

## Aykırı Değer Varlığında Cox Regresyon Analizi için Yeni Bir Yaklaşım

Nesrin ALKAN<sup>\*1</sup>, B. Barış ALKAN<sup>1</sup>

<sup>1</sup>Sinop Üniversitesi, Fen Edebiyat Fakültesi, İstatistik Bölümü, Sinop

(Alınış / Received: 29.03.2017, Kabul / Accepted: 13.07.2017, Online Yayınlanma / Published Online: 11.08.2017)

### Anahtar Kelimeler

Cox regresyon analizi,  
Çoklu değer atama yöntemi,  
Aykırı değer,  
Kayıp değer

**Özet:** Aykırı değer varlığı Cox regresyonun en önemli varsayımlarından olan orantılı hazard varsayımının ihlal olmasına ve doğru olmayan tahminlerin ortaya çıkmasına neden olur. Çünkü aykırı değerler, modelin parametrelerinin tahminleri üzerinde güçlü bir etkiye sahiptirler. Bu nedenle veri kümesinde aykırı değerlerin olması araştırmacılar için bir problemdir. Bu çalışmada, aykırı değerlerden dolayı orantılı hazard varsayımının ihlal edilmesi sonucu ortaya çıkan problemin çözümü farklı bir bakış açısıyla ele alınmıştır. Buna göre aykırı değer problemi bir kayıp değer problemi gibi düşünülüp çoklu değer atama yöntemi kullanılarak çözülmüştür. Sonuç olarak Cox regresyon analizinin orantılı hazard varsayımı tehlike altında ise kayıp veri problemlerinde üstün bir performans gösteren çoklu değer atama yöntemi ile elde edilen tahminler kullanılarak problemin çözülmesi önerilmektedir.

## A New Approach for Cox regression Analysis in The Presence of Outliers

### Keywords

Cox regression,  
Multiple imputation,  
Outliers,  
Missing value

**Abstract:** Outliers in data could lead to violation of proportional hazard assumption which is one of the most important assumptions of Cox regression and it leads to the emergence of inaccurate estimates. Because they have strong influence on the estimates of the parameters of model. For this reason, the presence of outliers in the data set is a problem for the researchers. The solution of the problem that result from a violation of assumptions was discussed with a different perspective. So the problem caused by outliers was transformed as an missing value problem and it was solved by multiple imputation method. Consequently, if the proportional hazard assumption of the Cox regression analysis is violated, using the estimates obtained by the multiple imputation method that shows superior performance in missing data problem is suggested to solve the problem.

### 1. Giriş

Hemen hemen bütün istatistiksel analizlerde, veri kümesinde yer alabilen kayıp değerler ve aykırı değerler problem oluşturur. Aykırı değer, diğer verilerle karşılaştırıldığında verinin geri kalanından farklı olan değerlerdir. Pek çok araştırmada aykırı değerli gözlemlerle karşılaşmakta ve varsayımları ihlal etmesi nedeniyle yok sayılarak analizler yapılmaktadır. Araştırmacıların büyük paralar ve zaman harcayarak elde ettikleri veriler gözlemlerin aykırı değer içermesi nedeniyle heba olmaktadır. Verilerdeki bu silme işlemiyle örnek genişliği küçülmekte ve bu durum istatistiksel gücün azalmasına neden olmaktadır. Tüm bu nedenlerden dolayı aykırı değer probleminin giderilmesi önem kazanmaktadır.

Aykırı değerler, parametre tahminleri üzerinde herhangi bir etkiye sebep olmayabilir, buna karşın sonuçlar üzerinde çok büyük bir etkisi de olabilir. Bu yüzden bir veri kümesinde aykırı değerlerin belirlenmesi istatistiksel analizler açısından çok önemlidir. Araştırmaların çoğunda aykırı değerler, modeldeki parametre tahminlerini etkileyebilir, seçilen modeli değiştirebilir ve modele dayanan tahminleri etkileyebilir düşüncesiyle veriden çıkarılarak, verilerin tekrar modellenmesi yoluna başvurmaktadır. Fakat bu da veri kümesinin küçülmesine ve aykırı değer bulunduğ gözlemin sahip olduğu diğer önemli bilgiler içeren bağımsız değişken değerlerinin de silinmesine neden olmaktadır [1]. Bu yüzden bu çalışmada veri kümesinde bulunan aykırı değerli gözlemin tümünü silmek yerine sadece aykırı değeri silerek, yerine

çoklu değer atama yöntemi ile değer atayarak aykırı değer probleminin çözülmesi amaçlanmaktadır.

