Available online at www.alphanumericjournal.com

alphanumeric journal

The Journal of Operations Research, Statistics, Econometrics and Management Information Systems

Volume 3, Issue 2, 2015

*2015.03.02.STAT.08*

# OUTLIERS IN SURVIVAL ANALYSIS

Durdu KARASOY*

Nuray TUNCER†

*Department of Statictics, Hacettepe University, Ankara*

*Ministry Of Finance, The Presidency of Tax Inspection Board, Ankara*

*Received: 02 November2015*

*Accepted: 27 December 2015*

**Abstract**

Survival analysis is a collection of statistical methods for analyzing data where the outcome variable is the time until the occurrence of an event of interest. Outliers in survival analysis calculated differently from classical regression analysis. Outlier detection methods in survival analysis are commonly carried out based on residuals and residual analysis. In survival analysis, there are different types of residuals that are Cox-Snell, Martingale, Schoenfeld, Deviance, Log-odds and Normal deviance residuals. There are methods which are DFBETA, LMAX and Likelihood Displacement values for detecting influential observations. The residuals are analyzed during the study which is applied on a stomach cancer data set and the outliers are detected. After omitting these outliers, model is set up again and results were found better..

*Keywords: Survival analysis, survival models, outliers, influential observations, residuals.*
*Jel Code: C10, C14, C19, C24*

# YAŞAM ÇÖZÜMLEMESİNDE AYKIRI DEĞERLER

**Özet**

Yaşam çözümlemesi, tanımlanan herhangi bir olayın ortaya çıkmasına kadar geçen sürenin incelenmesinde kullanılan istatistiksel yöntemler bütünüdür. Yaşam çözümlemesinde aykırı değerler klasik regresyonda kullanılan yöntemlerden farklı yöntemler kullanılarak hesaplanmaktadır. Yaşam çözümlemesinde aykırı değer belirleme yöntemleri artıklara ve artıkların analizine dayanmaktadır. Yaşam çözümlemesinde kullanılan başlıca artık türleri Cox-Snell, Martingale, Schoenfeld, Sapma, Log-odds ve Normal sapma artıklarıdır. Etkili gözlemleri belirlemek için kullanılan yöntemler ise DFBETA, LMAX ve Olabilirlik Değişim değerleridir. İncelenen artık türleri mide kanseri ile ilgili verilere uygulanmış ve aykırı değerler belirlenmiştir. Belirlenen aykırı değerler çıkarılarak model yeniden kurulmuş ve aykırı değerler çıkarıldığında sonuçların daha iyi olduğu görülmüştür.

*Anahtar Kelimeler : Yaşam çözümlemesi, yaşam modelleri, aykırı değerler, etkili gözlemler, artıklar*
*Jel Kodu : C10, C14, C19, C24.*

## 1. GİRİŞ

Eldeki veri kümesine bir model uyarlandıktan sonra, uygulanan modelin varsayımlarının sağlanıp sağlanmadığının kontrol edilmesi gereklidir. Modelin kontrolü için tanı yöntemleri, modelleme sürecinin önemli bir kısmıdır (Collett, 1994). Bu süreçlerin çoğu artıkların analizine dayanmaktadır. Yaşam çözümlemesinde özellikle Cox orantılı tehlikeler modelinde değişik amaçlarla kullanılabilecek değişik artık türleri vardır (Fitrianto & Jiin, 2013).

Yaşam çözümlemesinde en yaygın kullanılan artık türü Cox ve Snell (1968) tarafından önerilen Cox-Snell artıklarıdır (Cox & Snell, 1968). Bu artık türü modelin uygunluğunun kontrolü için kullanılır. Diğer bir artık türü olan Martingale artıkları Barlow ve Prentice (1988) tarafından önerilmiştir ve Cox-Snell artıklarının doğrusal

* durdu@hacettepe.edu.tr (Corresponding author)
† nuray.tuncer@vdk.gov.tr

dönüşümüdür. Martingale artıkları açıklayıcı değişkenlerin modele uyumunu belirlemede kullanılır (Barlow & Prentice, 1988). Benzer şekilde sapma artıkları da Martingale artıklarının dönüşümüdür. Aykırı değerler genellikle Therneau, Grambcsh ve Fleming (1990) tarafından önerilen sapma artıkları ile görüntülenir (Therneau, Grambcsh & Fleming, 1990). Flemming ve Harrington (1991) sapma artıklarının referans örnekleme dağılımına sahip olmadığına ve veri kümesinde durdurulmuş gözlemler olmadığında bile standart normal dağılım yaklaşımının tatmin edici olmadığına dikkat çekmiştir (Flemming & Harrington, 1991). Nardi ve Schemper (1999) bu problemin üstesinden gelmek için yeni artık türleri önermiştir. Bu yeni artık türlerinin, aykırı değerleri belirlemede daha doğru bir sonuç verdiği iddia edilmektedir. Bu yeni artıklar log-odds ve normal sapma artıklarıdır (Nardi & Schemper, 1999). Diğer bir artık türü ise Cox orantılı tehlikeler (Cox proportional hazards) modelinde orantılılık varsayımının testinde yaygın olarak kullanılan ve Schoenfeld (1982) tarafından önerilen Schoenfeld artıklarıdır (Schoenfeld, 1982).

Bu çalışmanın amacı, yaşam çözümlemesinde karşılaşılan aykırı değerleri tanımlama yöntemlerini ve uygulama alanlarını incelemektir. Bu amaçla artık türleri incelenmiş ve gerçek bir veri kümesi üzerinde uygulanmıştır.

## 2. YAŞAM ÇÖZÜMLEMESİ

Yaşam çözümlemesi, ilgilenilen herhangi bir olayın ortaya çıkmasına kadar geçen sürenin incelenmesinde kullanılan çözümleme yöntemleri topluluğudur. Geçen süre; araştırmanın başından ilgilenilen olay gerçekleşinceye kadar geçen yılları, ayları, haftaları, günleri ya da olay gerçekleştiği anda birimin yaşını ifade eder. Olay ise ölüm, hastalığa yakalanma, hastalığın kötüye gitmesi, iyileşme, işe dönme ya da birimin başına gelebilecek ilgilenilen herhangi bir olayı ifade eder (Kleinbaum & Klein, 2005).

Yaşayan bir organizmanın ya da cansız bir nesnenin belirli bir başlangıç zamanı ile başarısızlığı arasında geçen zamana "yaşam süresi" ya da "başarısızlık süresi" adı verilmektedir. Her bir birime ait yaşam süresi, tanımı gereği sürekli ve pozitif bir değere sahiptir (Elandt-Johnson & Johnson, 1980).

Yaşam çözümlemesinde kullanılan modellerin diğer istatistiksel modellerden temel farkı zaman değişkeninin yapısından dolayı durdurulmuş (censored) gözlem içeren veri kümesi için tasarlanmış istatistiksel yöntemler bütünü olmasıdır. Durdurma genel olarak, çalışmanın bitiş noktasına kadar olayın gözlenememesi (administrative censoring), çalışma bitmeden birimle ilgili bilgi alınamaması (lost to follow- up), başka bir olayla (başka nedenden ölüm, ilaç reaksiyonu gibi) karşılaşma (withdrawing) nedenleri ile meydana gelir**.** Bu gibi nedenlerle birimler daha fazla gözlemlenemez ve bu birimler "durdurulmuş gözlem" olarak ifade edilir (Kul, 2010).

### *2.1. Cox Orantılı Tehlikeler Modeli*

Yaşam çözümlemesinde en çok kullanılan model Cox orantılı tehlikeler modelidir. 1972 yılında Cox tarafından geliştirilen regresyon modeli ile yaşam çözümlemesinde önemli adımlar atılmış, Cox (1972)'un önerileri, Kalbfleisch ve Prentice (1980)'in katkıları ile bugünkü önemini kazanmıştır. Cox orantılı tehlikeler modeli; bir birimin yaşam süresi ile birden fazla açıklayıcı değişken arasındaki ilişkiyi ortaya çıkaran istatistiksel bir yöntemdir (Cox & Oakes, 1984).

Cox orantılı tehlikeler modeli, Cox modeli veya Cox regresyon modeli (Cox regression model) olmak üzere farklı şekillerde adlandırılabilen, dağılım bilgisi gerektirmeyen bir modeldir. Bu modelde, yaşam süresi ve bu süre üzerinde etkili olarak görülen açıklayıcı değişkenler yer alır. Açıklayıcı değişkenler, modeli toplamsal değil, çarpımsal olarak etkiler (Hosmer & Lemeshow, 1999).

Cox orantılı tehlikeler modeli,

$$h(t, \boldsymbol{X}) = h_0(t)\exp(\boldsymbol{\beta' X})$$

biçimindedir. Burada, $\boldsymbol{\beta}$ regresyon katsayıları vektörü, **X** açıklayıcı değişkenler vektörü ve $h_0(t)$ ise açıklayıcı değişkene sahip olmayan (**X=0** olan) bir birimin temel tehlike fonksiyonu olarak tanımlanmaktadır (Ata, Sertkaya & Sözer, 2007).

