DERLEME REVIEW

DOI: 10.5336/biostatic.2019-66838

Regresyon Analizinde Sağdan Sansürlü Veriler İçin Önerilen Çözüm Yöntemleri Üzerine Bir İnceleme

A Review for Solution Methods on Right-Censored Data in Regression Analysis

Ersin YILMAZ[a],
Dursun AYDIN[a]

[a] İstatistik Bölümü,
Muğla Sıtkı Koçman Üniversitesi
Fen Fakültesi,
Muğla, TÜRKİYE

Received: 08.05.2019
Received in revised form: 24.07.2019
Accepted: 20.08.2019
Available Online: 20.12.2019

Correspondence:
Ersin YILMAZ
Muğla Sıtkı Koçman Üniversitesi
Faculty of Science,
Department of Statistics, Mugla,
TÜRKİYE/TURKEY
ersinyilmaz@mu.edu.tr

ÖZET Sağdan sansürlü veri, başta klinik deneyler ve sağlık alanı olmak üzere biyoloji, endüstri, ekonomi, genetik ve bu alanlarla ilişkili birçok alanda karşımıza çıkmaktadır. Bu veri türünün en önemli karakteristik özelliği, ilgilenilen kişi veya nesne ile ilgili tamamlanmamış gözlemeler içermesidir. Modelleme çalışmalarında, tamamlanmamış gözlemler yanlı ve tutarsız sonuçlara neden olduğundan, bu sorunun çözülmesi için çeşitli yöntemler geliştirilmiştir. Bu çalışmada, literatürde var olan ve sağdan sansürlü verilerin modelleme sürecine dâhil olabilmesi için kullanılan birçok farklı yöntem incelenmiştir. Bu yöntemlerden bazıları; Kaplan-Meier ağırlıkları, Gaussian ve kNN yerine koyma yöntemi, Sentetik veri dönüşümleri olarak sıralanbilir. Genellikle, sağdan sansürlü veri noktaları bilinmediğinden veya kısmi olarak bilindiğinden, bu gözlemlere ait dağılımlar hakkında bazı varsayımlar kabul edilerek klasik istatistiksel analiz ve modelleme yöntemleri kullanılabilmektedir. Buna ek olarak, dağılım varsayımlarına dayanmayan bazı parametrik olmayan yöntemler kullanılarak da bu gözlemler tahmin edilebilmektedir. Bu iki ana başlık dışında, çok tercih edilmese de sansürlü veri noktalarının veri setinden atılması da mevcut yöntemlerden biri olarak söylenebilir. Bu çalışmada, önerilmiş en basit yöntemlerden en gelişmiş yöntemlere kadar, sağdan sansürlü verilerin regresyon analizine dâhil edilmesi için önerilen çözüm yöntemleri aşamalar halinde sunulmuştur ve bu yöntemlerin sansürün etkisini ne kadar yansıtabildiği anlatılmaya çalışılmıştır. Bu çalışmanın temel amacı, verilerin regresyon modeline eklenmesinden önce, verinin içerdiği sansür durumu için gerekli düzenlemelerin yapılmasını sağlayan yöntemlerin incelenmesidir. Elbette var olan bütün yöntemlerin incelenmesi mümkün olmadığından, literatürde en sık kullanılan yöntemler seçilmiştir.

Anahtar Kelimeler: Sağdan sansürlü veri; regresyon analizi; sansür çözüm yöntemleri; Kaplan-Meier ağırlıkları; *k*NN yerine koyma yöntemi

ABSTRACT Right-censored data is encountered in many areas related to biology, industry, economics, genetics and related fields, primarily clinical trials and health field. The most important characteristic of this data type is that it contains incomplete observations of the person or object of interest. In modeling studies, as incomplete observations result in biased and inconsistent results, several methods have been developed to solve this problem. In this study, many different methods which are used in the literature to be included in the modeling process of right and censored data are examined. Some of these methods can be ordered as follows: Kaplan-Meier weights, Gaussian and kNN imputation methods and synthetic data transformations. Generally, since the censored data points are unknown or can be partially known, classical statistical analysis and modeling methods can be used by assuming some assumptions about the distributions of these observations. In addition, these observations can be estimated using some nonparametric methods which are not based on distribution assumptions. Apart from these two main headings, it is possible to say that the censored data points are removed from the data set even if they are not preferred. In this study, the proposed solution methods for inclusion of right-censored data into regression analysis are presented in stages, from the simplest methods to the most advanced methods, and it is tried to explain how these methods can reflect the effect of censorship. The main purpose of this study is to examine the methods that allow the necessary arrangements for the censorship of the data before the data is added to the regression model. Of course, it is not possible to examine all the existing methods, the most commonly used methods are selected in the literature.

Keywords: Right-censored data; regression analysis; solution methods for censorship; Kaplan-Meier weights; *k*NN imputation Method

Sansürlü veri endüstri, biyoloji, kalite kontrol, risk yönetimi, demografi, teknoloji ve sağlık gibi çok farklı alanlarda karşımıza çıkmaktadır. Sansürlenmiş veri en temel ifadeyle, ilgilenilen kişi veya nesnenin, çalışma süresi boyunca tamamen gözlenmesinin mümkün olmadığı durumlarda ortaya çıkan bir veri türü olarak tanımlanabilir. Dolayısıyla veriyle alakalı sadece kısmi bir bilgi elde edilmiş olur. Literatürde tamamlanmamış gözlemlerden sonuç elde etmek amacıyla sağ-kalım analizi yöntemleri geliştirilmiş ve yaşam süresi verilerinden doğru ve tutarlı sonuçlar elde edilmeye çalışılmıştır. Bu çalışmada sansürlü veri türlerinde sağdan sansürlü verilere ve bu verilerin regresyon analizine doğru şekilde dâhil edilebilmesi için hem hâli hazırda önerilmiş parametrik ve parametrik olmayan çözüm yöntemleri incelenmiş hem de günümüzde popüler hale gelmiş sansürlenmiş gözlemleri tahmin etmeye yarayan yerine koyma yöntemleri incelenmiştir.

Sağdan sansürlenmiş veriler için sansür prosedürünü açıklamak gerekirse; birbirlerinden bağımsız olan yaşam süresi değişkeni Y ve sansür değişkeni C sırasıyla $F(t)=P[Y\leq t]$ ve $G(t)=P[C\leq t]$, $(t\in R)$ dağılımlarına sahip pozitif tanımlı rasgele değişkenler olsun. Sansürleme kuralına göre Y ve C değişkenlerinin değerleri arasından en küçük değere sahip olanlar sansüre göre güncellenmiş yaşam süresi değişkeni $Z=\min(Y,C)$ olarak seçilir. Buna göre sansür bilgisini içeren değişken $\delta=I(Y\leq C)$ şeklinde tanımlanır. Böylece ilgilenilen değişken artık Y yerine (Z,δ) değişkenleri olacaktır. Dolayısıyla yaşam sürelerinin tahmini, sağ-kalım fonksiyonu ve diğer tüm çıkarsamalar belirlenen yeni yaşam süresi değişkenine bağlı olarak yapılır.

Literatürde sağdan sansürlü verilerin tahmin edilmesi için yapılmış çeşitli çalışmalar vardır. Bunlardan önemli olanlardan bazıları şu şekilde sıralanabilir; Kaplan ve Meier, tamamlanmamış gözlemlerin dağılımlarının tahmin edilmesi konusunda ilk ve en önemli çalışmalardan birisini ortaya koymuşlar ve sağdan sansürlü verilerle ilgili gelecekte yapılacak birçok çalışmanın önünü açmışlardır.[1] Kaplan-Meier tahmin edicisi, sağ-kalım fonksiyonunu tahmin etmenin en yaygın tekniklerinden biri haline gelmiştir. Miller, bu tahmin ediciyi kullanarak Kaplan-Meier ağırlıklarını hesaplamış, en küçük kareler yöntemine dâhil etmiş ve doğrusal regresyon yöntemiyle sansürlü verileri tahmin etmiştir.[2] Stute,Kaplan-Meier ağırlıklarının asimptotik özelliklerini ortaya koymuş, merkezi limit teoremini sansür altında incelemiş ve ilk kez doğrusal olmayan modellere uygulayarak çıkarsamalar yapmıştır.[3-5]

Miller ve Halpern önerilen dört önemli yöntemi detaylı olarak karşılaştırmışlardır.[6] Karşılaştırdıkları dört yöntem şunlardır; Cox, Miller, Buckley ve James ve Koul vd. Cox, Cox-oransal yarı-parametrik regresyon modelini geliştirmiştir.[2,7-9] Bu modele göre risk fonksiyonu açıklayıcı değişkenlerin bir fonksiyonudur ve bilinmeyen regresyon katsayıları keyfi ve bilinmeyen bir zaman fonksiyonunu ile çarpılır.

