COX REGRESYON MODELİNDE DEĞİŞKEN SEÇİM YÖNTEMLERİ

VARIABLE SELECTION METHODS IN COX REGRESSION MODEL

PINAR AKBABA

PROF. DR. NİHAL ATA TUTKUN
Tez Danışmanı

Hacettepe Üniversitesi
Lisansüstü Eğitim-Öğretim ve Sınav Yönetmeliğinin
İstatistik Anabilim Dalı için Öngördüğü
YÜKSEK LİSANS TEZİ olarak hazırlanmıştır.

2022

Canım kızım Rüveyda Beray ve

doğacak oğlum Ömer Asaf'a…

ÖZET

COX REGRESYON MODELİNDE DEĞİŞKEN SEÇİM YÖNTEMLERİ

Pınar AKBABA

Yüksek Lisans, İstatistik Bölümü

Tez Danışmanı: Prof. Dr. Nihal ATA TUTKUN

Aralık 2022, 65 sayfa

Yaşam çözümlemesinde Cox regresyon modeli, yaşam süresi ile bir veya daha fazla açıklayıcı değişkenler arasındaki bağıntıyı modellemek için kullanılır. Modelleme yaparken Cox regresyon modelinin doğrusal bileşenine dahil edilecek ve edilmeyecek açıklayıcı değişkenleri belirlemek önemlidir. Bu değişkenlerin belirlenmesinde adımsal seçim yöntemleri, en iyi alt küme seçim yöntemleri ve küçültme yöntemleri kullanılmaktadır. Literatürde, adımsal seçim yöntemleri ve en iyi alt küme seçim yöntemleri sıklıkla kullanılırken, son yıllarda küçültme yöntemlerinin kullanımı da önem kazanmıştır.

Tez çalışmasında Cox regresyon modelinde değişken seçim yöntemleri için literatür hakkında genel bilgiler aktarılmış; adımsal seçim yöntemleri için ileriye doğru seçim, geriye doğru eleme, adımsal ileriye doğru seçim, adımsal geriye doğru eleme ve geliştirilmiş geriye doğru eleme yöntemleri; en iyi alt küme seçim yöntemleri için Akaike bilgi kriteri ve Bayesci bilgi kriteri; küçültme yöntemleri için Ridge regresyon modeli ve LASSO regresyon modeli ayrıntılı olarak incelenmiştir. Bu yöntemler literatürde yer alan böbrek kanseri veri kümesi üzerinde uygulanmış, elde edilen sonuçlar yorumlanmıştır.

Anahtar Kelimeler: Cox regresyon modeli, Değişken seçim yöntemleri, Adımsal seçim yöntemleri, Küçültme yöntemleri

ABSTRACT

VARIABLE SELECTION METHODS IN COX REGRESSION MODEL

Pınar AKBABA

Master of Science, Department of Statistics

Supervisor: Prof. Dr. Nihal ATA TUTKUN

December 2022, 65 pages

In survival analysis, the Cox regression model is used to model the relationship between survival time and one or more covariates. When modeling, it is important to determine which covariates should or should not be included in the linear component of the Cox regression model. Stepwise selection methods, best subset selection methods and shrinkage methods are used to determine these variables. While stepwise selection methods and best subset selection methods are frequently used in the literature, the usage of shrinkage methods has gained importance in recent years.

In the thesis, general information about the literature for variable selection methods in the Cox regression model is given; Forward selection, backward elimination, stepwise forward selection, stepwise backward elimination and augmented backward elimination methods for stepwise selection methods; Akaike information criterion and Bayesian information criterion for best subset methods; Ridge regression model and LASSO regression model were

examined in detail for shrinkage methods. These methods were applied on the kidney cancer dataset in the literature, and the results were interpreted.

Keywords: Cox regression models, Variable selection, Stepwise selection, Shrinkage methods

TEŞEKKÜR

Tezimin oluşmasında en büyük desteği veren, çalışmalarımın her aşamasında değerli bilgileri ve yardımlarıyla bana yol gösteren, sabırla ve büyük bir ilgiyle her zaman beni teşvik eden ve bana yol gösteren danışmanım Sayın Prof. Dr. Nihal ATA TUTKUN'a,

Ege Üniversitesi Fen Fakültesi İstatistik Bölümü'nde göreve başladığım günden itibaren yardımlarını esirgemeyen Bölüm Başkanı Sayın Prof. Dr. Ali MERT'e,

Tezimin değerlendirilme sürecinde değerli katkılarda bulunan Sayın Prof. Dr. Yasemin YAVUZ ve Sayın Prof. Dr. Duru KARASOY'a,

Hayatımın her aşamasında maddi manevi desteklerini esirgemeyen, her zaman güvenen ve yanımda olan sevgili annem Özlem SAYGI ve sevgili babam Rifat SAYGI'ya; kardeşlerim Burak SAYGI ve Mustafa Kaan SAYGI'ya,

Sadece akademik hayatımda ve yüksek lisans öğrenimim süresince değil; hayatımın her alanında bana yön gösteren, sabırla destekleyen, yardımlarını esirgemeyen ve bana benden çok inanan sevgili hayat arkadaşım Ömer AKBABA'ya,

sonsuz teşekkürlerimi sunarım.

İÇİNDEKİLER

ŞEKİLLER DİZİNİ

ÇİZELGELER DİZİNİ

KISALTMALAR

Kısaltmalar

K-M	Kaplan-Meier
TO	Tehlike Oranı
CRM	Cox Regresyon Modeli
AIC	Akaike Bilgi Kriteri
BIC	Bayesci Bilgi Kriteri
LASSO	En Küçük Mutlak Küçültme ve Seçim İşlemi

1. GİRİŞ

Yaşam çözümlemesi, bir olay (event) meydana gelene kadar geçen sürenin analizi için kullanılan istatistiksel yöntemler topluluğudur. Yaşam çözümlemesi terimi, ilk olarak ilgilenilen olayın *"ölüm"* olduğu durumlarda ortaya çıkmışsa da yaygın olarak *"hastalık, iyileşme, bozulma, boşanma vb."* gibi durumlar da olay olarak tanımlanmaktadır (Mukhopadhyay ve Singh, 2017).

Yaşam çözümlemesi, özellikle biyoistatistik, tıp ve demografi çalışmalarında kullanılmakla birlikte, bilgisayar bilimlerindeki hızlı gelişmelerin ve güçlü istatistiksel yazılım paketlerinin kullanım kolaylığının bir sonucu olarak yaşam çözümlemesi mühendislik, sosyoloji, ekonomi gibi birçok çalışma alanlarında da yaygın bir şekilde kullanılmaya başlanmıştır (Liu, 2012; Mukhopadhyay ve Singh, 2017).

Yaşam çözümlemesi, literatürde ilk olarak mortalite (ölüm) ve morbidite (hasta olma oranı) araştırmaları için kullanılmıştır. Yaşam verilerini analiz etmek için 1662 yılında İngiliz istatistikçi John Graunt tarafından ilk yaşam tablosu yöntemi yayınlanmıştır (Liu, 2012). Ancak 1958 yılında geliştirilen K-M yöntemi gibi modern yöntemler ile yaşam tablosunun yaşam çözümlemesindeki önemi azalmıştır (Eröz, 2019). Bu önerilen yeni yöntemlerin çoğu yaşam süresinin dağılımı hakkında hiçbir varsayım içermediğinden parametrik olmayan yöntemlerdir (Mukhopadhyay ve Singh, 2017). 1972 yılında ise Sir David Roxbee Cox tarafından geliştirilen Cox regresyon modeli ve kısmi olabilirlik yöntemi, yaşam verilerinin analizinde regresyon modeli ile karakterize edilen çok sayıda istatistiksel yöntem ve tekniklerin gelişmesini sağlamıştır. Cox regresyon modelinin en büyük katkısı, ölçülebilir açıklayıcı değişkenlerle ilişkili olarak karmaşık yaşam süreçlerini modellemek için esnek bir istatistiksel yaklaşımın sağlanmasıdır (Liu, 2012).