Cox orantılı tehlikeler modelinin temel varsayımı orantılı tehlikelerdir. Orantılı tehlikeler varsayımı, tehlikelerin oranının zamana karşı sabit olması anlamına gelmektedir. Orantılı tehlikeler varsayımını incelemek için grafiksel ya da sayısal yöntemler kullanılmaktadır. Orantılı tehlikeler varsayımının incelenmesinde en çok kullanılan yöntemler, log(-log) yaşam eğrileri, gözlenen ve beklenen yaşam eğrileri, Arjas grafikleri, modele zamana bağlı açıklayıcı değişkenlerin eklenmesi, Schoenfeld artıkları ile yaşam süresinin rankı arasındaki korelasyon testi biçiminde sıralanabilmektedir (Ata vd., 2007; Therneau & Grambsch, 2000).

### *2.2. Aykırı Değerler*

Eldeki veri kümesine bir model uyarlandıktan sonra, uygulanan modelin varsayımlarının sağlanıp sağlanmadığının kontrol edilmesi gereklidir (Collett, 1994).

Artıklar en yaygın kullanılan tanı yöntemleridir. Eğer uygulanılan model uygunsa artıklar için çizilen grafiklerde beklenilmeyen örüntüler oluşmaz. Doğrusal

regresyondaki artık değerleri en basit haliyle tahmin edilen değer ile gerçek değer arasındaki farktan hesaplanır. Bu durumda grafiklerin sıfır etrafında rasgele saçılımlı olması beklenir (Stepanova & Thomas, 2002).

Regresyon verilerinde aykırı değerler literatürde tartışılan bir problemdir. Aykırı değerler değişik nedenlerle ortaya çıkabilir. Örneğin, kaba (gross) hatalardan kaynaklanabilir. Bu hatalar kopyalama ya da veri girişi hataları, hatalı ondalık noktalama, çalışmadan elde edilen ölçümleri hatalı ölçeklendirme, farklı anlamdaki iki veriyi karıştırma, farklı kitleden hatalı alınan bir gözlem, geçici etkiler ya da donanım yetersizliği gibi hatalar olabilir. Bu tür hatalar, iyi koşullar altında özel ilgiyle gözlendiklerinden yüksek nitelikli verilerde nadiren rastlanır. Hampel v.d. (1986) verilerin tipik olarak %1-10 oranında bu tür hataya sahip olduklarını belirtmişlerdir (Noh, 2010).

### *2.3. Yaşam Çözümlemesinde Aykırı Değerler*

Yaşam çözümlemesinde verinin kendine has özelliklerinden dolayı birçok yazar aykırı değere özel bir anlam vermeyi denemiştir. Collet (1994) yaşam çözümlemesinde aykırı değerlerin, son derece uzun yaşam süresine sahip birimler olduğunu belirtmiştir (Collett, 1994). Therneau, Grambcsh ve Fleming (1990) ve Nardi ve Schemper (1999) aykırı değerleri çok erken ölen ya da çok uzun yaşayan birimlerle ilişkilendirmişlerdir (Therneau vd., 1990; Nardi & Schemper, 1999). Nardi ve Schemper (1999) daha sonradan yaşam çözümlemesinde aykırı değerleri "çok uzun yaşayan" ya da "çok erken ölen" birimler olarak belirtmişlerdir (Nardi & Schemper, 1999).

Yaşam çözümlemesi çalışmalarında toplanmış veriler aykırı değerler içerebilir. Aykırı değerler genellikle modele iyi uyum sağlamayan değerler olarak tanımlanır. Bu aykırı değerler, model tahminine göre "çok geç" ya da "çok erken" başarısız olan birimler olmasına göre büyük pozitif ya da negatif artık değerleri alabilir. Birimler uzun yaşam süresine sahip olabilir ama açıklayıcı değişkenlerin değerleri birimin daha erken ölmesi gerektiğini gösterebilir.

Yaşam çözümlemesinde aykırı değerler, modeldeki parametre tahminlerini etkileyebilir, tehlike oranını, seçilen modeli değiştirebilir ve modele dayanan tahminleri etkileyebilir. Bu aykırı değerler etkili gözlemler olarak tanımlanır. Etkili gözlemlere genelde uzun yaşam sürelerinde rastlanılmaktadır. Yaşam çözümlemesinde veri kümesindeki etkili gözlemlerin ve aykırı değerlerin ortaya çıkışının araştırılması oldukça önemlidir (Noh, 2010).

Yaşam çözümlemesinde, başlıca artık türleri; Cox-Snell artıkları, Martingale artıkları, Sapma artıkları, Schoenfeld artıkları, Log-odds artıkları ve Normal sapma artıkları olarak ifade edilebilir.

Aşağıda artık türlerinde kullanılan bazı eşitlikler verilmiştir:

$t_1 < t_2 < \ldots < t_D$ başarısızlık sürelerini ve $d_i$ , $t_i$ zamanındaki başarısızlık sayılarını göstersin. Birikimli temel tehlike oranı tahmin edicisi;

$$H_0(\mathrm{t}) = \sum_{t_i \le t} \frac{d_i}{\sum_{j \in R(t_i)} \exp(\beta \mathrm{Z}_j(\mathrm{s}))} \qquad (1)$$

biçimindedir. Burada $\beta$ , tahmini regresyon katsayılarını, $Z_j$ 'ler de açıklayıcı değişkenleri göstermektedir.

$$H_0(t) = \int_0^t h_0(u)\, d_u$$

olarak verilir ve bu gözlemlenen başarısızlık sürelerinde sıçramalar ile bir adım fonksiyonudur. Burada $h_0(u)$ temel tehlike fonksiyonunu göstermektedir.

#### *2.3.1. Cox-Snell Artıkları*

Cox-Snell artıkları, yaşam verilerinin çözümlenmesinde en yaygın kullanılan artık türüdür ve Cox ve Snell tarafından önerilmiştir (Cox & Snell, 1968). Bu artıklar, modelin uygunluğunu değerlendirmek için kullanılabilir.

i. gözlem için Cox-Snell artığı,

$$r_{C_i} = H_0(t_i) \exp(\beta' Z_i) \qquad (2)$$

biçimindedir. Burada $H_0(t)$ , Eşitlik (1)'de verildiği gibidir.

Eğer seçilen model veriye uygunsa ve $\hat{\boldsymbol{\beta}}$ değerleri $\boldsymbol{\beta}$ değerlerine yakınsa $r_{C_i}$ 'ler üstel dağılımlıdır (Gharibvand & Liao, 2008). Cox-Snell artıkları modele uyumu araştırmak için çok kullanışlıdır (Tableman & Kim, 2004).

Cox-Snell artıkları, doğrusal regresyon analizinde kullanılan artıklardan biraz farklı özelliklere sahiptir. Sıfır etrafında simetrik dağılmaz, sıfır ile sonsuz arasında değer alır yani negatif olamaz, uygun model uydurulduğunda Cox-Snell artıklarının üstel dağılımlı olduğu varsayıldığından oldukça çarpık bir dağılımı vardır ve i. artığın ortalaması ve vayansı bir olur. Diğer bir nokta ise eğer en büyük yaşam süresi durdurulmamış ise $r_{C_i}$ bu gözlem için tanımsızdır (Collett, 1994).

#### *2.3.2. Martingale Artıkları*

Martingale artıkları, Barlow ve Prentice tarafından önerilmiştir (Barlow & Prentice, 1988).

Zamana bağlı olmayan açıklayıcı değişkenlerle Cox modeli için, i. gözlemin $t_i$ zamanındaki değeri ve $\delta_i$ olay durumu ( $\delta_i = 0$ iken $t_i$ durdurma süresi ve $\delta_i = 1$ iken $t_i$ olay süresi) ise Martingale artığı,

$$r_{m_i} = \delta_i - H_0(t_i)\exp(\beta' Z_i) = \delta_i - r_{C_i} \qquad (3)$$

biçimindedir.

Eşitlik (3)'te görüldüğü gibi Martingale artıkları Cox-Snell artıklarının doğrusal dönüşümüdür.

Martingale artıkları ve dönüşümleri model belirlemede kullanılabilir. Açıklayıcı değişkenlere karşı çizilen Martingale artıkları grafiği modele dahil edilen açıklayıcı değişkenlerin fonksiyonel formunu belirlemede kullanılır. Modelin uygunluğu için değişkenlerin herhangi bir dönüşüme ihtiyacı olup olmadığını gösterir. En uygun biçimi belirlemek için model kurulduktan sonra Martingale artıklarına karşı değişkenlerin istenilen dönüşümlerinin grafiği çizdirilir. Eğer dönüşüm uygun ise grafikteki eğri yaklaşık olarak doğrusal olur. Ayrıca aykırı değerleri belirlemede de kullanılır (Lin & Ying, 1993).