Regresyon analizi söz konusu olduğunda, Buckley ve James, hataların dağılımı bilinmediğinde ve bağımlı değişken sağdan sansürlü olduğunda klasik en küçük kareler yöntemiyle regresyon modelinin tahmin edilmesi için yeni bir tahmin edici önermişlerdir.[8] Koul vd.ise bağımlı değişken sağdan sansürlü olduğunda ve hataların dağılımı bilinmediğinde, doğrusal regresyon modelinin parametre vektörünün tahmin edilmesi için yeni bir tahmin edici önermiştir.[9] Bahsi geçen iki çalışmada da sağdan sansürlü verilerin modele doğru şekilde eklenmesi için çözüm yöntemleri önerilmiştir.

Koul vd.önerdiği çözüm, son zamanlarda literatürde yer alan çalışmalarda en sık kullanılan çözüm yöntemlerinden biridir.[9] Kısaca şöyle açıklanabilir: Sansür değişkeni C'nin dağılımı $G(r)$ Kaplan-Meier tahmin edicisi veya farklı tahmin yöntemi kullanılarak tahmin edilir ve bu tahmin edilen dağılıma bağlı olarak sentetik gözlemler üretilir. Böylece sansürlü yanıt değişkeni yerine üretilen sentetik yanıt değişkeni kullanılarak tahmin yapılır. Bu yöntemin kullanıldığı çalışmalara örnek olarak Wang ve Li, Chen ve Khan, Dikta, Wang ve Zheng, Leurgans, Lai vd., Qin ve Jing gösterilebilir.[10-16]

Sağdan sansürlü veriler ile ilgili yapılan diğer çalışmalar şöyle sıralanabilir: Powell,en çok olabilirlik yöntemine bir alternatif önererek sansür altındaki standart doğrusal model tahmini için genelleştirilmiş en

küçük mutlak sapmalar tahmin edicisini önermiştir.[17] Bu yöntem normal dağılan hata terimleri varsayımına dayalı bir yöntemdir. Fan ve Gijbels regresyon ilişkisinin tamamen veriye göre karar verildiği ve klasik regresyon varsayımları dışında bir varsayımın yapılmak zorunda kalınmadığı bir yöntem önermiştir.[18] Bu yönteme göre sansürlü veri bir dönüşüme tabii tutulduktan sonra yerel ağırlıklandırılmış en küçük kareler regresyon yöntemine uygulanır. Ritov, yarı parametrik regresyon modelini sansür altında tahmin etmek için Buckley-James tipi bir tahmin edici geliştirmiştir.[19] Geliştirilen yöntem, Tsiatis, tarafından önerilen doğrusal rank testine dayanmaktadır ve hataların dağılımı bilindiğinde etkin sonuçlar vermektedir.[20] Breslow ve Crowley, Yaşam çizelgesi ve Kaplan-Meier tahmin edicisiyle sansürlü büyük örneklemler hakkında çıkarım yapmışlardır.[21]

Bu çalışmada regresyon modeline eklenecek sansürlü yanıt değişkeni için gereken düzenlemeler ele alınmıştır. Yapılacak incelemeler sansürün modele dâhil edilmesi için önerilen çözüm yöntemleri üzerine olacaktır. Sağdan sansürlü verilerin analizi için kullanılan ilk yöntemler parametrik regresyona dayanan yaklaşımlardır ve bilindiği üzere bu yöntemler bazı varsayımlara dayanan yaklaşımlardır. Örneğin sağdan sansürlü bağımlı bir değişken için sansür değişkenine ait dağılımın bilinmesi ve sansür için uygulanacak çözümlerin belirlenen dağılıma olarak gerçekleştirilmesi gibi. Buna örnek olarak sağ kalım fonksiyonunun parametrik bir dağılıma uyduğu varsayılarak yapılan regresyon analizleri örnek verilebilir. Elbette gerçek hayatta, değişkenler hakkında elde edilecek bilgiler oldukça sınırlıdır ve çoğunlukla karşılaşılan durumlarda değişkenler arasındaki ilişki belirsizdir ve sansür değişkenine ait dağılım hakkında kesin bir çıkarım yapmak oldukça zordur. Dolayısıyla, genellikle varsayımların ihlal edilmesi söz konusudur ki bu da yapılan tahminin güvenilirliğini ortadan kaldırır ve tahmin iyi bir tahmin olmaktan çıkar. Varsayımlar ihlal edilmediği takdirde ise, verinin varsayımlara uydurulması gerekir ki bu da veriden alınacak bilgiyi eksiltecek veya bozacaktır. Bu nedenle parametrik yöntemler yalnızca belirli şartları sağlayan veriler için uygulanabilir denilebilir. Buna alternatif olarak literatürde sıkı varsayımlardan bağımsız olan parametrik olmayan yöntemler önerilmiştir.

İkinci bölümde sağdan sansürlü verinin yapısı hakkında detaylı bilgi sunulmuştur. Üçüncü bölümde basit doğrusal regresyon modeli üzerinde çözüm yöntemlerinin nasıl uygulandıkları gösterilmiştir. Dördüncü bölümde çalışmadan elde edilmiş bilgiler derlenmiş, öneri ve sonuçlar sunulmuştur.

SAĞDAN SANSÜRLÜ VERİ

Sağdan sansürlü verilerin en sık kullanıldığı ve doğru analiz edilmesinin kritik öneme sahip olduğu alan sağlık alanı olduğundan çalışmanın geri kalanında, yaşam süreleri üzerinden açıklamalar yapılmıştır. Sağdan sansürlü verinin açıkça ifade edilebilmesi içinde klinik bir deney yapıldığı varsayılsın. Bu deneyde yer alan "n" adet hastanın yaşam süreleri takip ediliyor olsun. Hastaların gözlenmesi aşamasında, üç temel sebepten gözlemler sansürlenir;

(i) Hastanın çalışma süresince beklenen başarı durumunu göstermemesi (ölüm veya hastalık belirtisi vb.)

(ii) Hastanın bir süre sonra başka bir sebepten çalışmayı bırakıp gitmesi

(iii) Gözlem yapan kişinin hastanın izini kaybetmesi durumu ve artık hastanın gözlenemiyor olması.

Bu durumlarda çalışmada bulunan bazı yaşam süreleri hakkında yalnızca kısmi bir bilgi edinilebilir. Çalışmanın gerçek bitiş sürelerini "C" değişkeni temsil etsin. Bazı hastalarda çalışma sonuna kadar ilgilenilen olay gerçekleşmezse, en kötü ihtimalle o hastanın yaşam süresi C_i olacaktır. Bu durumda i'inci hastanın hakkında eksik bilgiye sahip olunur ve Y_i gözleminin C_i tarafından sansürlendiği söylenir. Bu durum tüm gözlemler için uygulanabilir. Dolayısıyla her bir kişi veya nesne için C_i değerleri ile Y_i değerleri arasından en küçük olanının seçilmesiyle yaşam sürelerini sansür durumlarıyla birlikte içeren yeni bir yanıt değiş-

keni ve sansür bilgisini içeren çift gözlemli değişken elde edilir ve çıkarsamalar elde edilen yeni değişkenler üzerinden yapılır;

$$Z_i = \min(Y_i, C_i) \; ve \; \delta_i = I(Y_i \leq C_i) \quad (1)$$

Böylece sansürü taşıyan gözlem çiftleri elde edilecektir. Burada δ_i'ler sansür bilgisini taşıyan gösterge fonksiyonudur. Sansür varsa "0" yoksa "1" değerini alır.

Sağdan sansürlü verilerin tahmin edilmesi ile ilgili yapılacak çalışmaların neredeyse hepsinde yapılacak çıkarsamaların doğruluğunun ve tutarlılığının sağlanması için yanıt değişkeni, sansür değişkeni ve açıklayıcı değişkenler için geçerli olması beklenen bazı özel varsayımlar belirlenmiştir. Bu varsayımlar;

V1. Y_i ve C_i'ler birbirinden bağımsızdır.

V2. $\boldsymbol{P}(Y_i \leq C_i | Y_i, X_i) = \boldsymbol{P}(Y_i \leq C_i | Y_i)$.