Model seçimi yaşam çözümlemesi için önemlidir. Yaşam çözümlemesinde model seçimine başlarken temelde "Hangi açıklayıcı değişkenlerin kullanılacağına nasıl karar verilir?" ve "Son modelin uygun olup olmadığına nasıl karar verilir?" şeklinde iki önemli sorunun cevabı araştırılmaktadır. Model seçimi için bilim bilgisi, deneme-yanılma, adımsal seçim yöntemleri, bilgi kriterleri, küçültme yöntemleri gibi çeşitli yaklaşımlar vardır. Tek değişkenli durumda K-M grafiklerine bakılarak ve orantılı tehlikeler varsayımı kontrol edilerek karar verilebilir. Değişken sayısı (p), modeldeki bilinmeyen parametrelerin sayısı

ondan büyük olduğunda (p>10) 2^p olası terim kombinasyonu vardır ve binden fazla açıklayıcı değişken kombinasyonu olabilir, bu yüzden olası modellere ilişkin hesaplamalar zor olmaktadır. Bu gibi durumlarda adımsal seçim (stepwise selection) yöntemleri ve küçültme (shrinkage) yöntemleri kullanılabilir (Heinze, Wallisch ve Dunkler, 2017).

Çoğu değişken seçim kriteri, cezalandırılmış en küçük kareler ve cezalandırılmış olabilirlik fonksiyonu ile yakından ilişkilidir. Akaike bilgi kriteri (AIC) (Akaike, 1974) ve Bayesci bilgi kriteri (BIC) (Schwarz, 1978) gibi bazı en iyi alt küme seçim kriterleri kolaylıkla yaşam çözümlemesi için genişletilmiştir. Hurvich, Simonoff ve Tsai (1998), parametrik olmayan regresyon modelleri için AIC'i, düzeltilmiş AIC (AIC_c) olarak genişletmişlerdir. Liang ve Zou (2008) ise AIC_c'yi yaşam çözümlemesine genişleterek AIC_{SUR}'i önermişlerdir. Volinsky ve Raftery (2000), BIC'i Cox regresyon modeli için genişletmiş ve BIC'deki birim sayısı yerine durdurulmamış olayların sayısının kullanılması için bir değişiklik önermişlerdir. Klasik değişken seçim yöntemleri, adımsal seçim yöntemleri ve en iyi alt küme seçimi gibi alt küme seçimini gerektirmektedir. Alt küme seçim yöntemleri, pratik olarak yararlı olmakla birlikte, teorik özelliklerinin anlaşılması biraz zordur. Alt küme seçiminin avantajlarını korumak ve alt küme seçiminin kararsızlığından kaçınmak için Tibshirani (1996), doğrusal regresyon modelleri ve genelleştirilmiş doğrusal modeller için LASSO değişken seçim yöntemini önermiştir. Daha sonra Tibshirani (1997), LASSO regresyon yöntemini Cox regresyon modeli için genişletmiştir. Ayrıca Hastie, Tibshirani ve Wainright (2015), LASSO regresyon yöntemi için istatistiksel öğrenme ile ilgili çalışmalarda bulunmuşlardır. Heinze, Wallisch ve Dunkler (2017) çalışmalarında değişken seçim yöntemlerini incelemişlerdir. Ekman (2017), değişken seçim yöntemleri için bir simülasyon çalışması yapmış ve LASSO yönteminin adımsal seçim yöntemlerine eşdeğer veya daha iyi sonuçlar verdiğini belirtmiştir. Koole (2017) ve Farooq ve Karami (2019) Cox Regresyon modeli için değişken seçim ve küçültme yöntemleriyle ilgili çalışmalarda bulunmuşlardır. Hastie ve ark. (2021), Wen ve ark. (2021), Thernau ve ark.(2022) ile Ripley ve ark.(2022) ise Rstudio'da çeşitli paket programlar geliştirmişlerdir.

Tezin ikinci bölümünde yaşam çözümlemesi hakkında temel kavramlar sunulmuş, Cox regresyon modeli hakkında bilgiler aktarılmıştır. Üçüncü bölümde Cox regresyon modeli için değişken seçim yöntemleri ayrıntılı olarak verilmiştir. Dördüncü bölümde literatürde yer alan

böbrek kanseri veri kümesine Cox regresyon modelleri için kullanılabilecek değişken seçim yöntemleri uygulanmış ve sonuçlar elde edilmiştir. Son bölümde ise tez çalışması özetlenmiş, adımsal seçim yöntemleri ile küçültme yöntemleri karşılaştırılmış ve ileriye dönük çalışmalar için öneriler sunulmuştur.

2. YAŞAM ÇÖZÜMLEMESİ

Yaşam çözümlemesi, bir başlangıç zamanından tanımlanmış bir olay gerçekleşene kadar geçen süreyi inceleyen istatistiksel bir yöntemdir (Ekman, 2017). Olay, yaşam çözümlemesi yöntemlerinin uygulandığı çalışma alanına göre ölüm, hastalık, iyileşme, nüks, boşanma, işten ayrılma, iflas, makine bozulması, vb. gibi durumlar olabilir. Genellikle bir olaya kadar geçen süre yıl, ay, hafta, gün, saat olarak ifade edilebilirken bir birimin yaşı da süre olarak ifade edilebilir (Petersson ve Sehlstedt, 2018). İlgilenilen olayın gerçekleşmesi başarısızlık, ilgilenilen olay gerçekleşene kadar geçen süre ise yaşam süresi veya başarısızlık süresi olarak tanımlanabilir.

Yaşam süresinin analizinde gözlemlenen temel sorun, bazı birimler için ilgilenilen olayın yaşam süresinin tam olarak bilinmemesidir (Ekman, 2017). Bu durum, durdurma (censoring) olarak adlandırılmaktadır (Kleinbaum ve Klein, 2012). Yaşam çözümlemesi yöntemlerinin en önemli özelliği ise, durdurulmuş birimleri eksik veri olarak ele almadan analiz edebilen tek yöntem olması ve yaşam verileri için normal dağılıma ilişkin hiçbir varsayımda bulunmamalarıdır (Hazra ve Gogtay, 2017).

2.1. Temel Kavramlar

Yaşam çözümlemesinde, yaşam verilerinin belirli özelliklerini tanımlamak için kullanılan bazı terimler aşağıda verilmiştir.

Durdurma

Yaşam çözümlemesi, durdurma adı verilen önemli bir kavramı dikkate alarak incelemeler yapmaktadır. Çalışmada yer alan bazı birimler için ilgilenilen olayın yaşam süresi tam olarak bilinmemektedir (Mukhopadhyay ve Singh, 2017). Çalışma sonlanmadan önce olayın meydana gelmemesi (administrative censoring), takibin yapılamaması (lost to follow up) veya ilgilenilen olay dışında bir nedenle çalışmadan çekilmesi (withdrawing) gibi çeşitli nedenlerle meydana gelen farklı durdurma türleri vardır (Kleinbaum ve Klein, 2012). Bunlar Çizelge 2.1'de verilmiştir (Ekman, 2017).