Martingale artıkları $\sum_{i=1}^{n} r_{m_i} = 0$ özelliğine sahiptir (Gharibvand & Liao, 2008). Büyük n için, $r_{m_i}$ 'ler sıfır ortalamalı kitleden ilişkisiz örneklerdir. Martingale artıkları sıfır etrafında simetrik dağılmaz, çarpıktır (Collett, 1994; Tableman & Kim, 2004). Martingale artıkları $-\infty$ ve 1 arasında değer alır ve durdurulmuş gözlemler ( $\delta_i = 0$ ) için negatif değerlidir.

Martingale artıkları bir değerinin yakınlarında değer alıyorsa beklenenden daha kısa yaşam süresi, büyük negatif bir değer alıyorsa beklenenden daha uzun yaşam süresi olduğu anlamına gelir (Collett, 1994).

*2.3.3. Sapma Artıkları*

Sapma artıkları Therneau, Grambch ve Fleming (1990) tarafından önerilmiştir (Therneau vd., 1990). Sapma artıkları olan $r_{d_i}$ 'ler Martingale artıklarından dönüştürülmüştür ve

$$r_{d_i} = sign(r_{m_i})\sqrt{2[-r_{m_i} - \delta_i \log(\delta_i - r_{m_i})]} \qquad (4)$$

biçimindedir.

Eşitlik (4)'te görüldüğü gibi, simetrik bir dağılım elde etmek için Martingale artıklarının dönüşümüyle elde edilir. Sonuç olarak sapma artıkları sıfır etrafında simetrik dağılır ve yaklaşık olarak 1 standart sapmaya sahiptir (Gharibvand & Liao, 2008; Gharibvand & Fernandez, 2008).

$sign(r_{m_i})$ Martingale artıklarının işaretini göstermektedir. Bu nedenle sapma artıkları Martingale artıkları ile aynı işarete sahiptir (Noh, 2010).

Grafiklerde, potansiyel aykırı değerler büyük mutlak değer sapma artıklarına karşılık gelir.

Durdurma yüzde yirmi beşten az oranda ya da yakın bir değerde ise, bu artıklar sıfır etrafında simetriktir ya da normal dağılıma oldukça yakındır. Yüzde kırktan daha fazla orandaki durdurma için ise sıfır etrafındaki artıklarla geniş noktalar kümesi normallik yaklaşımını bozar (Therneau vd., 1990; Tableman & Kim, 2004).

Martingale artıkları model uygun olsa bile çarpıktır ve bu çarpıklık artık grafiklerinin yorumlanmasını zorlaştırır. Sapma artık grafiklerinin yorumlanması ise daha kolaydır. Böylece bu artıklar diğer artıklardan aşırı derecede farklı olan yaşam sürelerine sahip birimlerin belirlenmesinde grafiksel bir araç olarak kullanılabilir. Birçok araştırmacı aykırı değerleri belirlemede sapma artık grafiklerini kullanmışlardır (Noh, 2010).

Sapma artıkları beklenenden daha uzun yaşam süresi olan gözlemler için negatif iken beklenenden daha kısa yaşam süresi olan gözlemler için pozitiftir. Çok büyük ya da çok küçük değerler olması bu değerlerin aykırı değer olduklarının göstergesi olabilir. Bu nedenle bu değerler dikkatle incelenmelidir (Gharibvand & Fernandez, 2008).

Başarısız olan gözlemlerde uyum yeterli ise sapma artıkları ak gürültüye benzer biçimde dağılır. Durdurulmuş gözlemler için ise sapma artıkları sıfır yakınlarında küme olarak yer alır (Klein & Moeschberger, 2003).

*2.3.4. 2.3.4. Schoenfeld Artıkları*

Cox orantılı tehlikeler modelinde kullanılan Cox-Snell, Martingale ve sapma artıklarının iki dezavantajı söz konusudur. Bu dezavantajlar, artıkların ağırlıklı olarak gözlenen yaşam süresine bağlı olmaları ve birikimli tehlike fonksiyonunun tahminini gerektirmesidir. Schoenfeld tarafından önerilen, skor artıkları olarak da adlandırılan Schoenfeld artıklarında bu sorunlar giderilmiştir (Schoenfeld, 1982). Bu yönüyle Schoenfeld artıkları, diğer artıklardan önemli bir farklılık göstermektedir. Bu artıklarda, her birimin artığı için tek bir değer yerine, tahmin edilmiş olan Cox orantılı tehlikeler modelinde yer alan her bir açıklayıcı değişken için birer tane olmak üzere değerler kümesi yer almaktadır (Collett, 1994; Yay, Çoker & Uysal, 2007).

Schoenfeld artıkları, değişkenin gerçek değeri ile ağırlıklı risk skorlarının ortalaması arasındaki farktır.

i. birim için Schoenfeld artığı,

$$S_i = X_i - \frac{\sum_{j \in R(t_i)} X_j \exp(\beta' X_j)}{\sum_{j \in R(t_i)} \exp(\beta' X_j)}$$

biçiminde tanımlanır. Burada $S_i$ , p açıklayıcı değişken sayısı olmak üzere, px1 boyutlu bir vektördür ve $S_i = (S_{i1}, ..., S_{ip})'$ şeklindedir. i. birim ve k. açıklayıcı değişken için Schoenfeld artıkları,

$$\hat{S}_{ik} = X_{ik} - \frac{\sum_{j \in R(t_i)} X_{jk} \exp(\beta' X_j)}{\sum_{j \in R(t_i)} \exp(\beta' X_j)}$$

biçimindedir. Burada $X_j$ j. birim için p sabit açıklayıcı değişken vektörüdür. $X_{jk}$ ise j. birimin k. açıklayıcı değişkenin değeridir. Bu nedenle, bu artık, $X_{jk}$ 'nın gözlenen değeri ile $t_i$ zamanında risk altındaki birimler üzerinden açıklayıcı değişken değerlerinin ağırlıklı ortalaması arasındaki farktır. $t_i$ zamanındaki risk kümesinde k. açıklayıcı değişkenli j. birim için kullanılan ağırlık,

$$\frac{\exp(\beta' X_j)}{\sum_{j \in R(t_i)} \exp(\beta' X_j)}$$

biçimindedir. Bu ifade, kısmi olabilirliği en büyüklemeye bu birimin katkısıdır.

Schoenfeld artıkları durdurulmamış gözlemler için tanımlıdır ve ayrıca her bir açıklayıcı değişkenin Scohenfeld artıkları toplamı sıfır olmalıdır (Gharibvand & Liao, 2008).

Schoenfeld artıkları sıfırda toplanır. İki düzeyli (0,1) değişkenler için bu artıklar -1 ile 1 arasında değer alır. Bu nedenle, artık grafiğinde iki kuşak olur; bir tanesi x=1 için sıfırın üstünde ve diğeri x=0 için sıfırın altında yer alır.

Orantılı tehlikeler varsayımını incelemek için Schoenfeld artıklarına dayanan bir test geliştirilmiştir. Belirli bir değişken için Schoenfeld artıkları ile birimlerin yaşam sürelerinin rankı arasındaki korelasyon kullanılarak orantılı tehlikeler varsayımı incelenebilir. Bu teste göre, orantılı tehlikeler varsayımının sağlanması için korelasyonun sıfıra yakın olması beklenmektedir (Ata vd., 2007).

Orantılı tehlikeler varsayımının geçerliliği zamana bağlı çizilen Schoenfeld artıkları grafiği ile de kontrol edilebilir. Çizilen Schoenfeld artıkları grafiği, yatay bir doğru etrafında seyrediyorsa, orantılı tehlikeler varsayımının sağlandığı söylenilir (Schoenfeld, 1982).

Hosmer ve Lemeshow regresyon katsayılarının kovaryans matrisine dayalı ölçeklendirilmiş Schoenfeld (scaled Schoenfeld) artıkları grafiğinin orantılı tehlikeler varsayımı için kullanılmasını önermiştir (Hosmer & Lemeshow, 1999).

Ölçeklendirilmiş Schoenfeld artıkları;

$$r_{kt}^{*} = m \sum_{i=1}^{p} V_{ik} S_{ik}$$

biçimindedir. Burada m toplam başarısız birim sayısını, V ise regresyon katsayılarından tahmin edilmiş kovaryans matrisini göstermektedir.

Ölçeklendirilmiş Schoenfeld artıkları grafiği zamana karşı her bir açıklayıcı değişken için çizilir. Ölçeklendirilmiş Schoenfeld artıkları etkili gözlemleri bulmak için kullanılır.

Açıklayıcı değişkenler sürekli olduğunda ölçeklenmiş Schoenfeld artıklarının kullanılması önerilmiştir (Terzi & Bek, 2005; Winnetti & Miscellanea, 2001).

*2.3.5. Log-Odds Artıkları*

Log-odds artıkları, Nardi ve Schemper (1999) tarafından verilmiş ve $L_i$ ile gösterilmiştir (Nardi & Schemper, 1999). Log-odds artıklarının dağılımı $E(L_i) = 0$ ortalama ve $V(L_i) = (\pi^2 / 3)$ varyans ile lojistik dağılımdır ve

$$L_i = \log\left[ S_i(T_i) / \left\{1 - S_i(T_i)\right\} \right]$$

biçimindedir. Bilinmeyen yaşam fonksiyonu yerine tahmin edicisi alındığında $L_i$ , $L_i$'ye yakınsar. Gözlenen yaşam süreleri $t_i \quad (1 \le i \le n)$ olduğunda, $L_i$ 'nin gözlenen değeri,

$$l_i = \log\left[ S_i(t_i) / \left\{1 - S_i(t_i)\right\} \right]$$

olur.