Burada X_i'ler regresyon modelinde kullanılan açıklayıcı değişken değerlerini temsil eder. Bu varsayımlar sağ-kalım analizlerinde kabul edilen genel varsayımlardır. İlk varsayım, Kaplan-Meier tahmin edicisi kullanıldığında sağdan sansürlü veri için kurulan modelin tanımlanabilirliğini sağlayan bağımsızlık varsayımıdır. Bu varsayım bozulduğunda düzgün bir model elde etmek için sansürlü veri hakkında daha fazla bilgi sahibi olunması gerekmektedir. İkinci varsayım ise; verilen olay (ölüm, belirti vb.) zamanı için açıklayıcı değişkenler veri sansürlü olsa da olmasa da elde edilenden daha fazla bilgi sağlayamayacaktır. Varsayımlar hakkında detaylı bilgi ve ispatlar için Stute, çalışmaları incelenebilir.[3-5] Ayrıca sansür varsayımlarına ilişkin diğer çalışmalar Tsiatis, Wei vd. ve Zhou olarak sayılabilir.[20,22,23]

SANSÜR İÇİN ÇÖZÜM YÖNTEMLERİ

Sağdan sansürlü yanıt değişkeninin regresyon modeline uygun şekilde eklenebilmesi için önerilen çeşitli yöntemlerin, model tahmini işleyişindeki durumunu görebilmek için basit doğrusal modeli kullanılmıştır. Buna göre doğrusal regresyon modeli aşağıdaki gibidir:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_p X_{ip} + \varepsilon_i, \; i = 1,2,\ldots,n \quad (2)$$

Burada Y_i'ler sansürlü yanıt değişkeninin değerlerini, $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \ldots, \beta_p)^T$ bilinmeyen regresyon katsayılarını, $\mathbf{X} = (X_{i1}, X_{i2}, \ldots, X_{ip})$ tamamı gözlenmiş açıklayıcı değişkenleri ve ε_i'ler rasgele hata terimlerini göstermektedir. Eşitlik (1)'de gösterildiği gibi yanıt değişkeni sansüre göre güncellenirse model aşağıdaki gibi tekrar yazılır:

$$Z_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_p X_{ip} + \varepsilon_i, \; i = 1,2,\ldots,n \quad (3)$$

Ve bu modelin matris ve vektör formu eşitlik (4)'te gibi gösterilir

$$\mathbf{Z} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad (4)$$

Tahmin edilmesi gereken model eşitlik (4)'te verilen modeldir. Regresyon modelinin sansürün etkisini içeren şekilde tahmin edilmesi için geçmişten bugüne önerilmiş çeşitli yöntemler vardır. Bu çalışmada bu yöntemlerden literatürde sıkça kullanılanları incelenmiştir. Bu yöntemler şu şekilde sıralanabilir: En ilkel yöntem olarak **Z** sansürlü gözlemlerden arındırılır (sansürlü gözlemler atılır) ve regresyon modeli kalan verilerle tahmin edilir. Bu yöntem, günümüzde artık kullanılmamakla birlikte, hem bilgi kaybına neden olduğu için hem de yanlı sonuçlara neden olduğu için bu çalışmada yalnızca kısaca bahsedilmiştir.

Yukarıda açıklandığı üzere, sansürlü yanıt değişkeninin tahmin prosedürüne uygun hale getirmek için önerilen yöntemler iki ana başlık altında toplanmaktadır: Parametrik ve parametrik olmayan çözüm yöntemleri. Çalışmamızda her iki durum da incelenmiştir.

PARAMETRİK ÇÖZÜM YÖNTEMLERİ

Ağırlıklı En Küçük Kareler Yöntemi

Sansürlü verilerin regresyon modeli ile tahmininde literatürde genellikle sentetik veri dönüşümleri kullanılmaktadır. Sentetik veri ile elde edilen tahminlerin varyanslarının doğru biçimde hesaplanabilmesi hatalarda değişen varyansın olmaması gerekmektedir. Genellikle sansürlü doğrusal modellerde değişen varyansa rastlanmaktadır. Bu durumu düzeltilmesi için ve tahmini daha etkin ve tutarlı hale getirebilmek için Miller sağdan sansürlü veriler için ağırlıklı en küçük kareler yöntemini önermiştir.[2] Burada şu belirtilmelidir ki ağırlıklı en küçük kareler yönteminde sansürün etkisi modele ağırlıklar yoluyla aktarılmaktadır ve ağırlıklar da sansür değişkeninin (C_i) dağılımı (G) tahmin edilerek veya biliniyorsa doğrudan kullanılarak hesaplanır. Dolayısıyla bu yöntem hem parametrik hem de parametrik olmayan yöntem olarak kullanılabilir. Bu durum sağ-kalım fonksiyonunun nasıl tahmin edildiğine bağlıdır.

Bu bölümde, sağ-kalım fonksiyonunun parametrik olarak tahmin edilmesi ele alınmıştır. Buna göre, sansür değişkeninin dağılımı için kullanılabilecek yöntemler ve fonksiyon tahminleri aşağıdaki tabloda gösterilmiştir;

TABLO 1: Ağırlıklı en küçük kareler yönteminde ağırlıkların belirlenmesi için kullanılabilecek parametrik dağılımlar.

Dağılım adı	Olasılık dağılımı
Weibull Dağılımı	$G(Y_i;\theta,p)=1-e^{\left\{-\left(\frac{Y_i}{\theta}\right)^p\right\}}$ $p > 0$ şekil parametresi $\theta > 0$ ölçek parametresi
Üstel Dağılım	$G(Y_i;\theta)=1-e^{-\theta Y_i}$ $\theta>0$ oran parametresi
Genelleştirilmiş Gama Dağılımı	$G(Y_i;d,p,\alpha)=\frac{\gamma\left(\frac{d}{p},\left(\frac{Y_i}{\alpha}\right)^p\right)}{\Gamma d/p}$ $\alpha > 0$ ölçek parametresi $d > 0$ ve $p>0$ şekil parametresi Eğer $d=p$ ise dağıım Weibull dağılımına yaklaşır. $\gamma(.)$ düşük tamamlanmamış gama dağılımı ve $\Gamma(.)$bilinen gama dağılımıdır.
Log-Normal Dağılım	$G(Y_i;\mu,\theta)=\Phi\left(\frac{\log(Y_i)-\mu}{\theta}\right)$ $\mu \geq 0$ şekil parametresi $\theta > 0$ ölçek parametresi $\Phi(.)$ Standart normal dağılıma ait dağılım fonksiyonu
Log-Lojistik Dağılım	$G(Y_i;\beta,\gamma)=\frac{1}{(1+(Y_i/\gamma)^{-\beta})}$ $\beta > 0$ şekil parametresi $\gamma > 0$ ölçek parametresi

Ağırlıklı en küçük kareler yöntemi ile model (4) tahmini aşağıdaki gibi yazılabilir;

$$\Sigma_{i=1}^{n} W_i(Z_i - \mathbf{X}_i\boldsymbol{\beta})^2,\ j = 1,\ldots,p \tag{5}$$

Burada W_i değerleri Tablo 1'de verilen dağılımlara göre aşağıdaki gibi hesaplanabilir;

$$W_i = \frac{\delta_{(i)}}{n-i+1}\left[1-G(Z_{(i)})\right]^{\delta(i)} \tag{6}$$

Burada $Z_{(i)}$'ler küçükten büyüğe sıralanmış yanıt değişkeni değerlerini göstermekte ve $\delta_{(i)}$'ler ise $Z_{(i)}$ değerleriyle ilişkili olarak sıralanmış değerlerdir. Ayrıca $G(Z_{(i)})$, Tablo 1'de verilmiş dağılım fonksiyonlardan veriyle ilişkili olarak belirlenen dağılım fonksiyonudur. Bu bağlamda, ağırlıklı en küçük kareler yardımıyla model parametrelerinin tahmini aşağıdaki gibi elde edilir;

$$\hat{\boldsymbol{\beta}}_W = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z} \tag{7}$$

Elde edilen bu tahmin edicinin varyansı ve uyum değerleri sırasıyla aşağıdaki gibi gösterilebilir;

$$Var(\hat{\boldsymbol{\beta}}_W) = \sigma^2(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1},\quad \hat{\sigma}^2 = \frac{||\mathbf{Z}-\mathbf{X}\hat{\boldsymbol{\beta}}W||^2}{n-p} = \frac{(\mathbf{Z}-\hat{\mathbf{Z}})^T(\mathbf{Z}-\hat{\mathbf{Z}})}{n-\text{zvzvvvvt}} \tag{8}$$

ve

$$\boldsymbol{E}(\mathbf{Z}|\mathbf{X}) = \hat{\mathbf{Z}} = \mathbf{X}\hat{\boldsymbol{\beta}}_W \tag{9}$$