Çizelge 2.1. Durdurma türleri

Durdurma Türü	**Tanım**
Soldan durdurma (Left censoring)	Belirli bir süreden önce ilgilenilen olayın olduğu bilinir ancak tam olarak ne zaman olduğu bilinmez.
Sağdan durdurma (Right censoring)	Belirli bir süreden sonra ilgilenilen olayın olduğu bilinir ancak tam olarak ne zaman olduğu bilinmez.
Aralıklı durdurma (Interval censoring)	İlgilenilen olayın ne zaman olduğu tam bilinmiyor, ancak iki değer arasında bir zaman içinde olduğu bilinmektedir.
Tip I durdurma (Type I censoring)	Çalışma, belirli bir zaman noktası belirlenerek sona erdirilmektedir. Bu, zamansal durdurma (time censoring) olarak da adlandırılır.
Tip II durdurma (Type II censoring)	Çalışma, sabit sayıda ilgilenilen olay meydana geldiğinde sona ermektedir. Bu, sayısal durdurma (failure censoring) olarak da adlandırılır.
Rastgele durdurma (Random censoring)	Durdurma zamanları rastgele belirlendiği durumlarda söz konusudur.

Durdurma bilgi içeren (informative) veya bilgi içermeyen (non-informative) olarak da ortaya çıkmaktadır. Durdurma, bir olayın meydana gelme olasılığı ile ilgili ise bilgi içeren durdurma; bir olayın meydana gelme olasılığı ile ilgili değilse bilgi içermeyen durdurma olarak adlandırılmaktadır. Örneğin, tedaviye devam eden bir hastanın sağlıkta bir bozulma veya iyileşme olması durumunda tedaviyi bırakması bilgi içeren durdurmaya; bir klinik araştırma söz konusu olduğunda bu çalışmadan ayrılan bir hastanın çalışmasıyla ilgisi olmayan nedenlerle bunu yapması bilgi içermeyen durdurmaya örnek olarak verilebilir (Baker, Wax ve Patterson, 1993; Ekman, 2017). Yaşam verilerinin analizi için kullanılan standart yöntemlerde, durdurmanın "bilgi içermeyen" olduğu varsayılır (Clark ve ark., 2003).

Yaşam fonksiyonu

Yaşam fonksiyonu (survival function), S(t), yaşam çözümlemesi için temel bir fonksiyondur ve birimin belirli bir t zamanından daha uzun süre yaşama olasılığını verir. T, ilgilenilen olay meydana gelene kadar geçen süreyi temsil eden, negatif olmayan bir rastlantı değişkenidir. Yaşam fonksiyonu, T'nin belirtilen t zamanını aşma olasılığını verir ve

$$S(t) = P(T > t) = 1 - F(t) = \int_t^{\infty} f(x)\, dx \tag{2.1}$$

biçiminde tanımlanır (Mukhopadhyay ve Singh, 2017). Yaşam fonksiyonu, tüm $t > x$ için $S(t) \leq S(x)$ olduğunda monoton olarak azalan bir fonksiyondur. Genellikle $t = 0$ iken $S(0) = 1$ ve $t = \infty$ iken $\lim_{t\to\infty} S(t) = 0$ olduğu varsayılır. (Ekman, 2017).

Çarpım-limit (product-limit) yöntemi olarak bilinen K-M tahmini ise, yaşam fonksiyonunu tahmin eden basit ve etkili bir yöntemdir.

Kaplan-Meier tahmini

Bir veri kümesinde durdurulmuş birimler bulunmadığında yaşam fonksiyonu tahmini için, deneysel yaşam fonksiyonu,

$$\hat{S}(t) = \frac{1}{n}\sum_{i=1}^{n} \mathrm{I}\{t_i < t\} \tag{2.2}$$

biçiminde tanımlanmaktır. Burada I gösterge fonksiyonudur ve t_i, i. birim için yaşam süresidir (i=1,2,…,n). Bu tahmin, Kaplan ve Meier (1958) tarafından durdurulmuş birimler için genişletilmiştir. $t_1 \leq t_2 \leq \cdots \leq t_m$, m tane farklı sıralı yaşam süreleri, d_i, t_i zamanındaki başarısızlıkların sayısı ve R_i, t_i zamanından hemen önce risk altındaki birimlerin sayısı olmak üzere K-M tahmini,

$$\hat{S}(t) = \prod_{i=t_i<t} \left(1 - \frac{d_i}{R_i}\right) \tag{2.3}$$

biçiminde tanımlanır. Böylece Eşitlik 2.3'de verilen K-M tahmini, durdurmanın olduğu ve durdurmanın olmadığı her iki durum için bir adım fonksiyonudur (Karasoy ve Ata Tutkun,

2016). Eğer birimler durdurulmuş birimleri içermezse K-M tahmini deneysel yaşam fonksiyonuna eşittir (Ekman, 2017).

Tehlike fonksiyonu

Tehlike fonksiyonu (hazard function), yaşam fonksiyonunun aksine, başarısızlığa odaklanır. Bu fonksiyon, başarısızlık hızı (failure rate), ani ölüm hızı (instantaneous death rate), koşullu başarısızlık hızı (conditional failure rate) ve ölümlülük gücü (force of mortality) olarak da adlandırılır (Karasoy ve Ata Tutkun, 2016).

Tehlike fonksiyonu, $h(t)$, birimin t zamanına kadar başarısızlık ile karşılaşmadığı bilindiğine göre ilgilenilen olayın meydana gelmesi için birim zaman başına anlık riski vermektedir.

Birimin t zamanına kadar başarısızlık ile karşılaşmadığı bilindiği göz önüne alındığında, (t, t+Δt) arasındaki küçük bir zaman diliminde başarısızlık riski olarak da adlandırılmaktadır (Mukhopadhyay ve Singh, 2017). Bu, olayın henüz gerçekleşmediğini ancak bir sonraki anda olayı yaşama riski olarak da düşünülebilir (Ekman, 2017). Bu anlamda tehlike, bir risk ölçüsüdür: t_1 ve t_2 zamanları arasındaki tehlike ne kadar büyükse, bu zaman aralığında başarısızlık riski de o kadar büyük olur (Mukhopadhyay ve Singh, 2017). Tehlike fonksiyonu, aşağıdaki gibi tanımlanır:

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t}.$$

(2.4)

Tehlike fonksiyonu ile yaşam fonksiyonu arasında yakın bir ilişki vardır ve

$$h(t) = \frac{f(t)}{S(t)} = -\frac{d}{dt}\log(S(t)), \quad (2.5)$$

ya da

$$S(t) = \exp\left[-\int_0^t h(x)dx\right] = \exp(-H(t)) \quad (2.6)$$

biçiminde verilir. Burada $H(t)$, birikimli tehlike fonksiyonu (cumulative hazard rate) olarak adlandırılmaktadır. Birikimli tehlike fonksiyonu, yaşam fonksiyonundan

$$H(t) = -\log S(t) \tag{2.7}$$

ile elde edilebilir.

Tehlike oranı

Tehlike oranı (hazard ratio, TO), lojistik regresyon çözümlemesinde kullanılan göreli riske (relative risk) benzemektedir.