İstatistiksel çalışmalarda karşılaşılan bir diğer problem kayıp değerlerdir. Kayıp değerli veri araştırmacılar için problem oluşturur. Çünkü geleneksel istatistiksel yöntemler ve programlar verinin tam olması durumu için tasarlanmıştır. Bu nedenle, kayıp değerli verinin çözümü iki şekilde yapılmaktadır. Bunlar kayıp değerli gözlemi silmek ve kayıp değere, değer atamaktır. Kayıp değerli veride denek veya değişkenlerin bilinen değerleri kullanılarak, kayıp değerlerin tahmin edilmesi değer atama olarak adlandırılır. Değer atama işleminden sonra kayıp değerler tamamlanarak yeni kayıp değer içermeyen veri seti oluşturulmuş olur. Oluşturulan bu yeni tamamlanmış veriye istenilen klasik istatistiksel analizler uygulanabilir. Verideki kayıp değerleri tahmin eden yöntemlerden biri çoklu değer atama yöntemidir. Çoklu değer atama yöntemi kayıp veri problemini çözen, örnek genişliği 100'den büyük olması durumunda kayıp değer oranı ne olursa olsun orijinal sonuçlara oldukça benzer sonuçlar veren bir yöntemdir [2].

Çalışmanın uygulama bölümünde Ondokuz Mayıs Üniversitesi Tıp fakültesinden alınan 174 akciğer kanserli hastanın sağkalım verisi kullanılmıştır. Schoenfeld artık analizi yöntemi ile hem verideki aykırı değerlerin tespiti hem de orantılı hazard varsayımının kontrolü yapılmıştır. Önerilen yöntemin başarılı olup olmadığını belirlemek için kullanılmak üzere öncelikle verinin orijinal hali analiz edilerek, orijinal parametre tahminleri ile karşılaştırılma imkanı sağlanmıştır. Daha sonra veri setinin %10'u aykırı değer içerecek şekilde düzenlenerek aykırı değerlerin varsayımı nasıl bozduğu ve varsayım ihlali durumunda parametreler üzerindeki etkilerinin gösterilmesi amaçlanmıştır. Ayrıca aykırı değerden kaynaklanan problemi çözmek için ilk olarak veri kümesinde yer alan aykırı değerler tespit edilip, bu değerler kayıp değer olarak düşünülmektedir. Daha sonra bu kayıp değerler yerine, son yıllarda yapılan kayıp veri analiz yöntemleri ile ilgili çalışmalarda üstün başarı gösteren çoklu değer atama yöntemi ile elde edilen tahminler atanarak veri setindeki aykırı değerler giderilmiş ve parametre tahminleri orijinal verinin tahminleri ile karşılaştırılmıştır. Schoenfeld artık analizi ve Cox regresyon analizi için R 3.3.1 programı, çoklu değer atama yöntemi için ise SAS 9.3 programı kullanılmıştır [3,4].

## 2. Materyal ve Metot

### 2.1. Cox Regresyon Analizi

Sansürlü verilerin yer aldığı sağkalım analizinde bağımlı değişken ile bağımsız değişkenler arasındaki neden-sonuç bağıntısını ortaya koymak için yararlanılan regresyon yöntemine Cox regresyon analizi adı verilir [5]. Sağlık alanındaki araştırmalarda hastalık tipi ve araştırma konusuna

göre belirlenen faktörleri dikkate alarak çözüm aramak gerekir. Birçok faktörü birlikte değerlendirmede ilk akla gelen yöntem çoklu regresyon analizidir. Fakat çoklu regresyon modelleri sonuç açısından olgular arasında zaman içinde oluşan farkı dikkate alarak değerlendirme yapma imkanına sahip değildir. Bu nedenle zaman içindeki değişimi dikkate alan ve sansürlü verilerle analiz yapmayı kolaylaştıran ve Cox [6] tarafından önerilen Cox regresyon modelinin kullanılması önemlidir.

Cox regresyonun temel varsayımı, modelde zaman eksenini boyunca herhangi iki bireyin hazard oranının (HR) sabit olmasıdır. Bu yüzden Cox regresyon modeli orantılı hazard modeli olarak da bilinmektedir.

Bağımsız değişkenler vektörü  $\mathbf{x}$  ve sağkalım süresi  $t$  olsun. Böylece bir bireyin bağımsız değişkenlere göre hazard fonksiyonu  $h(t;\mathbf{x})$  ile gösterilir. Bu durumda Cox Regresyon modelinin matematiksel ifadesi,

$$h(t;\mathbf{x})=h_0(t)\exp(\beta' \mathbf{x}) \quad (1)$$

eşitliği ile verilir. Burada  $h(t;\mathbf{x})$ , hazard fonksiyonu  $\mathbf{x}$ ,  $p \times 1$  boyutlu bağımsız değişkenler vektörü,  $\beta'$ ,  $1 \times p$  boyutlu bilinmeyen regresyon katsayıları vektörü  $h_0(t)$ , temel hazard fonksiyonudur. Cox Regresyon modelinin parametre tahmininde olabilirlik fonksiyonu yerine kısmi olabilirlik fonksiyonu kullanılır. Bu fonksiyon,

$$L_p(\mathbf{x} / \beta) = \prod_{i=1}^k \frac{\exp(\beta' \mathbf{x}_i)}{\sum_{j \in R(t_i)} \exp(\beta' \mathbf{x}_j)} \quad (2)$$

eşitliği ile verilir. Burada  $k$ , ilgilenilen olayın gerçekleşme sayısını,  $R(t_i)$  ise  $t_i$  süresinde risk altında olan tüm gözlemleri ifade eder [7].