Gaussian Yerine Koyma Yöntemi

Bu yöntem, ilk olarak Park vd. tarafından sansürlü zaman serilerinin tahmin edilmesi için kullanılmıştır.[24] Yanıt değişkeninin normal dağılıma sahip olduğu durumda, sansürlü gözlemlerin her birini tahmin etmeye yarayan ve eğer dağılım koşulu sağlanırsa başarılı sonuçlar veren bir yöntem olduğu söylenebilir. Model (4)'ten farklı olarak, basit bir otoregresif zaman serisi üzerinde kısaca nasıl çalıştığı gösterilmiştir. Otoregresif model aşağıdaki gibi yazılır;

$$Y_t = \phi_1 Y_{t-1} + \ldots + \phi_q Y_{t-q} + e_t,\ t = 1,2,\ldots,n \tag{10}$$

Burada Y_t kendi gecikmelerine bağlı durağan bir zaman serisini temsil etmektedir. e_t'ler sıfır ortalamalı ve sabit varyanslı rasgele hata terimleridir ve aşağıdaki gibi yazılır;

$$e_t = \theta_1 e_{t-1} + \ldots + \theta_p e_{t-p} + u_t \tag{11}$$

Burada $u_t \sim N(0,\sigma^2)$ dağılımına sahip hata terimleridir. Sansürlü yanıt değişken, eşitlik (1)'deki gibi sansüre göre güncellendikten sonra model tahmin yapılacağından ve model yeniden yazılır;

$$Z_t = \phi_1 Z_{t-1} + \ldots + \phi_q Z_{t-q} + e_t,\ t = 1,2,\ldots,n \tag{12}$$

Gaussian yerine koyma yönteminin uygulanması prosedürü için gerekli bazı notasyonal tanımlar yapılmalıdır. Buna göre sıfır ortalamalı ve durağan e_t hata terimlerinin $n\times n$ kovaryans matrisine ($\boldsymbol{\Sigma}$) sahip olduğu varsayılsın. Buna göre $Y_t \sim N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ dağılımına sahip olduğu söylenebilir. Eşitlik (1)'deki değişimden sonra ise $Y_t \sim TN_n(\boldsymbol{\mu},\boldsymbol{\Sigma};R_S)$ budanmış dağılımına sahip olduğu söylenir ki burada $TN_n(.;R_S)$, R_c aralığında budanmış normal dağılım anlamına gelir. Burada R_c, verilerin sansürlendiği değere göre değişiklik gösterir ve sansür değişkenine bağlı olarak $R_S = (0, C_t)$ olduğunda $\delta_t = 1$ ($gözlenmiş$) $R_S = [C_t,\infty)$ olduğunda ise $\delta_t = 0$

(*sansürlü*), şeklinde tanımlanır. Bu yöntemi uygulamadan önce, verilerin sansürlü ve gözlenmiş olarak ayrı ayrı ifade edilmesi gerekir. Buna göre verilerin tekrar düzenlenmesi için permütasyon matrisi P aşağıdaki gibi ifade edilir;

$$\mathbf{P}Y_t = \begin{pmatrix}\mathbf{P}_G\\ \mathbf{P}_S\end{pmatrix} Y_t = \begin{pmatrix}\mathbf{Y}_G\\ \mathbf{Y}_S\end{pmatrix} \tag{13}$$

Burada görüldüğü üzere $\mathbf{Y}_G$ gözlenen, $\mathbf{Y}_S$ sansürlenen kısmı temsil eder. Bu matrislere bağlı olarak elde edilecek olan çok değişkenli normal dağılım ise şöyle gösterilebilir;

$$PY_t \sim N_n\left(\boldsymbol{\mu}\begin{pmatrix}P_G\\ P_S\end{pmatrix}, \begin{bmatrix} P_G\boldsymbol{\Sigma}P_G' & P_G\boldsymbol{\Sigma}P_S' \\ P_S\boldsymbol{\Sigma}P_G' & P_S\boldsymbol{\Sigma}P_S' \end{bmatrix}\right) = N_n\left(\begin{pmatrix}\boldsymbol{\mu}_G\\ \boldsymbol{\mu}_S\end{pmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{GG} & \boldsymbol{\Sigma}_{GS} \\ \boldsymbol{\Sigma}_{SG} & \boldsymbol{\Sigma}_{SS} \end{bmatrix}\right) \tag{14}$$

Eşitlik (13) benzer şekilde gözlenebilen ve sansüre göre düzenlenmiş olan Z_t'ler için aynı işleyiş gerçekleştirilir;

$$\mathbf{P}Z_t = \begin{pmatrix}\mathbf{P}_G\\ \mathbf{P}_S\end{pmatrix} Z_t = \begin{pmatrix}\boldsymbol{Z}_G\\ \boldsymbol{Z}_S\end{pmatrix} \tag{15}$$

Böylece Park vd. önerdiği şekilde budanmış çok değişkenli normal dağılım yazılabilir;[24]

$$(\boldsymbol{Y}_S | \boldsymbol{Z}_G, \boldsymbol{Z}_S \in R_S) \sim TNn_S(\mathbf{M}, \boldsymbol{V}, R_S) \tag{16}$$

Burada n_S sansürlü verilerin sayısını göstermektedir ve TNn_S n_S boyutlu budanmış normal dağılımı temsil eder. R_S daha önce tanımlandığı gibi budama aralığını belirtir. **M** ve **V** sırasıyla ilgili aralıkta budanmamış kısımdan elde edilmiş koşullu ortalama ve varyansı temsil eder. Detaylı bilgi algoritmada verilmiştir.

Gaussian yerine koyma algoritması

Adım 1. Giriş parametreleri $\hat{\mu}^{(0)}$, $\hat{\rho}^{(0)}$ ve $\hat{\sigma}^{(0)}$ aşağıda gösterildiği gibi hesaplanır;

$\hat{\mu}^{(0)} = n^{-1}\Sigma_{i=1}^{n} Z_t^{(0)}$,

$\hat{\rho}^{(0)} = (\Sigma_{t=2}^{n}(Z_{t-1}^{(0)} - \bar{\bar{Z}}_{-n}^{(0)})^2)^{-1}\,(\Sigma_{t=2}\,(Z_t^{(0)} - \bar{\bar{Z}}_{-1}^{(0)})\,(Z_{t-1}^{(0)} - \bar{\bar{Z}}_{-n}^{(0)}))$,

$(\hat{\sigma}^{(0)})^2 = (n-3)^{-1}\sum_{t=2}^{n}\,(Z_t^{(0)} - \hat{\mu}^{(0)} - \hat{\rho}^{(0)}(Z_{t-1}^{(0)} - \hat{\mu}^{(0)}))^2$

Burada $\hat{\mu}^{(0)}$ notasyonu, sıfır iterasyon için tahminleri ifade eder. Her bir parametresi için aynıdır. Bazı hesaplamalar: $Z_{-n}^{(0)} = (n-1)^{-1}\Sigma_{t=1}^{n-1} Z_t$, $Z_{-1}^{(0)} = (n-1)^{-1}\Sigma_{t=2}^{n} Z_t$. Buna bağlı olarak ilk parametre vektörü $(\hat{\boldsymbol{\mu}}^{(0)}, \hat{\boldsymbol{\Sigma}}^{(0)})$ oluşturulur.

Adım 2. Budanmış normal dağılım için koşullu ortalama $\hat{\mathbf{M}}^{(0)}$ ve varyans $\hat{\boldsymbol{V}}^{(0)}$ hesaplanır ve $\boldsymbol{\Phi} = (\mathbf{M}, \boldsymbol{V})$ oluşturulur;

$$\hat{\mathbf{M}}^{(0)} = \hat{\mu}_S^{(0)} + \hat{\Sigma}_{SG}^{(0)}\,(\hat{\Sigma}_{GG}^{(0)})^{-1}(\boldsymbol{Z}_G - \mu_G^{(0)}),$$

$$\hat{\boldsymbol{V}}^{(0)} = \hat{\Sigma}_{SS}^{(0)} - \hat{\Sigma}_{SG}^{(0)}\,(\hat{\Sigma}_{GG}^{(0)})^{-1}\,\hat{\Sigma}_{GS}^{(0)}$$

Adım 3. $TNn_S(\hat{\mathbf{M}}^{(0)}, \hat{\boldsymbol{V}}^{(0)}, R_S)$ budanmış normal dağılımı kullanılarak Sağdan sansürlü gözlemlerin yerine kullanılacak olan $\boldsymbol{Z}_S^{(1)}$ vektörü oluşturulur.