Herhangi iki tahmin edici x ve x^* için tehlike oranı,

$$TO = \frac{h(t|x^*)}{h(t|x)} = \frac{h_0(t)\exp(\beta x^*)}{h_0(t)\exp(\beta x)} = \exp(\beta(x^* - x)) \tag{2.8}$$

biçiminde elde edilir. TO zaman içinde sabittir. Bilgi içermeyen durdurma ile ilgili varsayımı ile birlikte, orantılı tehlikeler varsayımı, yani TO'nun zaman içinde sabit olması, Cox regresyon modelinin temel varsayımlarıdır (Ekman, 2017).

2.2. Yaşam Fonksiyonlarının Karşılaştırılması

K-M grafikleri kullanılarak ilgilenilen değişkenin düzeyleri arasında yaşam olasılıkları arasında fark olup olmadığı gözlemlenebilir. Ancak düzey sayısı arttıkça, K-M grafiklerinin yorumu güçleşmektedir. Bu durumda düzeyler arasındaki farklılıkları incelemek için istatistiksel testler kullanılmaktadır. Bu amaçla yaygın olarak kullanılan testlerden biri log-rank testidir (Mukhopadhyay ve Singh, 2017). Bu test istatistiği için hipotezler;

$H_o: S_1(t) = S_2(t) = \cdots = S_k(t) \quad t \leq \tau$

$H_1: S_i(t) \neq S_j(t)$ i, j'den farklıdır.

biçimindedir. Burada τ, tüm grupların en az bir birimin risk altında olduğu en uzun zamandır.

Test istatistiği;

$$z_j = \sum_{i=t_{(i)}<\tau} \left(d_{ij} - R_{ij} \frac{d_i}{R_i} \right) \quad (2.9)$$

biçimindedir. Asıl amaç, sıfır hipotezi doğru olduğunda, yani yaşam olasılıkları arasında fark olmadığında, her gruptaki birimlerin gözlenen olay sayısı ile beklenen olay sayısını karşılaştırmaktır. Burada j. grup için olay (j = 1,..., k) z_j ve R_{ij}, j. grupta risk altındaki birimlerin sayısıdır (Ekman, 2017). Sıfır hipotezi altında, log-rank istatistiği iki grup olduğunda yaklaşık olarak bir serbestlik dereceli ki-karedir (Mukhopadhyay ve Singh, 2017).

2.3. Cox Regresyon Modeli

K-M tahmini ve log-rank testi ile belirli gruplar için yaşam fonksiyonu farklılıkları incelenebilir fakat yaşam fonksiyonunun bazı sürekli değişkenlerden nasıl etkilendiğini görmek ya da yaşam fonksiyonu iki veya daha fazla tahmin edicinin bir fonksiyonu olarak incelenmek istendiğinde uygun bir regresyon yöntemi kullanılmalıdır (Ekman, 2017).

Bu regresyon yöntemi, sonuç (bağımlı) değişkenini, yani olayın meydana gelmesine kadar geçen süreyi, ve açıklayıcı değişkenleri içermektedir. Yaşam verilerini analiz ederken bağımlı değişkenin biçimi ve verilerin durdurulmuş birimleri içerebilmesinden dolayı doğrusal regresyon kullanmak uygun değildir (Liu, 2012). Bunun yerine, yaşam verileri yapısına uyacak biçimde geliştirilmiş özel regresyon modelleri vardır, en yaygın olanlardan biri Cox regresyon modelidir (Cox, 1972).

Cox regresyon modeli, 1972 yılında Cox tarafından geliştirilmiştir. Açıklayıcı değişkenlere bağlı olarak tehlike oranı içeren bu modele Cox regresyon modeli (CRM) veya Cox orantılı tehlikeler modeli adı verilmektedir (Koole, 2017). Bu model, temel tehlike oranını etkisiz (unspecified) bırakırken orantılı tehlikeler varsayımını kullanarak açıklayıcı değişken etkilerinin sağlam (robust), tutarlı (consistent) ve etkili (efficient) tahminlerini elde etmektedir (Liu, 2012). Günümüzde sağlam ve tutarlı tahmin edicilerinin olması, tehlike oranının hesaplanabilmesi ve farklı bilim alanlarındaki çalışmalara uygulanabilirliği nedeniyle yaşam çözümlemesinde en çok kullanılan modellerden biridir. CRM'ye göre i. birimin tehlike oranı fonksiyonu,

$$h(t|x) = h_0(t)e^{\beta^T x} \quad (2.10)$$

biçiminde tanımlanır. $\boldsymbol{x} = (\boldsymbol{x_1}, \boldsymbol{x_2}, \ldots., \boldsymbol{x_p})^{\boldsymbol{T}}$ açıklayıcı değişkenler vektörüdür ve $\boldsymbol{\beta^T} = (\boldsymbol{\beta_1}, \ldots, \boldsymbol{\beta_p})$ modelin regresyon parametreleridir. Burada $h_0(t)$ temel tehlike (baseline hazard) fonksiyonu olarak adlandırılır ve tüm açıklayıcı değişkenlerin sıfır olduğu durumdaki tehlike oranıdır. Temel tehlike negatif olmadığı sürece, ifadenin üstel kısmının tahmini tehlike fonksiyonunun her zaman negatif olmamasını sağlamaktadır (Ekman, 2017).

K-M tahmini parametrik olmayan bir yöntem olduğundan verilerin temel dağılımı hakkında herhangi bir varsayımda bulunmaz. CRM ise temel tehlike fonksiyonunun belirtilmesi gerekmediğinden yarı parametrik bir modeldir (Ekman, 2017).

2.3.1. Olabilirlik Fonksiyonu Elde Edilmesi

2.3.1.1. Kısmi Olabilirlik Yöntemi

Kısmi olabilirlik yöntemi, başarısızlık süreleri arasındaki aralıkların açıklayıcı değişkenler ve tehlike oranı arasındaki ilişki hakkında herhangi bir bilgi sağlamadığı varsayımına dayanmaktadır (Collett, 1994). CRM, kısmi olabilirlik fonksiyonu için en çok olabilirlik algoritması kullandığından tahmin yaklaşımı, kısmi olabilirlik fonksiyonu olarak adlandırılmaktadır (Liu, 2012). Kısmi olabilirlik yönteminde, olabilirlik formülü yalnızca başarısız olan birimler için olasılıkları dikkate alırken, durdurulmuş birimler için olasılıklar dikkate alınmamaktadır. Bu nedenle CRM için bu olabilirlik, tüm birimler için olasılıkları dikkate almaz ve kısmi olabilirlik olarak adlandırılır (Kleinbaum ve Klein, 2012). Ayrıca kısmi olabilirlik fonksiyonuna bilgi sağlayan, başarısızlıklar arasındaki aralıklar değil, sıralı başarısızlık süreleridir (Box-Steffensmeier ve Jones, 2004).

k farklı başarısızlık süresine sahip n boyutundaki bir veri kümesinin kısmi olabilirlik fonksiyonunu elde etmek için, veriler ilk olarak $t_1 < t_2 < \cdots < t_k$ şeklinde sıralı başarısızlık süresine göre sıralanır, burada t_i, i. birim için başarısızlık süresini göstermektedir. Her bir durdurulmamış birim için ilgilenilen olay, farklı zaman dilimlerinde gerçekleşmektedir ve aynı anda gerçekleşen başka olay olmadığı varsayılmaktadır. Durdurma δ_i ile

gösterilmektedir. δ_i gösterge değişkeni, 0 değerini alması durumunda durdurulmuş birimleri; 1 değerini alması durumunda ise durdurulmamış birimleri ifade etmektedir. Sıralı başarısızlık süreleri, x açıklayıcı değişkenlerinin bir fonksiyonu olarak modellenmektedir (Box-Steffensmeier ve Jones, 2004).