### 2.2 Çoklu Değer Atama Yöntemi

Çoklu değer atama verilerdeki kayıp değer sorununun çözümü için Bayesci yaklaşımlar geliştirilen bir yöntemdir. Rubin (1987) çoklu değer atamayı, veri kümesindeki kayıp değerlerin olasılık dağılımının karşılık gelen mümkün değerle tamamlanması olarak tanımlamıştır [8].

Çoklu değer atama yönteminde her bir kayıp değer  $m$  kez tamamlanarak  $m$  tane tam veri seti oluşturulur. Çoklu değer atama yönteminde birden fazla bağımsız değişken için çoklu değer atama süreci gerçekleştirilir. Bu süreçte çok değişkenli kayıp değere sahip  $p$  boyutlu veri için kayıp değer yerine atama yapılmış veri setleri oluşturulur. Değer atanmış veri seti sayısı  $m$  ile gösterilir. Daha sonra bu  $m$  tam veri seti herhangi bir istatistiksel bir teknik ile analiz edilir. Son olarak  $m$  tam veri setinden elde

edilen sonuçların birleştirilmesi ile parametreler hakkında istatistiksel çıkarımlar elde edilir. Çoklu değer atama yönteminde temel kararlardan biri tamamlanmış veri seti sayısının ( $m$ ) belirlenmesidir. Rubin [8] ve Schafer [9], bu sayının üç-beş olmasını tavsiye etmişlerdir.

Bu yöntem değer atama ve birleştirme olmak üzere iki aşamada incelenebilir. Değer atama aşaması ise I ve P adımlarından oluşur. I adımında gözlenen değişkenlerden kayıp değerleri tahmin eden regresyon eşitlikleri oluşturmak için ortalama vektörünün ve kovaryans matrisinin elemanları kullanılır. P adımı ortalama vektörü ve kovaryans matrisinin sonsal dağılımını tanımlayan bir Bayes analizidir. Bu adımda I adımında oluşturulan tamamlanmış veri kullanılarak ortalama vektörü ve kovaryans matrisi tahmin edilir. Bu tahmin vektörü ve matrisinin her bir elemanına artık terimin eklenmesiyle yeni bir parametre seti oluşturulur. Sonsal dağılımlarından yeni parametreler çekilmesinden sonra I adımı güncellenmiş tahminleri kullanarak yeni regresyon katsayılar setini ve farklı atama değerlerini oluşturur. Bu yeni atamalar bir sonraki P adımına taşınır. Burada yeni parametre setleri oluşturulur [10].

Birleştirme aşamasındaki amaç,  $m$  farklı veri kümesine uygulanan tekniğin sonuçlarının tek bir kümede birleştirilmesidir. Rubin (1987) birleştirilmiş parametre tahminleri ve standart hatalar için formülleri özetlemiştir [8]. Örneğin birleştirilmiş parametre tahminleri analiz aşamasında elde edilen tahminlerin basit bir aritmetik ortalamasıdır [10].

### 2.3 Aykırı Değer

Aykırı değer, diğer verilerle karşılaştırıldığında verinin geri kalanından farklı olan değerlerdir. Veri setinde bulunan aykırı değerler, parametre tahmini üzerinde büyük bir etkiye sahip olabilirler. Aykırı değere sahip olan gözlemin veri kümesinden çıkarılması parametre değerlerini tamamen değiştirebilir.

Cox regresyonda kullanılan kısmi olabilirlik fonksiyonu, varsayımların ihlalinden etkilenmektedir. Aykırı değer varlığı, Cox regresyonun en önemli varsayımlarından olan orantılı hazard varsayımının ihlal olmasına neden olabilir.

Bu durumda yanlış tahminler elde edilir, bu da doğru olmayan sonuçlara yol açar ve modelin başarısızlığı anlamına gelir. Tüm bu nedenlerden dolayı veri kümesinde aykırı değer varlığı araştırmacılar için bir problemdir ve sağkalım analizinde veri kümesindeki aykırı değerlerin belirlenmesi konusu oldukça önemlidir. Aykırı değerlerin tespiti, artıkların analizine dayanmaktadır. Cox regresyon modelinde kullanılan pek çok artık analiz yöntemlerinden biri Schoenfeld tarafından önerilen, skor artıkları olarak da adlandırılan Schoenfeld artık analizidir [11].

