \log(gelir_i/100)$.
R_i	Stokastik	$Bernoulli(p_i)$	

[1] Stokastik düğümler için önsel beklenti dağılımları, [2] Mantıksal düğümler için kullanılan fonksiyonlar.

Analiz modeli ve kayıp veri modelinin Bayesci tahmin yöntemleri ile bileşik olarak modellenmesi WinBUGS programında [8] yapılmıştır. Toplam 50000 MZMC tekrarı sonucunda parametre değerlerinin farklı zincirler için birbirine yakınsandığı görülmüştür. İlk 5000 tekrarda yakınsama tam sağlanamayacağından, istatistiksel çıkarsamalar için kalan 45000 MZMC tekrarından elde edilen sonsal beklenti dağılımları kullanılmıştır.

3.3 Bulgular

Analiz modelinden elde edilen regresyon katsayılarının, farklı kayıp veri mekanizmalarına karşı ne kadar değişkenlik gösterdiği, duyarlılık analizi ile incelenmiştir. Duyarlılık analizinde rastlantısal olmayan kayıp veri mekanizmasının parametresi δ_1 için (-1.0, -0.75, -0.50, -0.25, 0, 0.25, 0.50, 0.75, 1.0) sabit değerleri tek tek ve ayrı ayrı tanımlanarak, bu değerler altında ilgilenilen `eğitimyılı` değişkenine ait β_1 katsayının ne kadar değiştiği irdelenmiştir (Aşağıda Şekil 2'de de ilgili gösterim sunulmuştur.). δ_1 katsayısı, kayıp verinin göreceli olasılık oranının logaritmasının, `log(gelir/100)` değişkenindeki bir birimlik artışla veya başka bir deyişle, bireyin gelir seviyesindeki 10 kat artışla, nasıl değiştiğini açıklar. Örneğin, δ_1 katsayısının 0.25 değerini alması, `log(gelir/100)` değişkenindeki bir birim artışın, kayıp veri oluşma olasılığını yaklaşık %28 arttığını belirtmektedir, (çünkü $e^{0.25} = 1.284$). Yine aynı şekilde, bu katsayının -0.25 değerini alması ise, `log(gelir/100)` değişkenindeki bir birim artışın, kayıp veri oluşma olasılığını yaklaşık %22 azalttığını belirtir, (çünkü $e^{-0.25} = 0.778$).



Şekil 2. Farklı δ_1 Varsayımları Altında β_1 Katsayısının Nokta Tahminleri ve %95 Bayesci Güvenilir Aralıkları

Analiz bulgularından da anlaşılacağı üzere, bireyin katıldığı formal eğitim yılı arttıkça, kazandığı aylık gelir miktarı da artmaktadır ($\beta_1 > 0$) (Sonuçlar, Çizelge 3'de verilmiştir.). Ancak eğitim yılı ile gelir seviyesi arasındaki bu pozitif ilişki, gelir seviyesi hakkındaki kayıp veri mekanizmasına oldukça duyarlıdır. Eğer Rubin'in [16] belirttiği gibi, yüksek gelirli kişiler diğer sorulara eksiksiz cevap verirken gelir düzeylerini söylemekten kaçındı ise, yani gerçekte gelir sorusunu cevaplamayan bireyler çoğunlukla yüksek gelirli bireyler ise, rastlantısal kayıp veri (RKV) varsayımı altında kayıp olan gelir düzeyi gözlemleri, gerçekte olduğundan daha düşük olarak tahmin edilir. Ayrıca, yüksek gelir seviyesindeki bireylerin eğitim düzeyinin de genel olarak daha yüksek olduğu düşünüldüğünde, RKV varsayımı altında yüksek eğitim seviyesindeki bireylerin gelir seviyesi gerçekte olduğundan daha düşük tahmin edilmektedir. Bu da, eğitimin gelir seviyesi üzerindeki etkisinin daha az olarak ölçülmesine neden olur. Ancak, δ_1 katsayısına pozitif değerler verildiğinde, yani bileşik model, gelir seviyesinin yüksek değerlerinin kayıp olma olasılığının daha fazla olduğu konusunda bilgilendirildiğinde, bu kez gelir düzeyinin kayıp değerleri RKV varsayımına göre daha yüksek olarak tahmin edilir. Bu da eğitimin gelir düzeyi üzerindeki etkisini, yani tahmin edilen β_1 katsayısını, RKV varsayımına göre arttırmaktadır. δ_1 katsayısının negatif değerlerinde ise, kayıp olan gelir düzeyi gözlemlerine daha küçük değerler atanacağından, β_1 katsayısı RKV varsayımına göre küçülmektedir.