t_i zamanında ilgilenilen olayı yaşama riski olan birimlerin sayısı $R(t_i)$, risk kümesini ifade etmek üzere j. birimin T_i zamanında başarısızlığın gerçekleşmesi olasılığı;

$$\Pr\left(t_j = T_i \middle| R(t_i) \right) = \frac{e^{\beta x_i}}{\sum_{j \epsilon R(t_i)} e^{\beta' x_j}} \tag{2.11}$$

biçiminde ifade edilmektedir. Burada payda, risk kümesindeki tüm birimlerin toplamını ifade etmektedir. Bu eşitlikteki koşullu olasılıkların çarpımı ile kısmi olabilirlik fonksiyonu,

$$L = \prod_{i=1}^{k} \left(\frac{e^{\beta x_i}}{\sum_{j \epsilon R(t_i)} e^{\beta' x_j}} \right) \tag{2.12}$$

biçiminde elde edilir. Buna karşılık gelen log- kısmi olabilirlik fonksiyonu ise,

$$logL = \sum_{i=1}^{k} \left[\beta' x_i - log \sum_{j \epsilon R(t_i)} e^{\beta' x_j} \right] \tag{2.13}$$

olur. Log-kısmi olabilirlik fonksiyonu maksimize edilerek, açıklayıcı değişkenlerin regresyon katsayılarını içeren bilinmeyen parametre vektörü β'ların tahmini elde edilebilir (Liu, 2012).

Kısmi olabilirlik yöntemi, sıralı başarısızlık sürelerine bağlı olduğundan tehlike oranları, başarısızlık süreleri ve olaylar arasındaki aralıklarla ilişkili değildir. Kısmi olabilirlik yöntemi, birimlerdeki toplam birim sayısından ziyade olay sayısına dayanmaktadır. Model

parametrelerinin tahmin edilmesinde Cox regresyon modelinin uygulanması için çok daha büyük bir örneklem gerektirmektedir (Liu, 2012).

2.3.1.2. Breslow Yöntemi

Breslow yöntemi (1974), birimlerdeki başarısızlık sürelerinin aynı olması (ilgilenilen olayın aynı anda gerçekleşmesi) durumunda bu sorunu ortadan kaldırmak için kullanılan bir yöntemdir. Kısmi olabilirlik fonksiyonunun Breslow yaklaşımı, hesaplama açısından diğer yöntemlere göre daha basit olduğu için yaygın olarak kullanılan bir yöntemdir. Aynı başarısızlık sürelerine sahip birimlerdeki, olayların meydana geliş sırasını belirlemek mümkün olmadığından Breslow yönteminde hangi olayın ilk gerçekleştiğine bakılmaksızın risk kümesinin boyutunun aynı olduğu varsayılmaktadır (Box-Steffensmeier ve Jones, 2004).

Breslow yöntemi, başarısızlık süresinde risk altındaki tüm durumlardan (aynı başarısızlık süresine sahip birimler de dahil) oluşan bir risk kümesinden, sıraları bilinmese de, sıralı olarak gerçekleştiğini varsayarak kısmi olabilirlik fonksiyonuna yakınsar (Box-Steffensmeier ve Jones, 2004). Breslow olabilirlik fonksiyonu;

$$L_{Breslow} = \prod_{i=1}^{k} \frac{e^{\beta' s_i}}{\left[\sum_{j \epsilon R(t_i)} e^{\beta' x_i}\right]^{d_i}} \tag{2.14}$$

biçiminde elde edilir. Eşitlik 2.14.'de s_i, aynı anda gerçekleşen olaylar için x açıklayıcı değişkenlerin toplamı ve d_i, t_i zamanında gerçekleşen birimlerin sayısıdır. Breslow yöntemi, aynı anda gerçekleşen olayların sayısı az olduğunda uygun olduğu ifade edilmiştir; ancak aynı anda gerçekleşen olayların sayısı arttıkça her periyotta risk kümesinin büyüklüğü artar ve kısmi olabilirlik fonksiyonuna yakınsaklığı azalır (Breslow, 1974; Box-Steffensmeier ve Jones, 2004).

2.3.1.3. Efron Yöntemi

Efron (1977), Breslow yöntemini geliştirerek aynı anda gerçekleşen olayların sırasına bağlı olarak risk kümesinin nasıl değiştiğini hesaba katan bir yöntem önermiştir;

$$L_{Efron} = \prod_{i=1}^{k} \left(\frac{e^{\beta' s_i}}{\left[\sum_{j \epsilon R(t_i)} e^{\beta' x_j} - (r-1) d_i{}^{-1} \sum_{j \epsilon D(t_i)} e^{\beta' x_j} \right]} \right) \quad (2.15)$$

Burada r, aynı anda meydana gelen olayların sayısını ve $D(t_i)$, t_i zamanında risk kümesindeki süreli olayların sayısını göstermektedir (Box-Steffensmeier ve Jones, 2004). Efron Eşitlik 2.15'de, kısmi olabilirlik fonksiyonunda bir sıralama getirerek paydanın ağırlığını azaltmaktadır. Bazı araştırmacılar, örneklem büyüklüğünün küçük olduğu veya durdurulmuş birimlerin fazla olması nedeniyle Efron yönteminin Breslow yönteminden daha uygun bir yaklaşım olduğunu belirtmektedir (Liu, 2012, Karasoy ve Kaplan, 2017).

2.3.2. Orantılı Tehlikeler Varsayımı

Orantılı tehlikeler varsayımı, kısmi olabilirlik fonksiyonunda temel tehlike fonksiyonunu etkisiz bırakan bir özellik olduğundan, CRM'nin geçerliliği buna dayanır (Ekman, 2017). Orantılı tehlikeler varsayımının güvenilirliğini incelemek için birçok farklı grafiksel yöntemler ve test istatistikleri kullanılmaktadır. Bu yöntemlerden en sık kullanılanları log-log grafiği ve Schoenfeld artıklarına dayanan test istatistiği aşağıda açıklanmıştır (Ekman, 2017).

Log-log grafiği yöntemi

Log-log grafiği, diğer tüm ilgili tahmin edicileri belirledikten sonra, belirli bir kategorik tahmin edici (x_i) için orantılı tehlikeler varsayımını grafiksel olarak kontrol etmektedir. Orantılı tehlikeler varsayımı geçerli olduğunda yaşam fonksiyonu, birikimli tehlike fonksiyonu ve CRM'den

$$S(t|x) = \exp\big(-H(t|x)\big) = \exp(-H_0(t)\exp(\beta^T x)) = S_0(t)^{\exp(\beta^T x)} \quad (2.16)$$

elde edilir. Burada $S_0(t)$, birim için temel yaşam fonksiyonudur. Bunun logaritması iki kere alınarak,

$$\log\big(-\log\big(S(t|x)\big)\big) = \beta^T x + \log\big(-\log\big(S_0(t)\big)\big) \quad (2.17)$$

elde edilir. Logaritması alınırken ilk logaritma negatiftir, dolayısıyla yaşam fonksiyonları [0,1] arasında değiştiğinden eksi işareti eklenir.

x_1 ve x_2 iki farklı birime veya gruba karşılık gelen x vektörünün iki farklı özelliği olarak düşünüldüğünde, karşılık gelen log-log eğrileri,

$$\begin{cases}\beta^T x_1 + \log\big(-\log\big(S_0(t)\big)\big) \\ \beta^T x_2 + \log\big(-\log\big(S_0(t)\big)\big)\end{cases}$$

olur. Dolayısıyla, iki birim kümesi için

$$log\big(-\log\big(S(t|x_2)\big)\big) = -log\big(-\log\big(S(t|x_1)\big)\big) = \beta^T(x_2 - x_1) \qquad (2.18)$$

elde edilir. Log-log grafiklerinin kullanımıyla ilgili bir sınırlama, açıklayıcı değişkenin kategorik olması veya uygun gruplama yoluyla kategorik hale getirilmesi gerekmesidir. Bu grafiksel yöntem yerine Schoenfeld artıkları yöntemini kullanmak bir alternatif olabilir (Ekman, 2017).