Adım 4. Eşitlik (15)'te gösterildiği şekilde yanıt değişkeni yeni üretilen sansürlü gözlemlerle tekrar oluşturulur:

$$\mathbf{Z}^{(1)} = P^{-1}\begin{bmatrix} Z_G \\ Z_S^{(1)} \end{bmatrix}$$

Adım 5. Yeni yanıt değişkeni kullanılarak μ, σ^2, ve ρ tekrar tahmin edilir ve koşullu ortalama ve varyanslar tekrar bulunır ve $\mathbf{\Phi}$ yeniden oluşturulur.

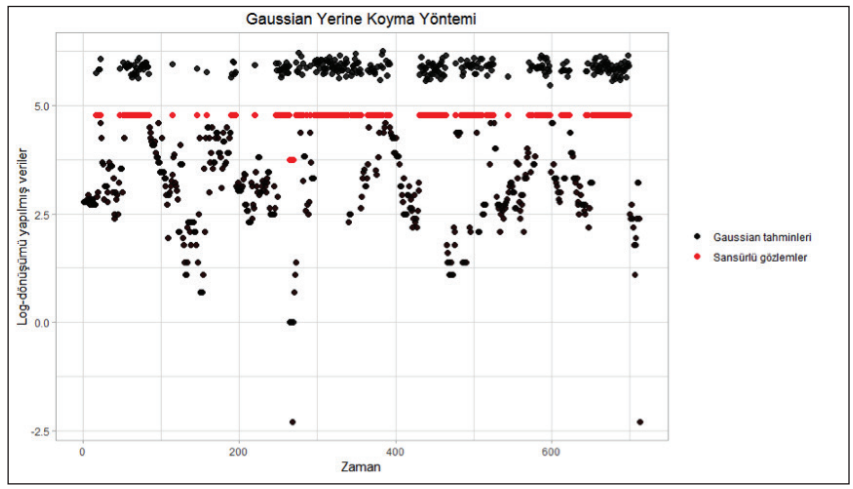
Adım 6. Adım 2-6, belirlenen yakınsama ψ<0.005 gerçekleşene dek tekrar edilir. Burada ψ k'ıncı iterasyon için aşağıdaki gibi hesaplanır;

$$\psi = \left[(\hat{\boldsymbol{\tau}}^{(k+1)} - \hat{\boldsymbol{\tau}}^{(k)})'(\hat{\boldsymbol{\tau}}^{(k+1)} - \hat{\boldsymbol{\tau}}^{(k)})\right] / (\hat{\boldsymbol{\tau}}^{(k)})'\hat{\boldsymbol{\tau}}^{(k)}$$

Burada $\hat{\boldsymbol{\tau}} = (\mu, \sigma^2, \rho)'$.

Yakınsama gerçekleştikten sonra elde edilen $\mathbf{Z}^{(k)}$, model (12) tahmininde kullanılacak olan ve sansürlü gözlemleri tamamlanmış olan yanıt değişkenidir. Model (12)'nin tahmini için kullanılacak yöntemler başka bir çalışmanın konusu olduğundan burada bahsedilmemiştir.

Örnek (Gaussian Yerine Koyma Yöntemi): Gaussian yerine koyma yönteminin nasıl çalıştığına ilişkin detaylı bilgi sağlamak için Şekil 1'de Ulusal atmosferik araştırma merkezi (NCAR) tarafından San Francisco eyaletinde 1989 yılının Mart ayı boyunca saatlik olarak toplanan sağdan sansürlü bulut yüksekliği verileri gösterilmiştir. Veri seti, 716 gözlemden oluşmaktadır ve orijinal veriler 100 fitlik ölçekteyken burada sürekliliğin sağlanabilmesi için log dönüşümü uygulanmıştır. Bu veride ölçüm sınırı 12.000 fit (log dönüşümlü veride 4.791) olduğundan bu değerin üzerindeki veriler gözlenememiştir. Veri setinden 293 veri sağdan sansürlü yani sansür yüzdesi %41.62'dir. Bu da verinin ağır sansür altında olduğunu gösterir. Aşağıdaki şekilde hem tamamlanmamış gözlemler (ölçüm sınırı nedeniyle) hem de Gaussian yerine koyma yöntemi ile tahmin edilmiş gözlemler gösterilmiştir. Yöntemin, sansürlü gözlemler için elde ettiği tahmin değerlerin, ağır sansür altındaki bir veri seti için tatmin edici olduğu söylenebilir.



ŞEKİL 1: Gaussian yerine koyma yöntemi ile tamamlanan bulut yüksekliği verileri.

PARAMETRİK OLMAYAN ÇÖZÜM YÖNTEMLERİ

Kaplan-Meier Ağırlıkları

İlk defa Miller tarafından önerilen bu yöntem, yukarıda bahsedilen ağırlıklı en küçük kareler yönteminde ağırlıkların parametrik dağılımlar ile değil, parametrik olmayan bir yöntem olan Kaplan ve Meier tarafından önerilen Kaplan-Meier tahmin edicisi ile elde edilmesi ile elde edilen yöntemdir.[1,2]

Buna göre Kaplan-Meier tahmin edicisine bağlı olarak ağırlıklar aşağıdaki gibi elde edilir;

$$W_i^{km} = \frac{\delta_{(i)}}{n-i+1} \Pi_{j=1}^{i-1} \left[\frac{n-j}{n-j+1}\right]^{\delta_{(i)}} \quad (17)$$

Buna bağlı olarak model (4)'ün tahmini eşitlik (7)'ye benzer şekilde şöyle yazılabilir;

$$\hat{\boldsymbol{\beta}}_{W^{km}} = (\mathbf{X}'\mathbf{W}^{km}\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z} \quad (18)$$

Bu yöntem, daha önce de belirtildiği gibi verilerin dağılımı hakkında herhangi bir bilgi yoksa parametrik yöntemlere göre daha anlamlı sonuçlar vermektedir. Ayrıca, burada hesaplanan Kaplan-Meier ağırlıkları ile ilgili Stute ve Wang detaylı olarak çalışmış ve bu ağırlıkların, sansürlü durumlarda büyük sayılar kanunu ile ilgili ilişkisini göstermiştir.[25] Buna ek olarak Miller,bağımlı değişkenin sağdan sansürlü olması durumunda regresyon katsayılarının $(\beta_0,\beta_1,\ldots,\beta_p)$, $n\to\infty$ olduğundan gerçek regresyon katsayılarına yakınsadığını göstermiştir.[2]

Örnek(Kaplan Meier Ağırlıkları): Kaplan-Meier ağırlıklarının tahmin performansını ölçmek için küçük bir simülasyon çalışması yapılmıştır. Bu çalışmada, doğrusal bir regresyon modeli aşağıdaki gibi üretilmiştir;

$$y_i = 1.5 + 3x_i+\varepsilon_i,\ i=1,\ldots,n \quad (19)$$

Burada y_i'ler tamamı gözlenmiş yanıt değerleri, $\boldsymbol{\beta} = (1.5, 3)^T$, $x_i \sim U[0,1]$ ve $\varepsilon_i \sim NIID(0,\sigma^2=1)$ olarak elde edilmiştir. $NIID$,normal, bağımsız ve aynı dağılımlı anlamına gelmektedir. Elde edilen yanıt değişkeninin sağdan sansürlenmesi için sansür değişkeni değerleri c_i ve sansür bilgisini içeren δ aşağıdaki gibi elde edilir;

$$\delta_i \sim bernoulli(CL) ve\ c_i \sim N(\mu_y,\sigma_y) I(c_i > y_i,\ \delta_i = 0)$$

Burada CL, sansür seviyesini gösterir ve bu çalışma için $CL = 0.20$ olarak belirlenmiştir. Buna göre eşitlik (1)'de gösterilen yeni yanıt değişken değerleri Z_i'ler elde edilir. Eşitlik (18)'de gösterilen Kaplan-Meier ağırlıklarıyla tahmin yapılabilir. Burada, örneklem büyüklüğü (n) 50 olarak belirlenmiş ve çalışma 1000 kez tekrar edilmiştir. Buna göre Tablo 2'de, $\hat{\beta}_{W^{km}}$ ve Kaplan-Meier ağırlıkları kullanılmadan doğrudan Z_i'ler ile yapılan tahminler $\hat{\beta}_{EKK}$ gösterilmiştir. Ayrıca yan ve varyanslar sunulmuştur.

Tablo 2'de elde edilen sonuçlara göre, Kaplan-Meier ağırlıklarıyle elde edilen tahminlerin nispeten klasik en küçük kareler yönteminden daha iyi olduğunu göstermektedir. Bu da Kaplan-Meier ağırlıklarının sansürün etkisini yanıt gözlemlerine daha iyi yansıttığını ispatlar niteliktedir. Daha detaylı bilgi ve benzer konuda daha geniş bir uygulama çalışması için Yilmaz and Aydın incelenebilir.[26]

TABLO 2: Kaplan-Meier ağırlıkları ile elde edilen model çıktıları.