Schoenfeld artıkları bağımsız değişkenin gerçek değeri ile ağırlıklı risk skorlarının ortalaması arasındaki farktır [12]. Oransal hazard varsayımının geçerliliğini kontrol amacıyla kullanılan Schoenfeld artıkları zamana karşı çizilir. Artıkların yatay bir doğru etrafında tesadüfi olarak yer alması, oransal hazard varsayımının sağlandığını gösterir. Schoenfeld artıkları,  $k$ . bağımsız değişken ve  $i$ . birim için

$$\hat{r}_{ki} = c_i \left[ X_{ki} - \frac{\sum_{j \in R(t_i)} X_{kj} \exp(\hat{\beta}' X_j)}{\sum_{j \in R(t_i)} \exp(\hat{\beta}' X_j)} \right] \quad (3)$$

eşitliği ile hesaplanır. Burada  $c_i$ ,  $i$ . birimin sansür durumunu gösterir. Hosmer ve Lemeshow (1999), ölçeklenmiş Schoenfeld artıkları grafiğinin oransal hazard varsayımı için kullanılmasını önermiştir [13]. Regresyon katsayılarının kovaryans matrisine dayalı ölçeklenmiş Schoenfeld artıkları ile aykırı değerleri bulmak için kullanılan ve zamana karşı her bir ortak değişken için çizilen grafikler elde edilir [14]. Ölçeklenmiş Schoenfeld artıkları,

$$\hat{r}_{ki}^* = m \sum_{i=1}^p V_{ki} \hat{r}_{ki} \quad (4)$$

ile hesaplanır. Eşitlikde  $m$  toplam başarısız birey sayısını,  $V$  ise regresyon katsayılarından tahmin edilmiş kovaryans matrisini göstermektedir.

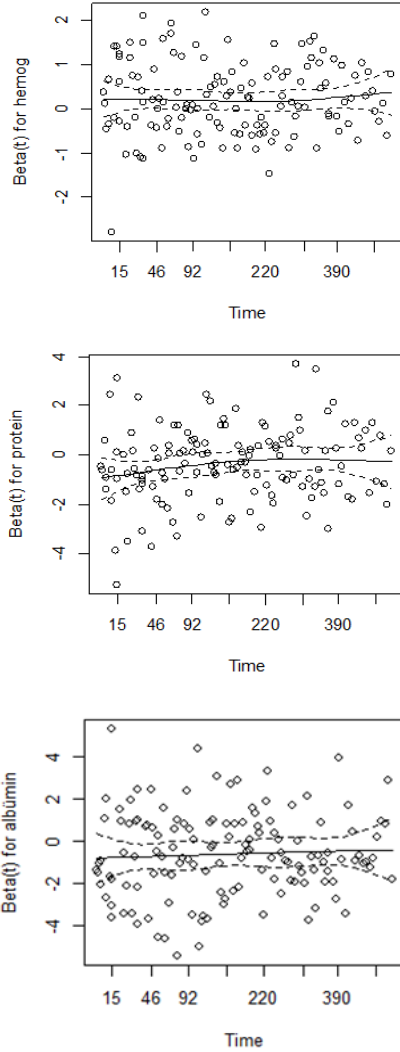
### 3. Bulgular

Aykırı değer nedeni ile orantılı hazard varsayımının ihlal edilmesi sonucu ortaya çıkan problemin çözümü farklı bir bakış açısıyla çoklu değer atama yöntemi ile çözülmesi amaçlanmıştır. Bu amaçla, Ondokuz Mayıs Üniversitesi, Tıp Fakültesinden alınan 174 akciğer kanserli hastanın sağkalım verisi kullanılmıştır. Veri setinde 174 hastanın 36'sı çalışma sona erdiğinde hala yaşamaktadır yani sansürlü hastalardır. Çalışmada yer alan hastaların sağkalım süreleri ortalaması 206 gün olarak hesaplanmıştır. Ayrıca çalışmada bağımsız değişkenler olarak incelenen hemoglobin 12.66 ortalama ile 6.8-17.6 arasında Protein 7.02 ortalama ile 4.2-9.2 arasında ve Albümin 3.73 ortalama ile 2- 7.2 arasında değerler almaktadır.

Veri setinde aykırı değer tespiti için ölçeklenmiş Schoenfeld artık analizi kullanılmış aykırı değere rastlanmamıştır. Fakat protein değişkeninde 17 tane gözlemin aykırı değer olma potansiyeline sahip olduğu gözlenmiştir. Çalışmada önerilen yöntemin, problem çözme konusunda başarısını göstermek amacıyla bu potansiyel değerlerin güçlü aykırı değerler olması için bu gözlemlerin değerlerine uç değerler verilmekte ve böylece veri setindeki aykırı değer yüzdesi %10 olarak belirlenmektedir.