Schoenfeld artıkları yöntemi

Açıklayıcı değişken sürekli ise, çok sayıda açıklayıcı değişken var ise veya birçok kategorik değişken var ise, log-log grafiğinin uygulanması zordur. Bu gibi durumlarda, Schoenfeld (1982) artıkları yönteminin orantılı tehlikeler varsayımının değerlendirilmesi için kullanılması daha uygundur (Machin, Cheung ve Parmar, 2006). Her bir birim için tek bir artık değeri yerine, her başarısızlık zamanında her bağımsız değişken için ayrı bir artık değeri vardır. k. açıklayıcı değişken ve i. başarısızlık zamanı için, Schoenfeld değeri;

$$r_{(i)k}{}^S = x_{(i)k} - \sum_{j\in R_i}(x_{jk} \times p(\hat{\beta}_k, x_{jk})) \qquad (2.19)$$

olarak tanımlanır; burada

$$p(\hat{\beta}_k, x_{jk}) = \frac{\exp(\hat{\beta}_k x_k)}{\sum_{j \in R_i} \exp(\hat{\beta}_k, x_{jk})} \quad (2.20)$$

biçimindedir ve t_i zamanında risk kümesinden $i.$ birimi seçme olasılığıdır. Dolayısıyla, Schoenfeld artığı, $t_{(i)}$ zamanında fiilen ölen birim için açıklayıcı değişken değeri x_{ik} arasındaki farktır, aynı zamanda belirlenen risk için bağımsız değişkenin beklenen değeridir. Burada beklenen değer, risk grubu için açıklayıcı değişkenler olarak, her bir birimin $t_{(i)}$ 'de fiilen ölme olasılığı ile ağırlıklandırılarak hesaplanır (Ekman, 2017).

Belirli bir değişken için orantılı tehlikeler varsayımını incelemek için geliştirilen Schoenfeld artıkları ile birimlerin yaşam süreleri rankı arasındaki korelasyon kullanılarak orantılı tehlikeler varsayımı incelenmektedir. Bu teste göre, orantılı tehlikeler varsayımının sağlanması için korelasyonunun sıfıra yakın olması beklenmektedir. Bu test istatistiği, orantılı tehlikeler varsayımının incelenmesi için kullanılan grafiksel yöntemlere göre daha nesnel bir kriter sağlamaktadır. Grafiksel yöntemler ise daha özneldir (Ata, Karasoy ve Sözer, 2007).

3. COX REGRESYON MODELİNDE DEĞİŞKEN SEÇİM YÖNTEMLERİ

Model seçimi, Cox regresyon çözümlemesinde önemli bir yere sahiptir. Uygulama yapılırken, modelde olası risk faktörleri olarak birden çok açıklayıcı değişken yer almaktadır. Model seçim sürecindeki ilk adım CRM'nin doğrusal bileşenine dahil edilecek olan açıklayıcı değişkenleri belirlemektir. En iyi olarak nitelendirilen tek bir model yerine, birçok iyi modelin kullanılması tercih edilir. Amaç, mümkün olan birçok modelin düşünülmesidir (Fan, Li ve Li, 2005; Karasoy ve Tutkun, 2016). Model seçimi, doğru modellerin oluşturulmasına veya tahmin ve sınıflama yaparken modele dahil edilecek değişkenlerin belirlenmesine yardımcı olmaktadır (Claeskens ve Hjort, 2008).

Cox regresyon modeli, belirli bir zamandaki riski p tane açıklayıcı değişkenlerin bir fonksiyonu olarak ifade etmektedir. Bu açıklayıcı değişkenlere karşılık gelen katsayılar ise olabilirlik fonksiyonları yardımı ile türetilmektedir. Bu adımlardan sonra modele dahil edilecek açıklayıcı değişkenlerin belirlenmesi gerekmektedir (Ekman, 2017).

Modelin yorumlanabilirliği ve tahminin doğruluğu için, açıklayıcı değişken sayısı p olmak üzere, p'nin büyük bir sayı olduğu varsayıldığında, $q < p$ açıklayıcı değişkenleri olan bir modelin seçilmesi uygundur. Bir modelde birçok açıklayıcı değişken olduğu durumda, bazı değişkenlerin sonuç değişkeniyle ilişkili olmadığı veya sonuç değişkeni üzerinde yalnızca küçük bir etkiye sahip olduğu görülmektedir. Bu açıklayıcı değişkenleri ve karşılık gelen etkileşim terimlerini de modelden çıkararak, daha kolay yorumlanan ve daha az karmaşık modeller elde edilebilir (Ekman, 2017).

Yaşam modellerinde genellikle birçok açıklayıcı değişken bulunmaktadır. Bu açıklayıcı değişkenlerin sayısı çok fazla ise aşırı parametreli ve yorumlanması zor bir model elde edilebilmektedir. Bu gibi durumlarda, modeli daha anlamlı ve uygun hale getirmek için açıklayıcı değişkenlerin sayısını sınırlandırmada kullanılan birkaç regresyon yöntemi vardır (Koole, 2017).

Bu yöntemlerden biri, en anlamlı değişkenlerin modele dahil edildiği veya modele katkısı olmayan değişkenlerin modelden çıkarıldığı kesikli bir süreç (a discrete process) olan değişken seçimidir. Çok sayıda açıklayıcı değişken olduğunda, en güçlü etkilere sahip olan değişkenlerin daha küçük bir alt kümesi belirlenmektedir (Koole, 2017).

Diğer bir yöntem ise, sürekli bir yöntem türü olan ve dolayısıyla değişkenlikten çok fazla etkilenmeyen küçültme yöntemidir. Bu yöntem sıfıra doğru küçülen katsayılar ve aynı zamanda tam olarak sıfır olan katsayılar üretmektedir (Koole, 2017). Cox regresyon modelinde değişken seçimi ve küçültme için LASSO regresyon yönteminin kullanılması ilk kez Tibshirani (1997) tarafından önerilmiştir.

3.1. Adımsal Seçim Yöntemleri

Açıklayıcı değişken sayısının fazla olması durumunda olası modellerin elde edilmesi için gerekli hesaplanmalar zor olduğu için modellerin belirlenmesinde adımsal seçim yöntemleri kullanılabilir. Ancak bu yöntemlerin bazı zayıf yönleri vardır. Birçok iyi model yerine olası modellerden tek bir modelin iyi olarak belirlenmesi, hiyerarşik ilkeyi dikkate almama eğiliminde olması, modeldeki değişkenleri belirlemede durdurma kuralının kullanılması gibi nedenlerle bu yöntemlerin model seçiminde kullanımları sınırlıdır (Karasoy ve Tutkun, 2016). Adımsal seçim yöntemleri, Alt bölüm 3.1.1-3.1.5'de ele alınmıştır.