	Tahmin değeri	Yan	Varyans
$\hat{\beta}_{W^{km}}$	[1.1705; **2.5622**]	[0.3294; **0.4377**]	[**0.0815**; **0.2540**]
$\hat{\beta}_{EKK}$	[**1.1823**; 2.4397]	[**0.3176**; 0.5602]	[0.0831; 0.2592]

Sentetik Veri Dönüşümleri

Önceki bölümlerde belirtildiği gibi, sansür söz konusu olduğunda ve verilerin dağılımı hakkında bilgi yoksa hem parametrik hem de parametrik olmayan yöntemler doğrudan uygulanamamaktadır. Sansürlü verilerin analizinde ana düşünce yanıt gözlemleri için uygun bir düzeltme ve dönüşüm yapmaktır. Bu konuda önerilmiş önemli birkaç yöntem vardır. Bu yöntemler genel olarak sentetik veri dönüşümü olarak ifade edilmektedir. Bunlardan literatürde en yaygın olanları Koul vd., Buckley ve James, Leurgans olarak gösterilebilir.[8,9,14] Bahsedilen çalışmaların hepsi sağdan sansürlü regresyon modellerinde, regresyon fonksiyonunun doğrusal olduğu durumları göz önüne almıştır. Regresyon fonksiyonunun şeklinin belirsiz olduğu durumlar için Fan ve Gijbels iki tip dönüşüm önermiştir.[18] Bunlar, yerel ortalama dönüşümü ve yeni sınıf (NC-new class) dönüşümü olarak bilinmektedir. Bu dönüşümler, Buckley ve James önerisine benzer biçimde fakat doğrusal olmama durumları için önerilmiştir.[8]

(i) Buckley-James Dönüşümü

Buckley ve James, bağımlı değişken sağdan sansürlü olduğunda, doğrusal regresyon model tahmini için Miller'dan biraz daha farklı bir yöntem önermiştir.[2,8] Bu yöntemde hatalar, belirli bir dağılıma sahip olmak zorunda değildirler. Yöntemde, en küçük karelerden elde edilen normal denklemler düzenlenerek Miller yöntemindeki tutarsızlıklar giderilmeye çalışılmıştır. Buckley ve James,çalışmalarında en küçük kareler yönteminden elde edilen normal denklemlerin üzerinde düzeltmeler yapılarak elde edilen regresyon yöntemlerine (Medyan regresyonu veya Rank regresyonu) benzer biçimde bu düzenlemeyi sansürlü gözlemlere uygun biçimde elde etmişlerdir.[8] Burada bağımlı değişkendeki sansürlenmeyi hesaba katabilmek için yapılan dönüşüm aşağıdaki gibidir;

$$Z_i^* = Z_i\delta_i + E(Z_i|Z_i>C_i)(1-\delta_i),\ \ i = 1,\ldots,n \tag{20}$$

Burada Z_i^* üretilen sentetik yanıt değişkeni, δ_i sansür işaret değişkenidir. Böylece, sansürlü yanıt değişkeni yerine elde edilen Z_i^* kullanılabilir denilir. Burada belirtmek gerekir ki eşitlik (20) da verilen beklenen değerin hesaplanabilmesi için kullanılacak dağılım fonksiyonu genellikle bilinmediğinde Kaplan-Meier tahmin edicisi ile elde edilir. Böylece regresyon katsayıları aşağıdaki gibi elde edilebilir;

$$\hat{\boldsymbol{\beta}}^* = (\Sigma_i^n Z_i(X_i-\bar{X}) + \Sigma_i^n Z_i^*(\hat{\boldsymbol{\beta}})(X_i-\bar{X}))/\Sigma_i^n X_i-\bar{X})^2 \tag{21}$$

Burada $Z_i^*(\hat{\boldsymbol{\beta}})$ aşağıdaki gibidir;

$$Z_i^*(\hat{\boldsymbol{\beta}}) = \hat{\boldsymbol{\beta}}X_i + \sum W_i^{km}\hat{\boldsymbol{\beta}}\ (Z_i-\hat{\boldsymbol{\beta}}X_i)$$

Burada W_i^{km} yukarıda tanıtılan Kaplan-Meier ağırlıklarıdır.

(ii) Leurgans Dönüşümü

Leurgans dönüşümü doğrusal regresyon modellerinde rasgele sağdan sansürlü gözlemlerin hesaba katılması için önerilmiş bir dönüşümdür. Bu dönüşüm için, klasik sansür varsayımları geçerlidir. Burada sansür değişkeni C_i, H dağılımına sahip olsun ve sansür değişkeni ile gerçek yaşam süreleri birbirinden bağımsız olsun (klasik sansür varsayımı). Klasik en küçük kareler yöntemi hata kareler ortalaması (HKO) ile gösterildiğinde aşağıdaki gibi gösterilir;

$$HKO(Z_i) = n^{-1}\Sigma_i^n(Z_i-\hat{\boldsymbol{\beta}}^T X_i)^2 \tag{22}$$

Yukarıdaki ifade, sansür olmadığında kullanılmaktadır. Bağımlı değişken üzerinde rasgele sağdan sansür söz konusu olduğunda, HKO yine hesaplanabilmektedir. Bu hesaplama için, sansür değişkeninin dağılımı veya bu dağılımının tahmini kullanılır. Yaşam süreleri ve sansür verileri birbirinden bağımsız olduğunda;

$$1\text{–}HKO(Z_i) = n^{-1}\Sigma_i^n (Z_i - \hat{f}_i)^2 \quad (23)$$

Burada $\hat{f}_i(r) = 1 + \Sigma_{j=1}^{i-1} w_j^L I(Z_j \le r) - A_i(Z_j < r)$ ile ifade edilir ve burada

$$w_j^L = \left[\frac{1}{1\text{–}\hat{H}(Z_j)} - 1\right] \Big/ \left[\frac{1}{1\text{–}\hat{H}(Z_j)}\right] \text{ ve } A_i = \frac{1}{1\text{–}\hat{H}(Z_j)}$$

Böylece sentetik veri noktaları $Z_i^* = \Sigma_i^n (Z_j - Z_{j-1}) A_i$ şeklinde hesaplanır ve en küçük kareler yöntemi ile model (4) tahmini gerçekleştirilebilir. Detaylı bilgi için Leurgans incelenebilir.[14]

(iii) Koul-Susarla-Van Ryzin (KSV) Dönüşümü

Literatürde en sık kullanılan sentetik veri dönüşümüdür. Daha önce bahsedildiği gibi Miller Kaplan-Meier ağırlıklarını kullanarak elde edilen en küçük kareler yöntemiyle doğrusal regresyon modeli tahmini yapmıştır.[2] Daha sonra Buckley ve James ortalamaya dayalı olarak başka bir tahmin edici geliştirmiştir.[8] Bahsedilen her iki tahmin edici de iterasyona dayalıdır ve asimptotik özelliklerine dair sezgisel bazı şeyler söylenmiş fakat matematiksel olarak desteklenememiştir. Koul vd., bu iki tahmin ediciden farklı olarak hem iterasyona ihtiyaç duymayan hem de asimptotik dağılım teorisi matematiksel olarak ispatlanmış bir sentetik veri dönüşümü ile tahmin yöntemi önermişlerdir.[9] Ayrıca belirtmek gerekir ki, bu tahmin edicinin hesaplanması ve teorisi kolaylıkla çok değişkenli modeller için esnetilebilmektedir.

Koul vd.tarafından önerilmiş olan bu dönüşüme dayalı olarak regresyon modelinin tahmin edilmesi için gereken varsayımlar şöyle sıralanabilir;[9]

A1. Model (4)'te ifade edilmiş olan hata terimleri $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ sıfır ortalamalı ve aynı dağılıma sahip hata terimleridir.

A2. Regresyon modelinde yalnızca yanıt değişkeninin sağdan sansürlü olması gerekir.