Çizelge 3. Duyarlılık Analiz Sonuçları (Regresyon Katsayıları ve Bu Katsayılara ait Standart Hatalar (SH) ve %95'lik Güvenilir Aralıkları (GA)).

	Değişkenler								
	Eğitim yılı			**Cinsiyet**			**Yaş**		
δ_1	β_1	**SH**	**%95 GA**	β_2	**SH**	**%95 GA**	β_3	**SH**	**%95 GA**
-1	0.037	0.009	(0.020, 0.055)	0.182	0.080	(0.026,0.342)	-0.002	0.002	(-0.005,0.004)
-0.75	0.039	0.008	(0.022, 0.056)	0.170	0.079	(0.013,0.326)	-0.001	0.002	(-0.005, 0.004)
-0.5	0.043	0.008	(0.026, 0.059)	0.168	0.076	(0.014, 0.311)	0.001	0.002	(-0.004, 0.005)
-0.25	0.046	0.008	(0.030, 0.063)	0.153	0.075	(0.008, 0.302)	0.001	0.002	(-0.004, 0.005)
0	0.048	0.008	(0.031, 0.063)	0.140	0.075	(-0.007, 0.286)	0.002	0.002	(-0.003, 0.005)
0.25	0.050	0.008	(0.033, 0.067)	0.129	0.075	(-0.018, 0.277)	0.002	0.002	(-0.003, 0.006)
0.5	0.052	0.008	(0.035, 0.068)	0.118	0.077	(-0.033, 0.272)	0.002	0.002	(-0.003, 0.006)
0.75	0.053	0.008	(0.036, 0.069)	0.104	0.078	(-0.050, 0.258)	0.003	0.002	(-0.002, 0.006)
1	0.053	0.009	(0.034, 0.070)	0.093	0.080	(-0.064, 0.252)	0.003	0.003	(-0.003, 0.007)

Diğer değişkenlerin katsayılarına ait detaylı sonuçlar da Çizelge 3'de verilmiştir. Buna göre sadece eğitim yılına ait regresyon katsayısı değil, aynı zamanda cinsiyet değişkeninin de regresyon katsayısı farklı kayıp veri varsayımları altında değişmektedir. Hatta bu değişim istatistiksel anlamlılık düzeyine de etki etmektedir. $\delta_1 \geq 0$ olduğunda, β_2 için %95% güvenilir aralığı 0'ı içermektedir. Bu da cinsiyetin gelir düzeyi için istatistiksel olarak anlamlı bir etken olmadığını belirtir. Ancak, düşük gelir seviyesindeki bireylerin kayıp veri olasılığı daha yüksek ise ($\delta_1 < 0$), erkekler aynı eğitim düzeyindeki ve yaştaki kadınlara göre daha çok aylık gelir etmektedir. Yaş değişkenine ait regresyon katsayısında, farklı kayıp veri mekanizmaları arasında büyük bir fark oluşmamıştır.

4. Sonuç ve Öneriler

Bu çalışmada, rastlantısal olmayan kayıp veri durumunda duyarlılık analizi, bir hanehalkı çalışmasından elde edilen verileri kullanmak suretiyle uygulamalı olarak ele alınmıştır. Bu uygulama kapsamında, bağımlı değişkeni kayıp veriler içeren bir doğrusal regresyon modeli ele alınmış ve farklı kayıp veri mekanizmaları altında bu regresyon modeline ait parametrelerin nasıl değiştiği incelenmiştir. Bunun için seçim modelleri çerçevesinde, Bayesci yaklaşım ile analiz modeli olan doğrusal regresyon modeli ile kayıp veri modeli olan lojistik regresyon modeli bileşik olarak modellenmiştir.

Duyarlılık analizi, bu bağlamda, regresyon modeli parametrelerine ait birer parametre tahmini vermek yerine, bu parametrelerin tahminlerinin farklı kayıp veri mekanizmaları altında ne ölçüde değiştiğine dair araştırmacıya yol gösteren bir yöntem olmaktadır. Farklı kayıp veri modelleri altında regresyon katsayılarının ciddi ölçüde değişmesi, bu veri setinde rastlantısal olmayan kayıp veri mekanizmasının gerçekçi bir varsayım olduğunu gösterir. Nitekim Çizelge 3'de, bu katsayıların farklı δ_1 katsayıları için oldukça farklı değerler aldığı görülmektedir ve rastlantısal olmayan kayıp veri varsayımının bu veri seti için gözönünde bulundurulması gereken bir yaklaşım olduğu açığa çıkmıştır.