3.1.1. İleriye Doğru Seçim Yöntemi

İleriye doğru seçim yöntemi, en iyi alt küme seçimine alternatif bir yöntemdir. En iyi alt küme seçim yöntemi p tahmin edicilerinin alt kümelerini içeren tüm 2^p olası modelleri dikkate alırken, ileriye doğru seçim yöntemi çok daha küçük bir model kümesini dikkate alır. İleriye doğru seçim yöntemi, hiçbir açıklayıcı değişken içermeyen bir modelle başlar ve ardından modele katkı sağlayan tüm açıklayıcı değişkenler modele girene kadar her seferinde bir açıklayıcı değişken modele eklenir. Yöntemin adımları aşağıdaki gibi özetlenebilir:

Adım 1: M_0 açıklayıcı değişken içermeyen temel modeli gösterir. Bu model ile sürece başlanır.

Adım 2: $k = 0,1, \dots, p-1$ olmak üzere M_k'ye bir açıklayıcı değişken ekleyen $p-k$ modelleri belirlenir.

Adım 3: $M_0, M_1, \dots, M_p$ modelleri arasından Akaike Bilgi Kriteri (AIC), Bayesci Bilgi Kriteri (BIC), çapraz doğrulama gibi yöntemlerinden biri kullanılır ve M_{k+1} olarak adlandırılan en iyi model seçilir (James ve ark., 2017).

3.1.2. Geriye Doğru Eleme Yöntemi

Alternatif bir diğer yöntem ise geriye doğru eleme yöntemidir. Bununla birlikte, ileriye doğru seçim yönteminden farklı olarak, tüm açıklayıcı değişkenleri içeren tam model ile başlanır ve ardından modeli en az geliştiren açıklayıcı değişkenleri teker teker modelden çıkarır. Yöntemin adımları aşağıdaki gibi özetlenebilir:

Adım 1: M_p, tüm açıklayıcı değişkenleri içeren tam modeli gösterir. Bu model ile sürece başlanır.

Adım 2: $k = p, p-1, \dots, 1$ olmak üzere $k-1$ tane açıklayıcı değişken için M_k'deki açıklayıcı değişkenlerden biri hariç tümünü içeren tüm k modelleri belirlenir.

Adım 3: Bu k modelleri arasından AIC, BIC, çapraz doğrulama gibi yöntemlerinden biri kullanılır ve M_{k-1} olarak adlandırılan en iyi model seçilir (James ve ark., 2017).

3.1.3. Adımsal İleriye Doğru Seçim Yöntemi

Adımsal ileriye doğru seçim yönteminde açıklayıcı değişken içermeyen boş bir model ile sürece başlanır. Bir ileriye doğru seçim adımı gerçekleştirilir. Her değişken eklendikten sonra bir geriye doğru eleme adımı gerçekleştirilir. Sonraki ileriye doğru seçim adımlarında, önceki adımlarda çıkarılan değişkenler yeniden değerlendirilir. Bu yöntem, hiçbir değişken modelden çıkarılmadığında ya da modele eklenmediğinde durdurulur (Heinze, Wallisch ve Dunkler, 2017).

3.1.4. Adımsal Geriye Doğru Eleme Yöntemi

Adımsal geriye doğru eleme yöntemine açıklayıcı değişkenlerin hepsinin yer aldığı tam model ile başlanır. Bir geriye doğru eleme adımı gerçekleştirilir ve her değişken çıkarıldıktan

sonra bir ileriye doğru seçim adımı gerçekleştirilerek aşamalı yaklaşım yapılır. Bu yöntem, hiçbir değişken modelden çıkarılmadığında ya da modele eklenmediğinde durdurulur (Heinze, Wallisch ve Dunkler, 2017)

3.1.5. Geliştirilmiş Geriye Doğru Eleme Yöntemi

Geriye doğru eleme yönteminin standartlaştırılmış halidir ve geriye doğru eleme yöntemini bir tahmin-içi değişim kriteri (change-in-estimate criterion) ile birleştirir. Özellikle epidemiyolojik araştırmalarda, açıklayıcı modellerde uyumlu değişkenleri seçmek için tahmin-içi değişim kriteri (change-in-estimate criterion) sıklıkla uygulanır. Tahmin-içi değişim kriteri genellikle göreli değişim (relative change) olarak ifade edilir. Bu yaklaşımı kullanarak, bir modelden anlamlı değişkenin çıkarılmasının anlamlı bir değişim tahminine yol açacağı ve anlamsız değişkenin çıkarılmasının anlamsız bir değişim tahminine yol açacağı gösterilebilir. Ayrıca Heinze, Wallisch ve Dunkler (2017), doğrusal, lojistik ve Cox regresyon modellerinde kullanılması için tahmin-içi değişim kriteri standartlaştırmışlardır, böylece tahmin-içi değişim tek bir ortak eşik değer τ ile karşılaştırılarak açıklayıcı değişkenlerin modelden çıkarılıp çıkarılmayacağına karar verilir. Açıklayıcı değişkenlerin $p > \alpha$ olsa bile hariç tutulmaları diğer herhangi bir değişkende standart bir tahmin değişiminin τ'den büyük olmasına neden oluyorsa, değişkenler dahil edilmelidir (Heinze, Wallisch ve Dunkler, 2017).

Modeldeki herhangi bir değişken, bir modelden çıkarılıp bırakılmamalıdır. Dunkler ve ark. (2014), standartlaştırılmış tahmin-içi değişim kriterinin geriye doğru seçim yöntemi ile birleştirilmesini ve geliştirilmiş geriye doğru seçim yönteminin uygulanmasını önermiştir. Geliştirilmiş geriye doğru eleme yöntemi, geriye doğru eleme yönteminden daha büyük modelleri seçme eğilimindedir ve daha az yanlı regresyon katsayılarına yol açtığı gözlenmiştir (Dunkler ve ark., 2014).

3.2. En İyi Alt Küme Seçim Yöntemleri

Modele hangi değişkenlerin dahil edileceğini bulmanın diğer bir yolu, en iyi alt küme seçimidir. p açıklayıcı değişkenlerinin her olası kombinasyonu (2^p tane terim) için ayrı bir

model elde edilerek en iyi performansı gösteren model seçilir. Burada genellikle AIC, BIC ve çapraz doğrulama gibi tekniklerinden biri kullanılır.

En iyi alt küme seçimi basit ve sezgisel bir yöntemdir. Ancak, değişken sayısı p büyük olduğu durumlarda olası modellerin sayısı hızla arttığından, hesaplama sorunları oluşmakta ve aşırı uyum gösterme riski artmaktadır. Bu nedenle, değişken seçimi için daha kısıtlı alternatiflerin kullanılması gerekebilmektedir.

Tüm adımsal seçim yöntemlerinde, farklı sayıda açıklayıcı değişkene sahip modeller karşılaştırılır. Bir modele parametre eklemek genellikle modeli oluşturmak için kullanılan verilere uyumu geliştirir ve daha yüksek olabilirlik verir. Farklı boyutlardaki modelleri karşılaştırmak için olabilirlik kullanılırsa aşırı uyumlu (overfitted) bir model oluşturma riskine girilir. Bu sorun, en iyi alt küme seçimi yöntemleri ile aşılabilir (Ekman, 2017).