Çalışmanın ilk aşamasında veri setinde aykırı değer olmaması durumunda yani verinin orijinal hali için Cox regresyon analizinin varsayım testi, ölçeklenmiş

Schoenfeld artık analizi ile yapılmakta ve daha sonra Cox regresyon analizi uygulanarak parametreler tahmin edilmektedir. Bu sonuçlar orijinal sonuçlar olarak düşünülüp elde edilen diğer sonuçlar için referans olarak kullanılarak karşılaştırılmıştır. Orijinal veri setinin her bir bağımsız değişkeni için ölçeklenmiş Schoenfeld artık grafikleri Şekil 1'de verilmektedir.



**Şekil 1.** Orijinal veri seti için ölçeklenmiş Schoenfeld artık analizi grafikleri

Şekil 1'de verilen grafikler incelendiğinde hemoglobin, protein ve albümin bağımsız değişkenleri için artıklar yatay bir doğru etrafında tesadüfi olarak yer almaktadır. Yani bütün değişkenler orantılı hazard varsayımını sağlamaktadır. Ölçeklenmiş Schoenfeld artık analizi ile grafiksel sonuçların yanı sıra her bir değişken için istatistiksel sonuçlar da hesaplanır. Ölçeklenmiş Schoenfeld artık analizinin her bir değişken için hesaplanan istatistiksel sonuçları Tablo 1'de verilmiştir.

Tablo 1 incelendiğinde sonuçlar, grafiksel sonuçlar ile benzer bulunmuş olup bütün değişkenler için orantılı hazard varsayımı sağlanır. Bu durum, orijinal

veri setinin Cox regresyon analizi ile test edilebileceğini ve elde edilen sonuçların güvenilir olacağını gösterir. Aykırı değer içermeyen verinin yani orijinal verinin Cox regresyon analizi sonuçları Tablo 2'de verilmiştir.

**Tablo 1.** Orijinal veri seti için Ölçeklenmiş Schoenfeld artık analizi istatistiksel sonuçları

Parametreler	rho	Ki-kare	p
Hemog	0.0198	0.0669	0.7959
Prot.	0.1412	2.5760	0.1085
Albü.	0.0561	0.3721	0.5418

**Tablo 2.** Orijinal veri seti için Cox regresyon analizi sonuçları

	$\beta$	Standart hata	Wald istatistiği	P değeri	Exp( $\beta$ )
Hemg.	0.206	0.063	10.599	0.001	1.229
Prot.	-0.394	0.132	8.875	0.003	0.674
Albü.	-0.606	0.183	10.999	0.001	0.545

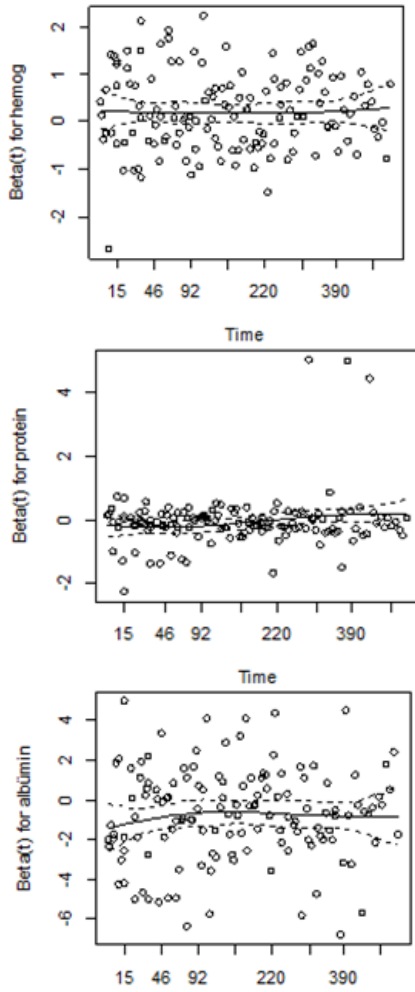
Tablo 2'de verilen sonuçlara göre hemoglobin, protein ve albümin değerleri sağkalım süresi üzerinde önemli bir etkiye sahiptir ( $p < 0.05$ ). Hemoglobin seviyesi normal seviyeden uzaklaştıkça hastanın ölüm riski 1.229 kat arttığı görülmektedir. Protein ve albümin seviyesinin düşük seviyeden normal seviyeye yükselmesi hastanın sağkalım süresini sırasıyla 1.48 ( $1/0.674$ ) ve 1.83 ( $1/0.545$ ) kat arttırmaktadır. Orijinal veri seti için Cox regresyon modeli, olabilirlik oran testi sonucu anlamlı ( $\chi^2 = 35.651$ ,  $p = 0.000 < 0.05$ ) bulunmuştur. Ayrıca model uyum kriteri olan Akaike bilgi kriteri (AIC) bu model için 1141,507 olarak hesaplanmıştır.