Durdurulmuş süre için $S_i\left(t_i^c\right)$ yaşam olasılığı bilinmemektedir. Dolayısıyla durdurulmuş yaşam süresi artıklarını uyarlamak için, $S_i(t_i^c)$ ile koşullu ortanca değeri $\frac{S_i(t_i^c)}{2}$ 'nin yer değiştirmesi önerilmiştir. Bu durumda $l_i^m$ log-odds artık değerleri,

$$l_i^c = \log\left[ S_i\left(t_i^c\right) / \left\{ 2\text{-}S_i\left(t_i^c\right) \right\} \right]$$

biçiminde elde edilir. $L_i$ 'nin koşullu dağılımından beklenen ortalama değerlerin sapması kullanıldığında ise daha karmaşık olmakta ve

$$l_i^m = l_i^c - \frac{1+\exp(l_i^c)}{\exp(l_i^c)} \log\left\{ 1+\exp(l_i^c) \right\}$$

biçiminde elde edilmektedir.

Son derece kısa yaşam süreleri için $l_i^m$ ( ya da $l_i^c$ ) sıfıra yakın olur ve beklenen $S_i\left(t_i\right) \cong 0.5$ olduğu varsayılır.

Log-odds artıkları lojistik dağılımlı olduklarından aykırı değerler bu dağılımlara dayanan kesim noktası ile belirlenebilir.

Kesim noktaları "çok erken ölen"=ÇEÖ ve "çok uzun yaşayan"=ÇUY ile gösterilirse log-odds artıkları için kesim noktaları,

$$R_{ÇEÖ,\mathrm{L}} = \left\{ l_i : \mathrm{l}_i > w_{1-\alpha} \right\}$$

$$R_{ÇUY,\mathrm{L}} = \left\{ l_i : \mathrm{l}_i < w_{\alpha} \right\}$$

biçimindedir. Burada $w_\alpha$ lojistik dağılımda $\alpha$ güven düzeyine karşılık gelen noktadır. Eğer birimlerin artık değerleri kesim noktalarını aşarsa bu gözlemler aykırı değerler olur (Nardi & Schemper, 1999; Nardi & Schemper, 2003).

*2.3.5. Normal Sapma Artıkları*

Normal sapma artıkları Nardi ve Schemper (1999) tarafından verilmiş ve $N_i$ ile ifade edilmiştir (Nardi & Schemper, 1999). Artıkların dağılımı standart normal dağılımdır.

Bilinmeyen yaşam fonksiyonu yerine tahmin edicisi alındığında $N_i$ 'nin olasılıkta yakınsaması $N_i$ ' dir. $\Phi$, normal birikimli dağılım fonksiyonudur.

Böylece,

$$N_i = \Phi^{-1}\left\{ S_i(T_i) \right\}$$

ve artık değerleri

$$n_i = \Phi^{-1}\left\{ S_i(t_i) \right\}$$

biçimindedir. Burada durdurulmuş süreler için $S_i\left(t_i^c\right)$ yaşam olasılığı bilinmemektedir.

Durdurulmuş yaşam sürelerinin artıklarının uyarlanması (accommodating) için çeşitli yollar vardır. Gerçek yaşam süresi, gözlenen durdurulmuş yaşam süresinden daha uzundur ve bilinmeyen doğru artıkların dağılımı $S_i(T_i)$ 'nin $[\,0,\ S_i\left(t_i^c\right)\,]$'de uniform dağılımı ile ilişkilidir. Böylece, $S_i\left(t_i^c\right)$, koşullu ortanca değeri $\frac{S_i\left(t_i^c\right)}{2}$ ile yer değiştirir. Sonuç olarak durdurulmuş süre için $N_i$,

$$n_i^c = \Phi^{-1}\left\{ \frac{S_i\left(t_i^c\right)}{2} \right\}$$

biçimindedir ve burada $n_i^c$, durdurulmuş süre $N_i$ 'nin ortalamasıdır ya da

$$n_i^m = -\frac{\exp(0.5(n_i^c)^2)}{\sqrt{2\pi}\ S_i(t_i^c)}$$

ile değiştirilmiş olabilir. Burada $n_i^m$ ,durdurulmuş süre $N_i$ 'nin ortancasıdır (Nardi & Schemper, 1999; Nardi & Schemper, 2003).

Durdurulmuş gözlemler için normal sapma artıkları koşullu ortalama ya da ortanca değeriyle yer değiştirir. Ancak bu durum artıkların anormal bir yığılma göstermesine neden olur. Böylece iyi uyum sağlayan model bile normal dağılımdan farklı bir dağılım gösterir (Nardi & Schemper, 2003).

Log-odds artıklarında olduğu gibi normal sapma artıklarında da son derece kısa yaşam süreleri için $n_i^m$ (ya da $n_i^c$ ) sıfıra yakın olur ve beklenen $S_i\left(t_i\right) \cong 0.5$ olduğu varsayılır (Nardi & Schemper, 1999).

Normal sapma artıkları normal dağılımlı olduklarından aykırı değerler bu dağılımlara dayanan kesim noktası ile belirlenebilir.

Normal sapma artıklarının kesim noktaları,

$$R_{ÇEÖ,N} = \left\{ n_i : n_i > z_{1-\alpha} \right\}$$

$$R_{ÇUY,N} = \left\{ n_i : n_i < z_{\alpha} \right\}$$

biçimindedir. Burada $z_\alpha$ standart normal dağılımda $\alpha$ güven düzeyine karşılık gelen noktadır. Aynı referans dağılıma sahip olduklarından benzer kesim noktaları sapma artıkları için de kullanılabilir (Nardi & Schemper, 1999; Nardi & Schemper, 2003).

### *2.4. Etkili Gözlemlerin Belirlenmesi*

Cox orantılı tehlikeler modelinde aykırı değerler üç madde altında incelenebilir:

- Örneklem ortalamasından büyük oranda farklı bir açıklayıcı değişken değerine sahip olanlar,
- Parametre tahminlerinde güçlü etkiye sahip olanlar,
- Kısmi olabilirlik fonksiyon değerinde ve böylece model yeterliğinde güçlü etkiye sahip olanlar.

İlk madde skor artıkları kullanılarak belirlenebilir. Model kestiriminden sonra, her gözlem için skor artıkları ilgili açıklayıcı değişken bakımından hesaplanır ve analiz edilen ortak değişkene karşı artıkların değerlerini gösteren bir grafik çizilir. Aykırı değerler bu tip analizlerle kolaylıkla belirlenebilir. Aykırı değerlerin belirlenmesinden sonraki aşama şüpheli gözlemlerin parametre tahminlerindeki etkilerinin şiddetinin tahmin edilmesidir. Şüpheli gözlem örneklem dışında bırakılarak model yeniden tahmin edilir. Parametre tahminindeki değişim,

$$\Delta\beta_{ki} = \beta_k - \beta_{k(-i)}$$

biçiminde hesaplanır. Burada $\beta_k$, k. açıklayıcı değişkenin modelin tüm örneklemdeki parametre kestirimidir ve $\beta_{k(-i)}$ ise i. gözlem çıkarıldıktan sonra hesaplanan örneklemdeki benzer bir değerdir. $\Delta_i$ k. elemanının vektörü ile iyi yaklaştırılmış olduğu kanıtlanmıştır ve

$$\Delta\beta_i = V(\beta)L_i$$

biçimindedir. Burada $L_i$ i. gözlem için skor artıkları vektörü ve $V(\beta)$ parametre tahminlerinin bir varyans-kovaryans matrisidir. Bu tanım ölçeklendirilmiş skor artıkları yada DFBETA artıkları olarak adlandırılır. Analiz edilen açıklayıcı değişkene karşı bu tip artıkların grafiği etkili gözlemlerin saptanmasında yararlıdır.

Eğer bu fark sıfıra yakın ise i. gözlemin tahmindeki etkisi çok küçüktür. Asıl modeldeki tüm değişkenler için bu süreç tekrarlanır. N gözlemli büyük veri setleri için bu farkları hesaplamak yani bu işlemi N kez tekrarlamak pratik bir yöntem değildir.

Bu gözlemlerin kısmi olabilirlik fonksiyonu değerindeki etkilerini tahmin etmek için skor artıkları,

$ld_i = \Delta\beta' V(\beta_i)^{-1}\Delta\beta_i = L_i' V(\beta)V(\beta)^{-1}V(\beta)L_i = L_i' V(\beta)L_i$

biçimindedir.

Bu istatistikler, olabilirlik değişim (likelihood displacement) olarak adlandırılır ve örneklemdeki i. gözlemin örneklem dışında bırakılmasından sonra kısmi olabilirlik fonksiyonunun logaritmasındaki değişiklik ile ilgilenir.