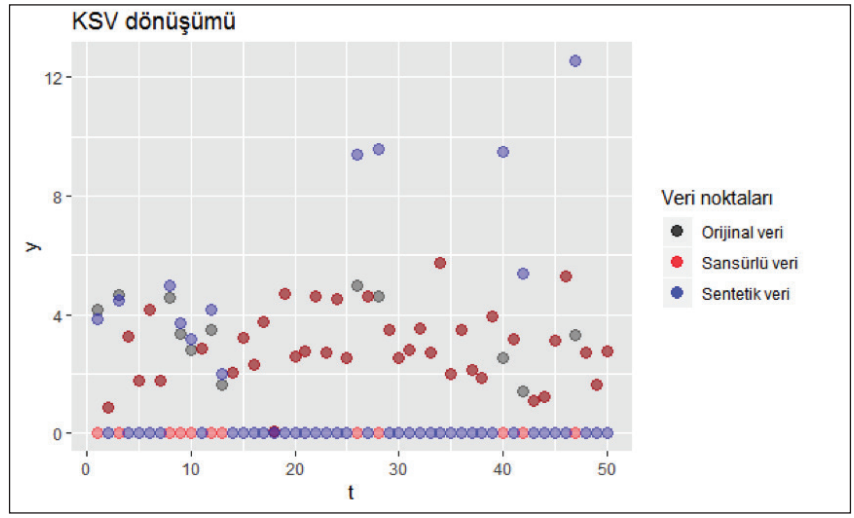
A3. Sansür değerleri C_i'ler ve hata terimleri ε_i'ler bağımsız olmalıdır.

Buna göre sentetik veri dönüşümü aşağıdaki gibi yazılabilir;

$$Z_i^* = \frac{\delta_i Z_i}{1\text{–}\hat{G}(Z_j)} \quad (24)$$

Burada $\hat{G}(Z_i)$ dağılımın Kaplan-Meier tahminidir. Bu yöntemin en önemli avantajı hem teorik hem de pratik olarak daha kolay hesaplanabilmesi olduğu söylenebilir. Bu yöntemin kullanıldığı son zamanlarda yapılan çalışmalar arasında Aydın and Yılmaz ve Aydın and Yılmaz gösterilebilir.[27,28]

Örnek (Sentetik Veri): Sentetik veri dönüşüm yöntemleri arasında en sık kullanılan yöntem, KSV dönüşümü olduğundan ve diğer dönüşüm yöntemleri de benzer şekilde çalıştığından bu bölümde yalnızca KSV dönüşüne yönelik bir örnek hazırlanmıştır. Şekil 2'de KSV dönüşümü ile elde edilen verilere karşılık, orijinal ve sansürlü gözlemler sunulmuştur.



ŞEKİL 2: KSV dönüşümü ile %20'si sansürlü üretilmiş yanıt değişkenleri için sansürlü veri (kırmızı), orijinal veri (siyah) ve sentetik veri (mavi).

Görüldüğü gibi, sentetik veri dönüşümünde sansürlü her bir gözleme sıfır değeri atanırken, bu gözlemlerin etkisi diğer veri noktalarına aktarılarak onların büyüklüğü arttırılmakta ve böylece sentetik veri ve orijinal verinin beklenen değerleri ($E(Y) \cong E(Y_{\hat{G}})$) eşitlenmeye çalışılmaktadır. Çok sık kullanılan bir yöntem olmasına karşın, noktasal tahminde oldukça yanlı sonuçlar vermektedir. Ayrıca verinin yapısına müdahil olunduğundan, ham verinin kendi özellikleri yitirilmiş olmaktadır.

*k*NN Yerine Koyma Yöntemi

Bu bölümde, sağdan-sansürlü veri noktalarını tahmin etmek için *k*NN yerine koyma yöntemi anlatılmıştır (Batista ve Monard).[29] Bu yöntem, Gaussian yerine koymaya benzer bir şekilde çalışsa da tamamen dağılımdan bağımsız bir yöntemdir. Bu yöntemin bazı avantajları şu şekilde sıralanabilir; Yerine koyulmak için elde edilen veriler sentetik yolla veya başka şekilde oluşturularak değil, verinin kendisi kullanılarak tahmin edilir. *k*NN açıklayıcı değişkenler hakkında daha fazla bilgi sağlar. Yöntem parametrik olmayan bir yöntemdir ve hiçbir varsayım kullanmaz. *k*NN hem kesikli hem de sürekli değişkenler için kullanılabilir.

Bu yöntem, veri noktalarının arasındaki mesafelere bağlı olarak benzerliğe dayalı çalışan bir yöntemdir. Genellikle bu mesafe, Öklid uzaklıkları ile ölçülür (Strike vd.,):[30]

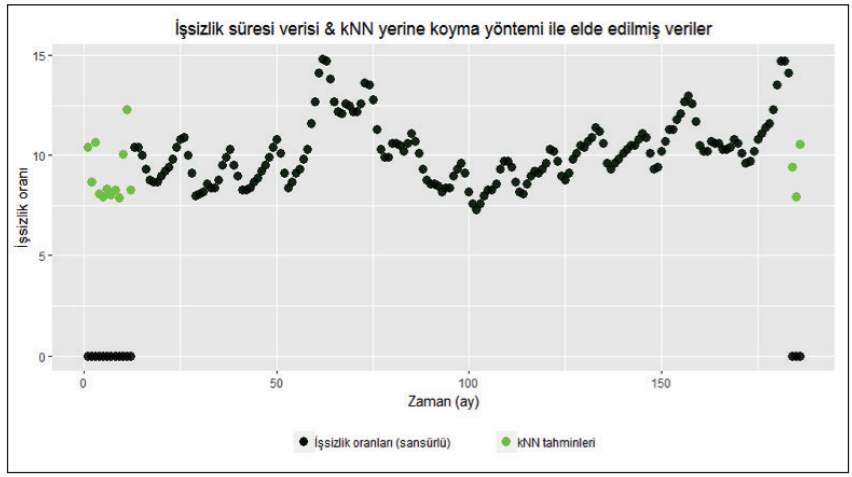
$$d_E(x,y) = \sqrt{\Sigma_i^n (x_i - z_i)^2} \qquad (25)$$

*k*NN yerine koyma yönteminin uygulanabilmesi için geliştirilen algoritma Tablo 3'de verilmiştir;

TABLO 3: *k*NN için algoritma.

Algoritma: *Sağdan−sansürlü veriler için kNN yerin koyma yöntemi*

Girdi: Sağdan sansürlü veriseti z_i
Sansür işaretçisi δ_i
En yakın komşu sayısı k
Açıklayıcı değişken değerleri x_i
Çıktı:Sansürlü verilerin tahminlerini içeren yeni değişken $\mathbf{y}^{knn} = (y_1^{knn},..., y_n^{knn})^T$
1 **başla**
2 **for** ($i = 1$ to n)
3 **Eğer** ($\delta_i = 0$) **yap** (if data point is censored)
4 **for**($j = 1$ to n)
5Her bir sansürlü gözlem için x_j ve x_i için eşitlik 24'te verilen öklit uzaklıkları bulunur
6 Mesafeler küçükten büyüğe sıralanır
7 **for** ($j = 1$ to k)
8 Sıralanan mesafelerle ilişkili ilk k adet sansürlü olmayan gözlem alınır.
9 i'nci sansürlü veri tahmini (y_i^{knn}) en yakın k adet y_i değerinin ortalaması alınarak hesaplanır.
10 Tahmin edilen gözlemler (y_i^{knn}) sansürlü gözlemler ($z_i, \delta_i = 0$)ile yer değiştirilir.
11 $\mathbf{y}^{knn} = (y_1^{knn},..., y_n^{knn})^T$ oluşturulur.
12 **Son**



ŞEKİL 3: İşsizlik süresi oranları verisi (sansürlü-siyah) sansürlü gözlemlerin kNN yerin koyma yöntemi ile tahminleri (yeşil).

Böylece Model (4)'te yanıt değişkeni yerine $\mathbf{y}^{knn}$ kullanılabilir ve sansürün etkisi modele eklenmiş olur. Tahmin işleyişi, $\mathbf{y}^{knn}$ bağlı olarak değişiklik göstermeden devam eder.

Örnek (*k*NN Yöntemi): *k*NN yerine koyma yönteminin nasıl çalıştığının görülebilmesi için Türkiye'de 2014-2019 yılı aylık işsizlik süresi oranlarından oluşan veri-seti kullanılmıştır. Bu veri seti sağdan sansürlü diğer veri setlerinden biraz farklı olarak sıfırdan itibaren sağdan sansürlenmiştir. Dolayısıyla, gözlenebilen verilerle sansürlü veriler arasında yüksek bir fark gözlenmektedir. Bu da her ne kadar veri seti %8 (hafif) sansür içeriyor olsa da, veriyi tamamlamayı zorlaştırmaktadır.