Çalışmanın ikinci aşamasında aykırı değer içeren veri seti kullanılarak Cox regresyon analizinin varsayım testi için ölçeklenmiş Schoenfeld artık analizi uygulanmakta ve daha sonra Cox regresyon analizi uygulanarak parametreler tahmin edilmektedir. Aykırı değer içeren veri setinin her bir bağımsız değişkeni için ölçeklenmiş Schoenfeld artık grafikleri Şekil 2'de verilmektedir.

Şekil 2'de verilen grafikler incelendiğinde hemoglobin ve albümin bağımsız değişkenleri için artıklar yatay bir doğru etrafında tesadüfi olarak yer almaktadır. Yani bu değişkenler orantılı hazard varsayımını sağlamaktadır. Fakat protein değişkeni varsayımı sağlamamakta ve aykırı değer içermektedir. Ölçeklenmiş Schoenfeld artık analizi ile her bir değişken için hem grafiksel sonuç hemde istatistiksel sonuçlar hesaplanır. Ölçeklenmiş Schoenfeld artık analizinin her bir değişken için hesaplanan istatistiksel sonuçları Tablo 2'de verilmiştir.

Tablo 3 incelendiğinde istatistiki sonuçların, grafiksel sonuçları desteklediği görülmektedir. Buna göre hemoglobin ve albümin orantılı hazard varsayımını sağlarken protein değişkeni varsayımı sağlamamaktadır. Varsayımı sağlamayan, artık değerler içeren bir veri seti için parametre

değerlerini tahmin etmek için Cox regresyon analizi uygulanmış ve sonuçlar Tablo 4’de verilmiştir.



Şekil 2. Aykırı değer içeren veri seti için ölçeklenmiş Schoenfeld artık analizi grafiği

Tablo3. Ölçeklenmiş Schoenfeld artık analizi istatistiksel sonuçları

Parametreler	rho	Ki-kare	p
Hemog	0.00695	0.00828	0.92749
Prot.	0.16422	6.23134	0.01255
Albü.	0.0458	0.33360	0.56355

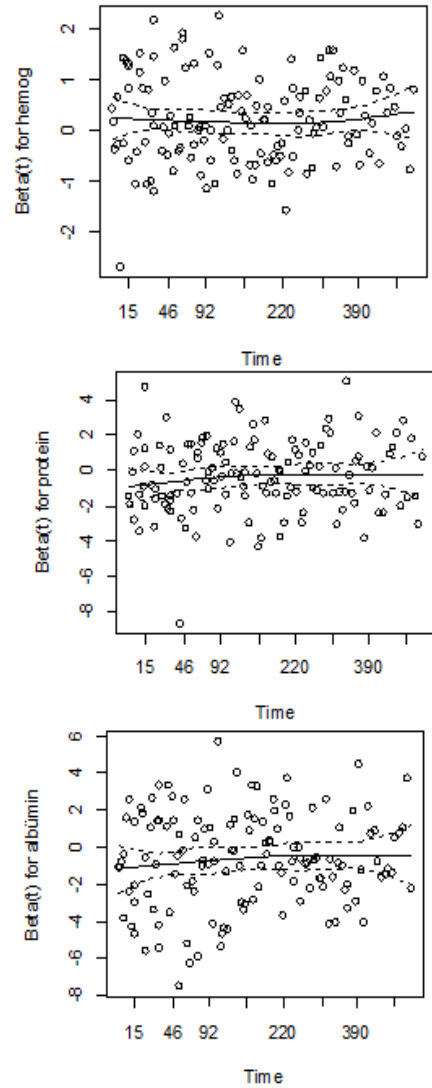
Tablo 4. Aykırı değer olması durumunda Cox regresyon analizi sonuçları

	$\beta$	Standart hata	Wald istatistiği	P değeri	Exp( $\beta$ )
Hemg.	0.203	0.064	10.143	0.001	1.225
Prot.	-0.082	0.058	2.049	0.152	0.921
Albü.	-0.798	0.186	18.448	0.000	0.450

Aykırı değer olması durumunda model anlamlı ( $\chi^2 = 29.936$ ,  $p=0.000 < 0.05$ ) olarak bulunmasına rağmen Akaike bilgi kriteri 1147.222 olarak elde edilmiştir. Bu değer, aykırı değerlerin etkisi ile orijinal verinin AIC değerinden daha büyük hesaplanmıştır. Tablo 4’de verilen sonuçlara göre hemoglobinin ve albümin değerleri sağkalım süresi üzerinde önemli bir etkiye sahip ( $p<0.05$ ) iken protein değişkeninin sağkalım süresini etkilemediği belirlenmiştir. Bu sonuç protein değişkeninde bulunan aykırı değerlerin parametre

tahminini ve modeli etkilediğini göstermektedir. Yani, orijinal verinin sonuçlarına göre modelde olması gereken ve sağkalım süresini etkileyen bir değişkenin, modelde olmasının anlamsız olduğu sonucu elde edilmektedir. Bu durum, araştırmacının yanlış model kurmasına ve bu modeli kullanarak yanlış kararlar vermesine sebep olacaktır.