Olabilirlik değişim değerleri i. gözlemin etkisini bu gözlem çıkarıldığında log olabilirlik modelindeki değişim yaklaşımıyla hesaplar.

i. değişken için olabilirlik değişim değeri,

$$2\left\{\log L(\beta) - \log L(\beta_i)\right\}$$

biçimindedir. Burada $\beta$ tüm model için hesaplanan tahmini ve $\beta_i$ i. gözlem çıkarıldıktan sonra hesaplanan tahmin değerini gösterir. L(.) tüm veriden tahmin edilen kısmı olabilirlik değeridir. L(.) hesaplanırken tüm veri kullanılır ama $\beta_i$ parametre tahminleri i. gözlem çıkarılarak elde edilir. $\beta$ uygun bir çözüm verir ve olabilirlik değişimi hiçbir zaman negatif değildir.

Ayrıca $L_i' V(\beta)L_i$ matrisi için özdeğerler bulunur. Yüksek özdeğerler ile ilişkili özdeğerler LMAX istatistikleri olarak adlandırılır. Olabilirlik değişim de LMAX da özet istatistiklere (örneğin Martingale artıkları) karşı çizilir. Yüksek derecede etkili gözlemler bu tarz grafiklerle kolayca belirlenebilir.

Etkili gözlemlerin belirlenmesinde kullanılan iki alternatif yöntem LMAX ve olabilirlik değişim değerleri, DFBETA değerlerinin aksine gözlemlerin etkisini bütün olarak katsayılar vektöründen ölçer. Böylece çok değişkene karşı sadece tek bir değer elde edilir.

Artıkların kullanımı veriye ve araştırmacıya bağlı olsa da geleneksel ve tavsiye edilen kullanım şekilleri vardır. Bunlar;

- Cox-Snell artıkları modele uyumu araştırmak için kullanılır.
- Martingale artıkları modele dahil edilen açıklayıcı değişkenlerin fonksiyonel formunu belirlemede ve bazen de aykırı değerleri belirlemede kullanılır.
- Sapma artıkları modelin doğruluğunu test etmede ve aykırı değerleri belirlemede kullanılır.
- Schoenfeld ve ölçeklendirilmiş Schoenfeld artıkları orantılı tehlikeler varsayımının kontrolünde kullanılır.
- Normal sapma ve log-odds artıkları aykırı değerleri belirlemede kullanılır.
- Olabilirlik değişim ve Lmax değerleri etkili gözlemleri belirlemede kullanılır. DFBETA değerleri

de etkili gözlemleri belirlemede kullanılır ancak işlemleri değişkenler bazında teker teker yapar.

## 3. UYGULAMA

Ankara Onkoloji Hastanesi'nde Ocak 1990 ve Kasım 1995 tarihleri arasında mide kanseri tanısı konulan ve cerrahi tedavi geçiren, yaşları 29 ile 84 arasında değişen 118 hastaya ait veriler Eroğlu v.d. tarafından incelenmiş ve Kaplan-Meier yöntemiyle yaşam olasılıkları bulunmuş, Cox orantılı tehlikeler modeli elde edilmiştir (Eroğlu, Altınok, Özgen & Sertkaya, 1997).

Bu çalışmada ise bu veriler, yaşam çözümlemesinde aykırı değerleri belirlemede kullanılan artık türlerinin uygulama alanlarını göstermek ve elde edilen sonuçları karşılaştırabilmek amacıyla kullanılmıştır.

Analizler için STATA 12 programı deneme sürümü kullanılmıştır.

Açıklayıcı değişkenler olarak yaş, cinsiyet, kilo kaybı, anemi, tümörün midedeki lokalizasyonu, lenf nodu diseksiyonunun genişliği, hastalığın evresi ve adjuvan kemoterapi alınmıştır. Bu hastaların yaş ortalaması 56.70'dir ve %58.5'i erkektir. Hastaların %55.1'i hastalığın 3. evresinde ve %19.5'i 4. evresindedir. Hastaların %71.2'si kemoterapi almıştır. Hastaların 52 tanesinde başarısızlık gözlenmiştir. Başarısızlık, ölüm olarak alınmıştır. Açıklayıcı değişkenlerin düzeyleri ve sıklıkları Çizelge 1'de verilmiştir.

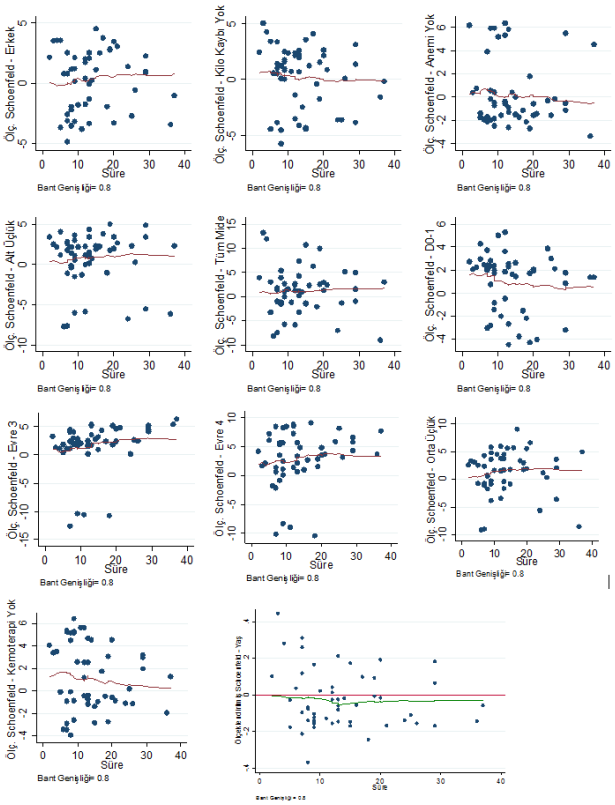
Çizelge 1. Açıklayıcı değişkenler ve düzeyleri

| Değişkenler | Düzeyler | Sıklıklar (%) |
|---|---|---|
| Cinsiyet | 1. Erkek | 69 (58.5) |
| | 2. Kadın | 49 (41.5) |
| Kilo Kaybı | 1. Yok | 90 (76.3) |
| | 2. Var | 28 (23.7) |
| Anemi | 1. Yok | 31 (26.3) |
| | 2. Var | 87 (73.7) |
| Lenf nodu diseksiyonunun genişliği (Diseksiyon) | 1. D0-1 | 62 (52.5) |
| | 2. D2-3 | 56 (47.5) |
| Tümörün midedeki lokalizasyonu (Lokal) | 1. Üst üçlük | 21 (17.8) |
| | 2. Orta üçlük | 25 (21.2) |
| | 3. Alt üçlük | 62 (52.5) |
| | 4. Tüm mide | 10 (8.5) |
| Hastalığın evresi (Evre) | 1. Evre1+Evre2 | 30 (25.4) |
| | 2. Evre3 | 65 (55.1) |
| | 3. Evre4 | 23 (19.5) |
| Adjuvan kemoterapi (Kemoterapi) | 1.Yok | 34 (28.8) |
| | 2. Var | 84 (71.2) |
| Yaş | 56,403 ±1,044 | |

Yaşam süresine değişkenlerin etkisi araştırılmak istenildiğinde Cox orantılı tehlikeler modelinin kullanılabilmesi için değişkenlerin orantılı tehlikeler varsayımını sağlaması gerekmektedir. Orantılı tehlikeler varsayımın sağlanıp sağlanmadığı araştırılmıştır.

Orantılı tehlikeler varsayımını incelemek için Schoenfeld artıkları kullanılabilir. Schoenfeld artıkları, açıklayıcı değişkenin gerçek değeri ile ağırlıklı risk skorlarının ortalaması arasındaki farktır.

Süreye karşı çizdirilen grafikte eğri sıfır etrafında yaklaşık olarak doğrusal ise varsayımın sağlandığı sonucuna ulaşılır.




**Şekil 1.** Açıklayıcı değişkenler için ölçeklendirilmiş Schoenfeld artığı grafikleri

Şekil 1 incelendiğinde grafiklerde eğrinin yaklaşık olarak doğrusal olduğu yani değişkenlerin orantılı tehlikeler varsayımını sağladığı görülmektedir.

Orantılı tehlikeler varsayımını incelenmek için kullanılan diğer bir yöntem olan, yaşam süresi rankının Schoenfeld artıkları ile ilişkisi de incelenmiştir. Çizelge 2 incelendiğinde tüm değişken düzeylerinde p-değeri>0.05 olduğu için orantılı tehlikeler varsayımının sağlandığı görülmüştür.