Şekilde görüldüğü gibi, *k*NN yöntemi, en yakın komşu noktalardan faydalandığı için bu tip veriyi tamamlamakta zorluk çekmemiştir. Elbette verinin orijinal değerleri bilinmediğinden noktasal olarak tahminin kalitesi ölçülememektedir. Fakat bu konuda detaylı olarak yapılmış bir başka çalışma olan Ahmed vd. incelenebilir.[31]

SONUÇ VE ÖNERİLER

Bu çalışmanın sağdan sansürlü veriler ile ilgili önerilen tüm yöntem ve analizleri derinlemesine barındırmadığı açıktır fakat hem sağ-kalım analizleri hem de diğer sansürlü veri içeren alanlarda çalışmaya başlayacak olan araştırmacılar için önemli bir yol gösterici olacağına inanıyoruz. Bu çalışmada sağdan-sansürlü verilerin regresyon modeline eklenmesinde sansür etkisinin nasıl modele uygun şekilde dâhil edileceğine dair hem parametrik hem parametrik olmayan yöntemler tartışılmıştır. Bu teknikler literatürde ilgili alanlarda geniş bir kullanım alanına sahip yöntemlerdir. Örneğin; parametrik olmayan *k*NN yerine koyma yöntemi, genetik verilerin analiz edilmesinde son dönemde oldukça sık kullanılmaya başlanmıştır.

Temel olarak çalışmadan elde edilecek öneriler şu şekilde sıralanabilir;

• Sağdan sansürlü verilerin modellenmesinde, yanıt değişkeni sansürlü olduğunda ne parametrik ne de parametrik olmayan hiçbir yöntem doğrudan kullanılamamaktadır. Bu durum oldukça yanlı sonuçlara sebep olacaktır. Bu nedenle bu çalışmada tanıtılan yöntemlerden veya bu yöntemlerden türetilmiş herhangi bir çözüm yöntemi kullanılmalıdır.

• Detaya inildiğinde sansürlü verinin dağılımının bilindiği ve bilinmediği durumlar için çözüm yöntemleri ikiye ayrılmaktadır. Ayrıca verinin toplandığı alan da önemli hâle gelmektedir. Örneğin; sağlık alanında toplanan yaşam süresi verilerinin, Weibull veya genelleştirilmiş gama dağılımına uyması beklenirken, sansürlü zaman serilerinde veya bir cihazın ömrünün incelenmesinde verilerin dağılımı hakkında belirsizlikler ortaya çıkabilmektedir. Bu çalışmada her iki durum için de başlıca yöntemler tanıtılmıştır.

• Yukarıda belirtildiği gibi, gerçek hayatta verilerin dağılımı ile ilgili bilgi çoğunlukla sağlanamadığından Kaplan-Meier ağırlıkları, *k*NN yerine koyma yöntemi ve sentetik veri dönüşümleri parametrik yöntemlere görece daha sık kullanılmaktadır.

Finansal Kaynak

Bu çalışma sırasında, yapılan araştırma konusu ile ilgili doğrudan bağlantısı bulunan herhangi bir ilaç firmasından, tıbbi alet, gereç ve malzeme sağlayan ve/veya üreten bir firma veya herhangi bir ticari firmadan, çalışmanın değerlendirme sürecinde, çalışma ile ilgili verilecek kararı olumsuz etkileyebilecek maddi ve/veya manevi herhangi bir destek alınmamıştır.

Çıkar Çatışması

Bu çalışma ile ilgili olarak yazarların ve/veya aile bireylerinin çıkar çatışması potansiyeli olabilecek bilimsel ve tıbbi komite üyeliği veya üyeleri ile ilişkisi, danışmanlık, bilirkişilik, herhangi bir firmada çalışma durumu, hissedarlık ve benzer durumları yoktur.

Yazar Katkıları

Fikir/Kavram: *Dursun Aydın;* ***Tasarım:*** *Ersin Yılmaz;* ***Denetleme/Danışmanlık:*** *Dursun Aydın;* ***Veri Toplama Ve/Veya İşleme:*** *Ersin Yılmaz;* ***Analiz Ve/Veya Yorum:*** *Dursun Aydın, Ersin Yılmaz;* ***Kaynak Taraması:*** *Ersin Yılmaz;* ***Makalenin Yazımı:*** *Ersin Yılmaz;* ***Eleştirel İnceleme:*** *Dursun Aydın.*

KAYNAKLAR

1. Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. J Am Stat Assoc. 1958;53(282):457-81. [Crossref]
2. Miller RG. Least squares regression with censored data. Biometrika. 1976;63(3):449-64. [Crossref]
3. Stute W. Consistent estimation under random censorship when covariables are present. J Multivar Anal. 1993;45(1):89-103. [Crossref]
4. Stute W. The central limit theorem under random censorship. Ann Stat. 1995;23(2):422-39. [Crossref]
5. Stute W. Nonlinear censored regression. Statistica Sinica. 1999;9:1089-102.
6. Miller R, Halpern J. Regression wtih censored data. Biometrika. 1982;69(3):521-31. [Crossref]
7. Cox DR. Regression models and life-tables (with discussion). Journal of Royal Statistics Society B. 1972;34(2):187-220. [Crossref]
8. Buckley J, James I. Linear regression with censored data. Biometrika. 1979;66(3):429-36. [Crossref]
9. Koul H, Susarla V, Van Ryzin J. Regression analysis wtih randomly right-censored data. Ann Statist. 1981;9(6):1276-85. [Crossref]
10. Wang QH, Li G. Empirical likelihood semiparametric regression analysis under random censorship. J Multivar Anal. 2002;83(2):469-86. [Crossref]
11. Chen S, Khan S. Semiparametric estimation of a partially linear censored regression model. Econometric Theory. 2001;17(3):567-90. [Crossref]
12. Dikta G. The strong law under semiparametric random censorship models. J Stat Plan Inference. 2000;83:1-10. [Crossref]
13. Wang QH, Zheng ZG. Asymptotic properties for the semiparametric regression model with randomly censored data. Science in China Series A. 1997;40(9):945-57. [Crossref]
14. Leurgans S. Linear models, random censoring and synthetic data. Biometrika. 1987;74(2):301-9. [Crossref]
15. Lai TL, Ying Z, Zheng ZK. Asymtotic normality of a class of adaptive statistics with applications to synthetic data methods for censored regression. J Multivar Anal. 1995;52(2):259-79. [Crossref]
16. Qin G, Jing B. Asymptotic properties for estimation of partial linear models with censored data. J Stat Plan Inference. 2000;84(1-2):95-110. [Crossref]
17. Powell JL. Least absolute deviations estimation for the censored regression model. J Econom. 1984;25(3):303-25. [Crossref]
18. Fan J, Gijbels I. Censored regression: local linear appriximations and their applications. J Am Stat Assoc. 1994;89(426):560-70. [Crossref]
19. Ritov Y. Estimation in a linear regression model with censored data. Ann Stat. 1990;18(1):303-28. [Crossref]
20. Tsiatis AA. Estimating regression parameters using linear rank tests for censored data. Ann Stat. 1990;18(1):354-72. [Crossref]
21. Breslow N, Crowley J. A large sample study of the life table and product limit estimates under random censorship. Ann Statist. 1974;2(3):437-53. [Crossref]
22. Wei LJ, Ying Z, Lin DY. Linear regression analysis of censored survival data based on rank tests. Biometrika. 1990;77(4):845-51. [Crossref]
23. Zhou M. Asymptotic normality of the 'synthetic data' regression estimator for censored survival data. Ann Statist. 1992;20(2):1002-21. [Crossref]
24. Park JW, Genton MG, Ghosh SK. Censored time series analysis with autoregressive moving average models. Can J Stat. 2007;35(1):151-68. [Crossref]
25. Stute W, Wang JL. The strong law under random censorship. Ann Statist. 1993;21(3):146-56. [Crossref]
26. Yılmaz E, Aydın D. A comparison of two methods for estimating censored linear regression models. International Journal of Statistics in Medical and Biological Research. 2017;1, 1-8.

27. Aydın D, Yılmaz E. Modified spline regression based on randomly right-censored data: a comparative study. Commun Stat Simul Comput. 2018;47(9):2587-611. [Crossref]

28. Aydın D, Yılmaz E. Modified estimators in semiparametric regression models with right-censored data. J Stat Comput Simul. 2018;88(8):1470-98. [Crossref]

29. Batista G, Monard M. An analysis of four missing data treatment methods for supervised learning. Applied Artificial Intelligence, 2003;17(5-6):519-33. [Crossref]

30. Strike K, El Emam K, Madhavji N. Software cost estimation with incomplete data. IEEE Transactions on Software Engineering. 2001;27:890-908. [Crossref]

31. Ahmed SE, Aydın D, Yılmaz E. Nonparametric regression estimates based on imputation techniques for right-censored data. Proceedings of the Thirteenth International Conference on Management Science and Engineering Management. 2019;109-20. [Crossref]