Aykırı değer varlığı, sonuçları olumsuz anlamda etkilemesi nedeniyle araştırmacılar için bir problemdir. Bu problemin çözümü için bu çalışmada, öncelikle veri setinde bulunan aykırı değerler silinerek kayıp değerli veri seti elde edilir ve daha sonra bu kayıp değerlere çoklu değer atama yöntemi ile her bir kayıp değer için değer atanarak 5 tamamlanmış veri seti oluşturur. Çoklu değer atama (ÇDA) ile tamamlanmış veri setlerinin her biri için Schoenfeld artık analizi ile varsayım testi yapılmış ve tüm veri setleri için artık değerinin olmadığı ve varsayımın sağlandığı sonucu bulunmuştur. Bu veri setlerinden herhangi biri için Schoenfeld artık analizi grafiksel sonucu Şekil 3’de istatistiksel sonuç ise Tablo 5’de verilmiştir.



Şekil 3. ÇDA sonucu tamamlanan veri seti için Ölçeklenmiş Schoenfeld artık analizi grafiği

Şekil 3’de verilen grafikler incelendiğinde hemoglobin, protein ve albümin bağımsız değişkenleri için artıklar yatay bir doğru etrafında tesadüfi olarak yer almaktadır. Yani bütün değişkenler orantılı hazard varsayımını sağlamaktadır.

**Tablo 5.** ÇDA sonucu tamamlanan veri seti için Ölçeklenmiş Schoenfeld artık analizi istatistiksel sonuçları

Parametreler	rho	Ki-kare	p
Hemoglobin	0.00493	0.00421	0.9483
Protein	0.09631	1.35789	0.2439
Albümin	0.08614	1.18031	0.2773

Tablo 5 incelendiğinde sonuçlar, grafiksel sonuçlar ile benzer bulunmuş olup bütün değişkenler için orantılı hazard varsayımı geçerlidir ve veri setinde aykırı bir değer bulunmamaktadır. Böylece aykırı değerlerden kaynaklanan problem giderilmiştir ve bu veri setinin parametre tahmininin orijinal verinin parametre tahminine benzerliğinin tespiti için çoklu değer atama yönteminde atanmış değerlerden oluşan 5 veri setine ayrı ayrı Cox regresyon uygulanmıştır ve elde edilen birleştirilmiş sonuçlar Tablo 6’da verilmiştir.

**Tablo 6.** ÇDA yöntemi sonrası Cox regresyon analizinin birleştirilmiş sonuçları

	$\beta$	Standart hata	%95 güven aralığı		p
Hemg.	0.1886	0.0639	0.06307	0.3139	0.0032
Protein	-0.358	0.1315	-0.6167	-0.101	0.0064
Albümin	-0.626	0.2024	-1.0229	-0.229	0.0020

Tablo 6’da verilen sonuçlara göre hemoglobin, protein ve albümin değerleri sağkalım süresi üzerinde önemli bir etkiye sahiptir ( $p < 0.05$ ). Hemoglobin seviyesi normal seviyeden uzaklaştıkça hastanın ölüm riski ( $e^{0.1886}$ ) 1.208 kat arttığı görülmektedir. Protein ve albümin seviyesinin düşük seviyeden normal seviyeye yükselmesi hastanın sağkalım süresini sırasıyla 1.43 ( $1/\exp(-0.358)$ ) ve 1.87 ( $1/\exp(-0.626)$ ) kat arttırmaktadır. Tüm bu sonuçlar, orijinal verinin Cox regresyon sonuçlarına çok benzemektedir. Böylece aykırı gözlemin sebep olduğu parametre tahmininde ve model üzerindeki olumsuz etki ortadan kaldırılmıştır.

Çalışmada kullanılan orijinal veri, aykırı değerli veri ve çoklu değer atama (ÇDA) ile tamamlanmış veri setleri için ayrı ayrı elde edilen analiz sonuçlarının benzerliklerini ve farklılıklarını daha rahat görmek için Tablo 7 oluşturulmuştur.

Tablo 7’ye göre aykırı değer olması durumunda protein değişkeni önemsiz bulunurken, problem çoklu değer atama yöntemi ile giderildikten sonra protein değişkeni önemli hale gelmiştir. Ayrıca parametre tahminleri de orijinal verinin sonuçlarına benzer şekilde bulunmuş olması önerilen yaklaşımın

ne kadar doğru bir yöntem olduğunun da önemli bir göstergesidir.