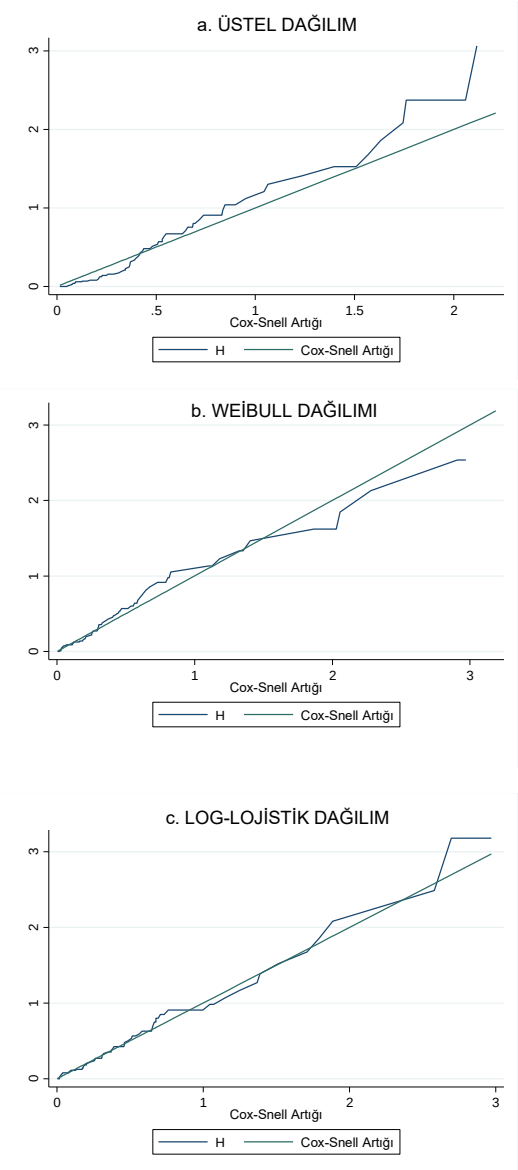
**Çizelge 2.** Yaşam süresi rankının Schoenfeld artıkları ile ilişkisi

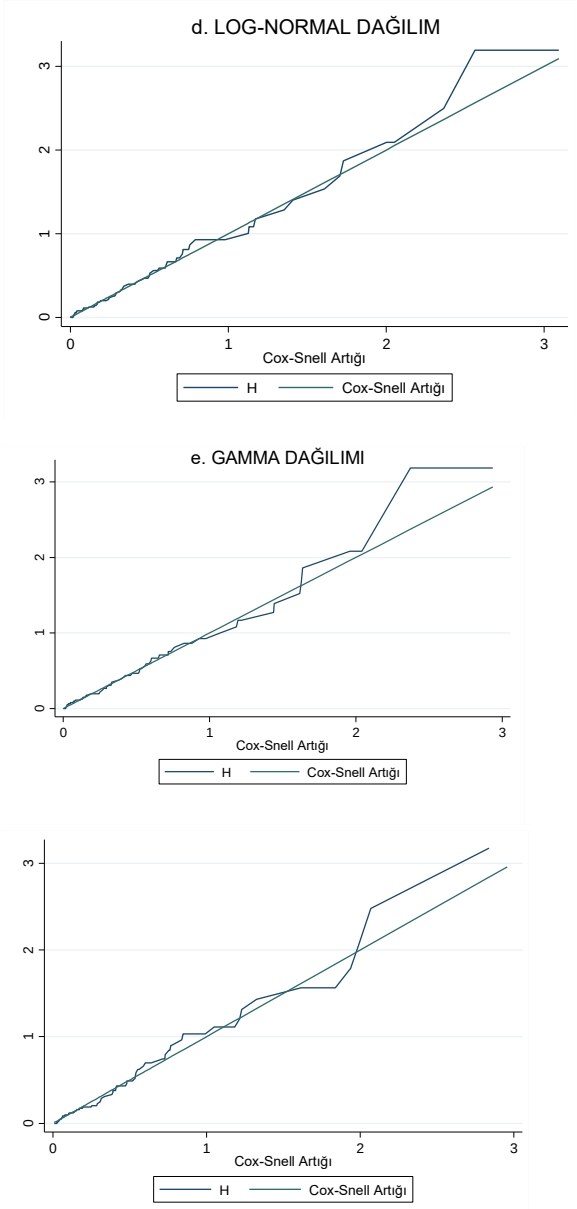
| Değişken | p-değeri |
|---|---|
| Yaş | 0,21 |
| Cinsiyet- Erkek | 0,6 |
| Kilo kaybı- Yok | 0,21 |
| Anemi- Yok | 0,55 |

| Değişken | p-değeri |
|---|---|
| Diseksiyon- D0-1 | 0,09 |
| Lokalizasyon-Orta Üçlük | 0,37 |
| Lokalizasyon-Alt Üçlük | 0,79 |
| Lokalizasyon-Tüm mide | 0,81 |
| Evre- Evre3 | 0,08 |
| Evre- Evre4 | 0,19 |
| Kemoterapi- Yok | 0,28 |

Cox-Snell artıkları modele uyumu araştırmak için kullanılır.

Verilerin dağılımı Easyfit programı ile incelendiğinde bilinen bir dağılıma uygunluk göstermese de parametrik bir dağılım gösterdiği görülmüştür. Bu nedenle, parametrik yaşam çözümlemesi dağılımlarının uygunluğu da Cox orantılı tehlikeler modeli ile birlikte Cox-Snell artıkları grafikleri ile incelenmiştir.




Şekil 2. a-e. Parametrik modeller için ve Cox orantılı tehlikeler modeli için Cox-Snell artığı grafikleri

Model uygun ise birikimli tehlikeye karşı Cox-Snell artığı grafiği yaklaşık olarak bir eğimli bir doğru olacaktır. Şekil 2 incelendiğinde log-lojistik modelin ve log-normal modelin daha iyi sonuç verdiği söylenebilir. Bu durumda Akaike bilgi kriterine (AIC) göre karar verilmelidir. Modellere ait AIC değerleri Çizelge 3'te verilmiştir.

Çizelge 3. Modellerin karşılaştırılması

| Model | -2log(L) | AIC |
|---|---|---|
| Cox Orantılı Tehlikeler | 385.364 | 418.36 |
| Üstel | 190.56 | 214.56 |
| Weibull | 174.20 | 200.20 |
| Log-normal | 170.98 | 196.98 |
| Log-lojistik | 172.26 | 198.26 |
| Genelleştirilmiş Gamma | 170.84 | 198.84 |

Çizelge 3 incelendiğinde AIC kriterine göre log-normal modelin veriye en uygun olduğu sonucuna ulaşılmıştır. Bu nedenle veriye log-normal regresyon modeli uygulanmış ve elde edilen sonuçlar Çizelge 4'de verilmiştir.

Çizelge 4. Log-normal regresyon modelinin sonuçları

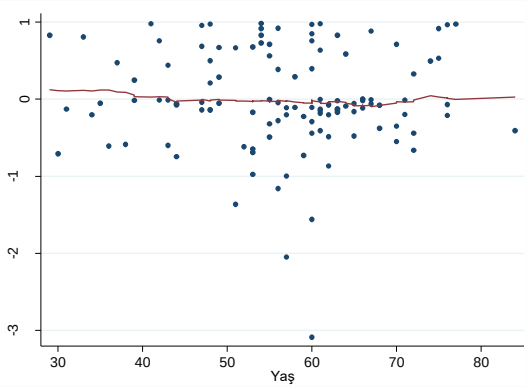
| Değişken | Tahmin | Standart Hata | *p*-değeri |
|---|---|---|---|
| Yaş | 0,0139 | 0,0084 | 0,096 |
| Cinsiyet- Erkek | -0,106 | 0,2105 | 0,616 |
| Kilo kaybı- Yok | -0,212 | 0,2253 | 0,347 |
| Anemi- Yok | -0,028 | 0,2368 | 0,907 |
| Diseksiyon- D0-1 | -0,723 | 0,2099 | 0,001* |
| Lokalizasyon- Orta Üçlük | -0,853 | 0,3169 | 0,007* |
| Lokalizasyon- Alt Üçlük | -0,482 | 0,2859 | 0,092 |
| Lokalizasyon- Tüm mide | -0,705 | 0,3791 | 0,063 |
| Evre- Evre3 | -1,262 | 0,2951 | 0,000* |
| Evre- Evre4 | -1,835 | 0,3457 | 0,000* |
| Kemoterapi- Yok | -0,835 | 0,212 | 0,000* |
| Sabit | 5,0182 | 0,6803 | 0,000* |
| Lnsigma | -0,23 | 0,1021 | 0,024* |
| Sigma | 0,7947 | 0,0811 | |
| -2log(L) | 170,995 | (p=11) | |

*p-değeri<0.05 olduğundan anlamlıdır.

Log-normal model için Martingale, sapma, log-odds ve normal sapma artık değerleri hesaplanarak aykırı değerleri tespit etmek amaçlanmıştır.

Martingale artıkları modele dahil edilen açıklayıcı değişkenlerin fonksiyonel formunu belirlemede kullanışlıdır. Eğer değişken modele uygun ise grafikteki eğri yaklaşık olarak doğrusal olur.