Bu çalışmada kullanılan veri setinde kayıp verilerin neden oluştuğuna dair herhangibir önbilgi mevcut değildir. Rastlantısal olmayan kayıp veri mekanizması altında, eğer mümkünse, veri toplayanlarla veya araştırmayı yürüten uzmanlarla görüşülerek verilerin neden kayıp veri olarak ölçüldüğüne dair bilgi toplanabilir. Bu bilgiler önsel beklenti dağılımı olarak kayıp veri modeline dâhil edilebilir. Böylece kayıp olan değerlerin, gerçek değerlerine en yakın tahminler elde edilerek, regresyon parametrelerinin daha doğru bir şekilde tahmin edilebileceği düşünülmektedir.

Kaynaklar

[1] P.D Allison, 2002, *Missing Data*. Sage Publications Inc, California.
[2] P. Diggle, *M. G.* Kenward, 1994, Informative drop-out in longitudinal data-analysis, *Applied Statistics*, 43, 49–93.
[3] A. Gelman, J.B. Carlin, H.S. Stern, D. B. Rubin, 2004, *Bayesian Data Analysis*, Chapman & Hall, Florida.
[4] A. Gelman, D. Rubin, 1992, Inference from Iterative Simulation using Multiple Sequences, *Statistical Science*, 7, 457-511.
[5] J. Geweke, 1992, *Evaluating the Accuracy of Sampling-Based Approaches to Calculating Posterior Moments*, J. M. Bernardo, J. M. Berger, A. P. Dawiv, A. F. Smith, Bayesian Statistics, (s. 196-193), University Press, Oxford.
[6] J.J. Heckman, 1979, Sample Selection Bias as a Specification Error, *Econometrica*, 47, 153-161
[7] F.T. Juster, J.P. Smith, 1997, Improving the Quality of Economic Data: Lessons from the HRS and AHEAD, *Journal of the American Statistical Association*, 92, 1268-1278.
[8] D. J. Lunn, A. Thomas, N. Best, D. Spiegelhalter, 2000, WinBUGS – A Bayesian modelling framework: Concepts, structure, and extensibility. Statistics and Computing, 10, 325–337.
[9] R. Little, 2008, *Selection and Pattern-Mixture Models*, G. Fitmaurice ve diğerleri, Advances in Longitudinal Data Analysis (18. Bölüm). CRC Press, London.
[10] B. Mandal, E. A. Stasny, 2004, Imputing Missing Income Data and Weighting Data with Imputed Income, *Proceedings of the 2004 Joint Statistical Meetings, American Statistical Association*, 3962- 3980.
[11] G. Molenberghs, M.G. Kenward, 2007, *Missing Data in Clinical Studies.* John Wiley and Sons, Chichester, UK.
[12] J. Moore, L. Stinson, E. Welniak, 1999, *Income Reporting in Surveys Cognitive Issues and Measurement Error*, M. D. Sirkin ve diğerleri (Ed.), Cognition and Survey Research (10. Bölüm). John Wiley & Sons, New York.
[13] C. Nicoletti, F. Peracchi, F. Foliano, 2011, Estimating income poverty in the presence of missing data and measurement error, *Journal of Business & Economic Statistics*, 29, 61–72.
[14] T. E. Raghunathan, 2004, What do we do with missing data? Some options for analysis of incomplete data, *Annual Review of Public Health*, 25, 88–117.
[15] D. B. Rubin, 1976, Inference and Missing Data, *Biometrika*, 63, 581-592.
[16] D.B. Rubin, 2004, *Multiple imputation for Nonresponse in Surveys.* John Wiley & Sons, New York.
[17] D.B. Rubin, F. Little, 2002, *Statistical Analysis with Missing Data, (Ed.),* John Wiley and Sons, New Jersey.
[18] J. A. Sterne, I. R. White., J. B. Carlin, M. Spratt, P. Royston, M. Kenward, et al., 2009, Multiple imputation for missing data in epidemiological and clinical research: Potential and pitfalls. *British Medical Journal*, 338, b2393.
[19] H. Thijs, G. Molenberghs, B. Michiels, G. Verbeke, D. Curran, 2002, Strategies to fit pattern-mixture models. *Biostatistics*, 3, 245-265