**Tablo 7.** Yapılan analizlerin özet sonuçları

		$\beta$	SE	p
Orijinal veri için Cox Reg.	Hemg.	0.206	0.063	0.01
	Prot.	-0.394	0.132	0.03
	Albü.	-0.606	0.183	0.01
		$\beta$	SE	p
Aykırı değerli veri için Cox Reg.	Hemg.	0.2040	0.0637	0.0013
	Prot.	-0.083	0.0578	0.1486
	Albü.	-0.801	0.1860	0.0002
		$\beta$	SE	p
ÇDA Sonrası Cox Reg.	Hemg.	0.189	0.064	0.0032
	Prot.	-0.359	0.132	0.0064
	Albü.	-0.626	0.202	0.0020

#### 4. Tartışma ve Sonuç

Cox regresyonun en önemli varsayımlarından olan orantılı hazard varsayımının ihlal olmasına neden olabilen aykırı değerler, veri setinde bulunan parametre tahmini üzerinde büyük bir etkiye sahip olabilirler. Böyle bir durumda da aykırı değer varlığı doğru olmayan sonuçlara yol açmaktadır. Bu nedenle, aykırı değer bir problemdir ve bu problemin çözümü için literatürde farklı uygulamalar vardır. Bu çalışmada, var olan literatüre katkıda bulunmak amacıyla problemin çözümü ile ilgili olarak çoklu değer atama yöntemine dayanan farklı bir yaklaşım önerilmiştir. Buna göre aykırı değer problemi, bir kayıp veri problemi gibi düşünülüp çoklu değer atama yöntemi kullanılarak giderilmiştir. Çoklu değer atama yöntemi, kayıp veri probleminin çözümünde çok etkili bir yöntemdir. Alkan, N. ve ark.(2013), tarafından yapılan çalışmada hem farklı kayıp oranlarında hem de farklı örnek genişliklerinde çoklu değer atama yönteminin gerçek veriye yakın sonuçlar verdiği simülasyon çalışması ile verilmiştir.

Bu çalışmada yapılan analizler sonucu çoklu değer atama sonrası hem varsayım sağlanmış hem de uygulanan Cox regresyon analizi sonuçları orijinal verinin Cox regresyon sonuçlarına çok benzer bulunmuştur. Sonuç olarak eğer aykırı değer nedeniyle Cox regresyon analizinin varsayımları tehlike altında ise kayıp veri problemlerinde üstün bir performans gösteren çoklu değer atama yöntemi ile elde edilen tahminler kullanılarak problemin çözülmesi önerilmektedir.

#### Kaynakça

- [1] Alkan, B.B., Atakan C., Alkan, N. 2015. A Comparison of Different Procedures for Principal Component Analysis in the Presence of Outliers, Journal of Applied Statistics, Vol. 42, No. 8, 1716-1722.
- [2] Alkan, N., Terzi, Y., Cengiz, M. A., Alkan B B. 2013. Comparison of Missing Data Analysis Methods in Cox Proportional Hazard Models. Türkiye Klinikleri Journal of Biostatistic, 5(2), 49-54.

- [3] R Development Core Team 2011. R: A language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna
- [4] SAS Institute, SAS 9. 3 2011. Output Delivery System: User's Guide, SAS institute, North Carolina
- [5] Sınayış, B., Erkut, H., 2009. Türkiye'de organizasyonel sürekliliği etkileyen faktörlerin incelenmesi, İTÜ Mühendislik Dergisi, 8 (1), 95-104.
- [6] Cox, D.R., 1972. Regression models and life tables. Journal of the Royal Statistical Society, 34, 187-220.
- [7] Kleinbaum, D.G., Klein, M., 1996. Survival Analysis, A Self Learning Text. Springer, 124 s, USA.
- [8] Rubin, D.B. 1987. Multiple Imputation for Nonresponse in Surveys, Wiley&Sons, New York.
- [9] Schafer, J.L. 1997. Analysis of Incomplete Multivariate Data, Chapman&Hall, London.
- [10] Enders, C.K., 2010. Applied Missing Data Analysis. Guilford Pres, s:165-286, New York.
- [11] Schoenfeld, D., 1982. Partial residuals for the proportional hazards regression model, Biometrika, 69, 239-241.
- [12] Yay, M., Çoker, E., Uysal, Ö., 2007. Yaşam Analizinde Cox Regresyon Modeli ve Artıkların İncelenmesi, Cerrahpaşa Tıp Dergisi, 38(4).
- [13] Hosmer D.W., & Lemeshow S. 1999. Applied survival analysis: regression modeling of time to event data. Wiley, John Wiley&Sons, Incorporated, Canada.
- [14] Karasoy D., Tuncer N., 2015. Outliers in survival analysis", Alphanumeric Journal, 3(2), 139-152.