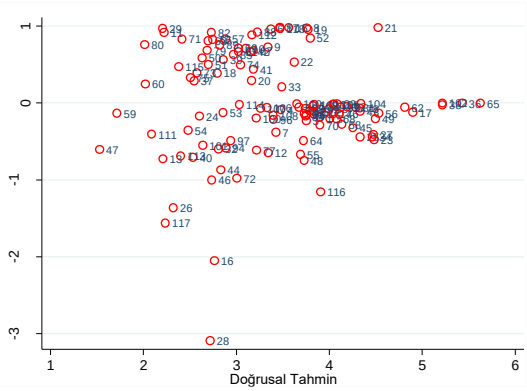
Log-normal regresyon modelinde yaş değişkeni için Martingale artığı grafiği Şekil 3'te verilmiştir.




Şekil 3. Log-normal regresyon modelinde yaş değişkeni için Martingale artığı grafiği

Şekil 3 incelendiğinde eğrinin yaklaşık olarak doğru olduğu yani yaş değişkenin modele uygun olduğu, bir dönüşüme ihtiyaç olmadığı görülür.

Martingale artıklarına karşı doğrusal tahmin grafiği aykırı değerleri belirlemede kullanılabilir. Log-normal regresyon modelinde doğrusal tahmine karşı Martingale artığı grafiği Şekil 4'te verilmiştir.

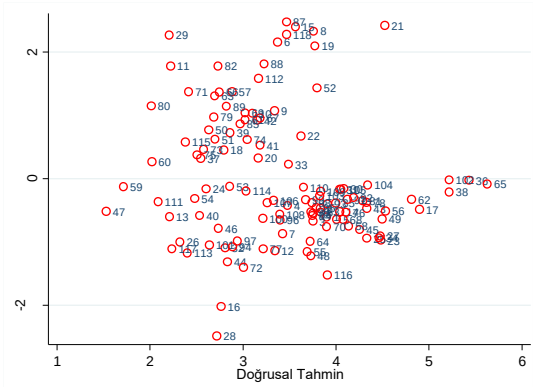



Şekil 4. Log-normal regresyon modelinde doğrusal tahine karşı Martingale artığı grafiği

Şekil 4 incelendiğinde 16., 21., 29. ve 28. gözlemlerin aykırı değerler olabilecekleri görülür.

Simetrik bir dağılım göstermesinden dolayı sapma artıklarının doğrusal tahmine karşı grafikleri aykırı gözlemlerin belirlenmesinde ve modele uyumun sağlanmasında kullanımı daha yaygındır.

Log-normal modelde doğrusal tahmine karşı sapma artığı grafiği Şekil 5'te verilmiştir.




Şekil 5. Log-normal regresyon modelinde doğrusal tahmine karşı sapma artığı grafiği

Şekil 5 incelendiğinde 6., 8., 15., 16., 19., 21., 28., 29., 87.,118. ve 128. gözlemlerin aykırı değerler olabileceği görülür. Sapma artıklarının referans dağılımı olan normal dağılımın kesim noktalarına göre karşılaştırılma yapıldığında da aynı sonuca ulaşılır.

Tüm değişkenler için log-odds ve normal sapma artıkları hesaplanmış ve hesaplanan değerler kesim noktalarıyla karşılaştırıldığında bu değerlerin dışında kalan gözlemler aykırı değer olarak belirlenmiştir. Buna göre bulunan aykırı değerler Çizelge 5'de verilmiştir.

Çizelge 5. Log-normal regresyon modelinde log-odds ve normal sapma artık değerleri

| Gözlem | Süre | Durum | $S_i(t_i)$ | $S_i(t_i^c)$ | $l_i^c$ | $l_i^m$ | $n_i^c$ | $n_i^m$ |
|---|---|---|---|---|---|---|---|---|
| 11 | 3 | Ölmüş | 0,983 | | 1,763 | 1,763 | 2,121 | 2,121 |
| 16 | 39 | Yaşıyor | - | 0,1982 | -0,607 | -1,143 | -0,848 | -2,884 |
| 17 | 51 | Yaşıyor | - | 0,1982 | -0,607 | -1,143 | -0,848 | -2,884 |
| 23 | 69 | Yaşıyor | - | 0,1982 | -0,607 | -1,143 | -0,848 | -2,884 |
| 25 | 57 | Yaşıyor | - | 0,1982 | -0,607 | -1,143 | -0,848 | -2,884 |
| 27 | 63 | Yaşıyor | - | 0,1982 | -0,607 | -1,143 | -0,848 | -2,884 |
| 28 | 58 | Yaşıyor | - | 0,1982 | -0,607 | -1,143 | -0,848 | -2,884 |
| 29 | 2 | Ölmüş | 0,991 | - | 2,068 | 2,068 | 2,387 | 2,387 |
| 34 | 65 | Yaşıyor | - | 0,1982 | -0,607 | -1,143 | -0,848 | -2,884 |
| 38 | 37 | Yaşıyor | - | 0,1982 | -0,607 | -1,143 | -0,848 | -2,884 |
| 45 | 44 | Yaşıyor | - | 0,1982 | -0,607 | -1,143 | -0,848 | -2,884 |
| 48 | 44 | Yaşıyor | - | 0,1982 | -0,607 | -1,143 | -0,848 | -2,884 |
| 49 | 44 | Yaşıyor | - | 0,1982 | -0,607 | -1,143 | -0,848 | -2,884 |
| 55 | 39 | Yaşıyor | - | 0,1982 | -0,607 | -1,143 | -0,848 | -2,884 |
| 56 | 37 | Yaşıyor | - | 0,1982 | -0,607 | -1,143 | -0,848 | -2,884 |
| 58 | 36 | Yaşıyor | - | 0,2123 | -0,569 | -1,108 | -0,798 | -2,584 |
| 62 | 34 | Yaşıyor | - | 0,2256 | -0,536 | -1,078 | -0,753 | -2,349 |
| 64 | 33 | Yaşıyor | - | 0,2256 | -0,536 | -1,078 | -0,753 | -2,349 |
| 65 | 33 | Yaşıyor | - | 0,2256 | -0,536 | -1,078 | -0,753 | -2,349 |
| 68 | 29 | Yaşıyor | - | 0,2256 | -0,536 | -1,078 | -0,753 | -2,349 |
| 70 | 29 | Yaşıyor | - | 0,2256 | -0,536 | -1,078 | -0,753 | -2,349 |
| 116 | 73 | Yaşıyor | - | 0,1982 | -0,607 | -1,143 | -0,848 | -2,884 |

Çizelge 5 incelendiğinde log-odds artık değerlerine göre aykırı değer bulunamadığını ve normal sapma artık değerlerine göre ise 11., 16., 17., 23., 25., 27., 28., 29., 34., 38., 45., 48., 49., 55., 56., 58., 62., 64., 65., 68., 70. ve 80. gözlemlerin aykırı değer olabileceği görülür.

Yöntemlerce ortak çıkan değerleri aykırı değerler olarak belirsek 16., 28. ve 29. gözlemler aykırı değerdir. Bu değerler çalışmadan çıkarılıp log-normal regresyon modeli yeniden elde edilmiş ve sonuçlar Çizelge 6'da verilmiştir.

Çizelge 6. Aykırı değerler çıkarıldıktan sonra log-normal regresyon modeli sonuçları

| Değişken | Tahmin | Standart Hata | *p*-değeri |
|---|---|---|---|
| Yaş | 0,0113 | 0,0075 | 0,130 |
| Cinsiyet- Erkek | 0,0250 | 0,1948 | 0,898 |
| Kilo kaybı- Yok | -0,2218 | 0,2038 | 0,285 |
| Anemi- Yok | 0,1483 | 0,2217 | 0,503 |
| Diseksiyon- D0-1 | -0,7043 | 0,1913 | 0,000* |
| Lokalizasyon- Orta Üçlük | -0,6854 | 0,2871 | 0,017* |
| Lokalizasyon- Alt Üçlük | -0,2456 | 0,2618 | 0,348 |
| Lokalizasyon- Tüm mide | -0,4317 | 0,3453 | 0,211 |
| Evre- Evre3 | -1,2598 | 0,2704 | 0,000* |
| Evre- Evre4 | -1,7469 | 0,3136 | 0,000* |
| Kemoterapi- Yok | -0,7758 | 0,1965 | 0,000* |
| Sabit | 4,7470 | 0,6141 | 0,000* |
| Lnsigma | -0,3389 | 0,1024 | 0,001* |
| Sigma | 0,7126 | 0,0730 | |
| -2log(L) | 153,102 | (p=11) | |

*p-değeri< 0.05 olduğundan anlamlıdır.

Çizelge 4 ve Çizelge 6 incelendiğinde, aykırı değerler çıkarıldıktan sonra kurulan log-normal regresyon modeli ile bu değerler çıkarılmadan önceki log-normal regresyon modeli karşılaştırıldığında standart hataların genel olarak azaldığı görülmüştür. İki model eşit parametre sayısına sahip olduğu için -2log(L) değerleri kullanılarak karşılaştırma yapılabilir. Aykırı değer çıkarılmadan önceki modelin -2log(L) değeri 170,995 iken çıkarıldıktan sonraki modelin -2log(L) değeri Çizelge 6'dan görüldüğü gibi 153,102'dir. Buna göre aykırı değerlerin çıkarılması sonucunda elde edilen modelin daha uygun bir model olduğu söylenebilir.

Log-normal model için adımsal seçim yöntemi uygulanırsa Çizelge 7'deki sonuçlar elde edilir.

Çizelge7. Log-normal regresyon modeli için adımsal seçim yöntemi sonuçları

| Değişken | Tahmin | Standart Hata | *p*-değeri |
|---|---|---|---|
| Dissek1 (D0-1) | -0,718 | 0,211 | 0,001 |
| Evre3 | -1,251 | 0,276 | 0,000 |
| Evre4 | -1,632 | 0,327 | 0,000 |
| Kemoterapi Yok | -0,779 | 0,214 | 0,000 |
| sabit | 4,994 | 0,318 | 0,000 |
| Lnsigma | -0,170 | 0,102 | 0,096 |
| Sigma | 0,843 | 0,863 | |
| -2log(L) | 183,184 | | |

Çizelge 7 incelendiğinde diseksiyon, evre ve kemoterapi değişkenleri yaşam süresini etkileyen faktörler olarak bulunmuştur. Lenf nodu diseksiyonunun değeri D0-1 olanların yaşam süresi D2-3 olanlara göre 2 (1/exp(-0,718)) kat daha kısa, evre 3 olanların evre 1+evre 2 olanlara göre yaşam süresi 3,5 (1/exp(-1,251)) kat, evre 4 olanların evre 1+evre 2 olanlara göre yaşam süresi 5 (1/exp(-1,632)) kat daha kısadır. Kemoterapi almayanların yaşam süresi ise alanlara göre 2 (1/exp(-0,779)) kat daha kısadır.

## 4. SONUÇLAR

Bu çalışmada, yaşam çözümlemesi ve aykırı değerler hakkında genel bilgiler, kavramlar, fonksiyonlar ve modeller verilmiş, yaşam çözümlemesinde aykırı değerleri belirlemede kullanılabilecek yöntemler incelenmiştir.

Aykırı değerlerin belirlenmesi modele uyum için oldukça önemlidir ve yaşam çözümlemesinde aykırı değerleri belirleme yöntemleri artıklara dolayısıyla artıkların analizine dayanmaktadır. Bu amaçla literatürde karşılaşılan artık türleri incelenmiştir. Bu yöntemlerin uygulaması 118 gözlemli ve 8 açıklayıcı değişkenli mide kanseri verisi üzerinde yapılmıştır.

Uygulamada ilk olarak yaşam çözümlemesinde en çok kullanılan model olan Cox orantılı tehlikeler modeli uygulanmış ve modele uyumlu olduğu görülmüştür. Cox orantılı tehlikeler modelinin temel varsayımı olan orantılı tehlikeler varsayımı da Schoenfeld artıkları ile incelenmiş ve bu varsayımın sağlandığı görülmüştür. Verilerin dağılımı bilinen bir dağılıma uygunluk göstermese de parametrik bir dağılım gösterdiği görüldüğünden parametrik yaşam çözümlemesi modelleri incelendiğinde log-normal dağılımın veriye uygun bir model olduğu sonucuna ulaşılmıştır.

Veriye uygun olduğu bulunan log-normal regresyon modeli elde edilmiştir. Log-normal regresyon modeli için de aykırı değerler belirlenmeye çalışılmıştır. Aykırı değerleri belirlemek için Martingale, sapma, log-odds ve normal sapma artıkları hesaplanmış ve grafikleri çizdirilmiştir. Aykırı değer olabilecek gözlemler tespit edilmiştir. Martingale artıkları simetrik dağılmadığından grafiklerinin yorumlanması zordur. Ama sapma artıkları simetriktir ve referans dağılımı normal dağılım olduğundan anlamlılık düzeyine göre kesim noktası yaklaşık olarak 2 belirlenip sapma artık grafiğinde +2 ve -2 değerlerinin dışındaki gözlemlerin aykırı değerler olabilecekleri sonucuna ulaşılmıştır. Aynı referans dağılımına sahip normal sapma artıkları için de kesim noktası aynı alınmış ve benzer yorumlara ulaşılmıştır. Log-odds artıkları incelendiğinde ise aykırı değerler olabilecek gözlemler bulunamamıştır. Aykırı değer olabileceği düşünülen üç gözlem çalışmadan çıkarılıp model yeniden kurulmuş ve aykırı değerler çıkarıldıktan sonra elde edilen modelin daha iyi olduğu görülmüştür.

## References


Ata, N., Sertkaya, D., Sözer, M.T., (2007). "Orantılı Tehlike Varsayımının İncelenmesinde Kullanılan Yöntemler ve Bir Uygulama", Eskişehir Osmangazi Üniversitesi Mühendislik Mimarlık Fakültesi Dergisi, XX, S.1.

Barlow ,W. E., Prentice, R. L., (1988). Residuals for relative risk regression, Biometrika, 75, 65 – 74.

Collett, D., (1994). Modelling Survival Data in Medical Research, Chapman &Hall/CRC.

Cox, D. R., Snell, E. J., (1968). "A General Definition of Residuals", Journal of the Royal Statistical Society, 30, 2, 248-275.

Cox, D.R., Oakes, D., (1984). Analysis of Survival Data, Chapman and Hall, London.

Elandt-Johnson, R. C., Johnson, N. L., (1980). Survival Models and Data Analysis, John Wiley& Sons, Inc, New York.

Eroğlu, A., Altınok, M., Özgen, K., Sertkaya, D., (1997). "A Multivariate Analysis of Clinical and Pathological Variables in Survival After Resection of Gastric Cancer", Türkiye Klinikleri Medical Research, 15, 1, 15-20.

Fitrianto, A., Jiin, R. L. T., (2013). "Several Types of Residuals in Cox Regression Model: An Emprical Study", International Journal of Mathematical Analysis, 7,73, 2645-2654.

Flemming, T. R., Harrington, D. P., (1991). Couting Processes and Survival Analysis, Wiley, New York.

Gharibvand, L. and Fernandez, G., (2008). Advanced Statistical and Graphical Features of SAS® PHREG, SAS GLOBAL Forum 2008 Conference proceedings San Antonio TX.

Gharibvand, L., Jeske, D.R., Liao, S., (2008). Evaluation of a Hospice Care Referral Program Using Cox Proportional Hazards Model, Western Users of SAS Software Conference, Universal City, CA.

Hosmer, D. W., Lemeshow, S., (1999). Applied Survival Analysis: Regression Modeling of Time to Event Data, Wiley&Sons, New York.

Klein, J. P., Moeschberger, M. L., (2003). Survival Analysis: Techniques for Censored and Truncated Data., Springer, New York.

Kleinbaum,D.G., Klein,M., (2005). Survival analysis: A Self-Learning Text, Second Edition, Springer.

Kul, S., (2010) "The Use of Survival Analysis for Clinical Pathways", International Journal of Care Pathways,14, 23–26.

Lin, D. Y., Wei, L. J., Ying, Z., (1993). "Checking the Cox Model with Cumulative Sums of Martingale-Based Residuals", Biometrika, 80, 3, 557-572.

Nardi, A., Schemper, M., (1999). "New Residuals for Cox Regression and Their Application to Outlier Screening", Biometrics, 55, 2, 523-529.

Nardi, A., Schemper, M., (2003). "Comparing Cox and Parametric Models in Clinical Studies", Statistics in Medicine, 22, 3597-3610.

Noh, N. A., (2010). Detecting Outliers and Influential Observations in Survival Model, Master Thesis, University of Malaya, Institute of Mathematical Sciences, Kuala Lumpur.

Schoenfeld, D., (1982). "Partial Residuals for the Proportional Hazards Regression Model", Biometrika, 69, 239-241.

Stepanova, M., Thomas, L., (2002). "Survival Analysis Methods for Personal Loan Data", Operations Research, 50, 2, 277-289.

Tableman, M., Kim, J.S., (2004). Survival Analysis Using S: Analysis of Time-to-Event Data, Chapman & Hall/CRC.

Terzi, Y., Cengiz, M.A., Bek, Y., (2005). "Cox Regresyon Modelinde Oransal Hazard Varsayımının Artıklarla İncelenmesi ve Akciğer Kanseri Hastaları Üzerinde Uygulanması", Türkiye Klinikleri Tıp Bilimleri Dergisi, 25, 770-775.

Therneau, T. M., Grambcsh, P. M., Fleming, T. R., (1990). "Martingale-based Residuals for Survival Models", Biometrika, 77, 1, 147-160.

Therneau, T. M., Grambsch, P. M., (2000). Modeling Survival Data: Extending Cox Model, Springer, New York.

Winnett, A, Sasieni, P., Miscellanea: (2001). "A Note on Scaled Schoenfeld Residuals for the Proportional Hazards Model", Biometrika, 88, 565-71.

Yay, M., Çoker, E., Uysal, Ö., (2007). "Yaşam Analizinde Cox Regresyon Modeli ve Artıkların İncelenmesi", Cerrahpaşa Tıp Dergisi, 38, 139 – 145.