3.2.1. Akaike Bilgi Kriteri

Veriler genellikle farklı şekillerde yorumlanabilirler. Bazen daha fazla parametreye sahip daha basit yaklaşımlar ya da gelişmiş yaklaşımlar olabilir. Birçok açıklayıcı değişken olduğunda bir sonuç değişkeni üzerindeki etkilerini modellemek için açıklayıcı değişkenlerin tamamı kullanılabilir, bu da sonuçları yorumlamayı kolaylaştırır. Bir aday listesi arasından bir model seçmek için Akaike bilgi kriteri genellikle en popüler bir yaklaşımdır (Claeskens ve Hjort, 2008).

Burada tüm veriler modele uyum sağlayacak biçimde kullanılır ve modeldeki parametre sayısı için bir ceza terimi (penalty term) getirilerek aşırı uyum göstermeden kaçınılır. Akaike bilgi kriteri, Akaike (1974) tarafından önerilmiştir ve aşağıdaki gibi tanımlanır:

$$AIC = -2\log(L\big((\hat{\beta})\big) + kp. \tag{3.1}$$

İlk terim, negatif log-olabilirliğinden oluşmaktadır. Bu, verilere iyi uyum sağlayan modeller için daha küçüktür. İkinci terim, modeldeki tahmin edicilerin sayısını k katına çıkaran bir ceza terimidir. Cox regresyon modeli için $k = 2$ olarak ele alınmaktadır. En iyi model, en düşük AIC değerine sahip model olarak tanımlanır. Bu nedenle, artırılmış cezayı giderebilmek için yeterli bilgiye sahip olması gereken modele bir açıklayıcı değişken eklenir (Ekman, 2017).

Maksimum log-olabilirliğinin örneklem miktarı ile doğrusal olarak artması ve ceza teriminin parametre sayısıyla orantılı olmasından dolayı AIC, fazla sayıda açıklayıcı değişken içeren modeller seçmektedir (Claeskens ve Hjort, 2008). Hurvich, Simonoff ve Tsai (1998), parametrik olmayan regresyon modelleri için AIC yerine düzeltilmiş (corrected) Akaike (AICC)'yi önermişlerdir. Düzeltilmiş AIC aşağıdaki gibi tanımlanmıştır:

$$AIC_C = -2\log(L((\hat{\beta})) + \frac{n(p+1)}{n-(p+2)}. \quad (3.2)$$

Burada ceza terimindeki n, toplam birim sayısını ifade etmektedir. Therneau ve Grambsch (2003) ise CRM için n yerine, durdurulmamış olay sayısı, r, ile değiştirilmesi gerektiğini belirtmişlerdir;

$$AIC_C = -2\log(L((\hat{\beta})) + \frac{r(p+1)}{r-(p+2)}. \quad (3.3)$$

Liang ve Zou (2008), küçük örneklem boyutları için AIC'yi geliştirmek için Hurvich ve Tsai (1989) tarafından regresyon modelleri için geliştirilen AIC_C seçim yöntemini yaşam çözümlemesine genişletmişler ve geliştirilmiş (improved) Akaike (AICSUR) önermişlerdir. AICSUR, n birim sayısı ve p açıklayıcı değişken sayısı olmak üzere aşağıdaki gibi tanımlanmıştır:

$$AIC_{SUR} = AIC + \frac{2(p+2)(p+3)}{n-p-3}. \quad (3.4)$$

Geliştirilen bu yöntem ile küçük örneklem boyutları için AIC'den büyük ölçüde daha iyi performans gösterdiğini; orta ve büyük örneklem boyutlarında ise AIC ile rekabet ettiğini göstermişlerdir (Liang ve Zou, 2008).

3.2.2. Bayesci Bilgi Kriteri

Model seçimine yönelik bir diğer yaklaşım Schwarz (1978) tarafından önerilen Bayesci bilgi kriteridir. Bayesci bilgi kriteri (BIC), model seçiminde dikkate alınan tüm modellerin kapsamı içinde olan gerçek bir modelin varlığını üstlendiği durumlar için geliştirilmiştir. BIC tarafından yönlendirilen seçim ile seçilen model "gerçek" veri oluşturma modeline (noktasal bir şekilde) yaklaşır. BIC aşağıdaki gibi hesaplanır:

$$BIC = -2\log(L((\hat{\beta})) + \log(n)\, p. \tag{3.5}$$

İlk terim AIC'deki ile aynıdır. Ancak ikinci terim için ceza terimi olan k değeri, log(n) ile değiştirilir; burada n birim sayısıdır. Herhangi bir $n > 7$ için $\log(n) > 2$ olduğundan, BIC istatistiği genellikle birçok açıklayıcı değişkenli modellerde AIC istatistiğinden daha ağır bir ceza verir. Bu nedenle BIC istatistiğinin kullanılması genellikle daha az açıklayıcı değişken içeren modellerde kullanılır. Sonuç olarak, herhangi bir uygun örneklem boyutu için BIC'nin ceza faktörü AIC'den daha büyüktür ve BIC daha küçük modeller seçmektedir (Ekman, 2017).

BIC'nin ceza terimi, çok sayıda açıklayıcı değişken için daha büyük ceza değeri vermektedir. Volinsky ve Raftery (2000), BIC, Cox regresyon modeline genişletmişlerdir. Gözlem sayısı n yerine durdurulmamış olay sayısını kullanmayı önermişlerdir. Düzeltilmiş BIC,

$$BIC_C = -2\log(L((\hat{\beta})) + \log(r)\, p \tag{3.6}$$

biçiminde yazılabilir. Burada r, durdurulmamış gözlem sayısını ifade etmektedir (Farooq ve Karami, 2019).

3.2.3. Çapraz Doğrulama

Bir regresyon modeli oluştururken, mevcut verileri en iyi tanımlayanı aramak olağandır. Bu yöntemde, log-olabilirlik ile ölçülen varyasyon maksimize edilir. Ancak, modelin gelecekteki verileri ne kadar iyi tahmin ettiğini değerlendirmek için tahmin değerinin ölçülmesi gerekir (Verweij ve Houwelingen, 1993).

Çapraz doğrulamada n tane birim olduğu ve verileri tanımlamak için bir regresyon modeli kullanıldığı varsayılmaktadır. Log-olabilirliği $l(\beta)$ ile gösterilir, burada β regresyon katsayısıdır. i. birimin log-olabilirliği

$$l_i(\beta) = l(\beta) - l_{(-i)}(\beta) \tag{3.7}$$

biçiminde tanımlanır. Burada $l_{(-i)}(\beta)$, i. birimin dışarıda bırakıldığındaki log-olabilirlik değeridir. $\hat{\beta}_{(-i)}$, $l_{(-i)}(\beta)$'yi maksimize eden β değeridir. Olabilirliğin bileşenleri doğrusal ve lojistik regresyon modellerinde olduğu gibi bağımsız ise $l_i(\beta)$, i. bileşen ile $\sum_{i=1}^{n} l_i(\beta) = l(\beta)$ katkısına eşittir. Çapraz doğrulanmış log-olabilirlik (cross-validated likelihood) değeri,

$$cvl = \sum_{i=1}^{n} l_i\left(\hat{\beta}_{(-i)}\right) \tag{3.8}$$

biçiminde tanımlanır (Verweij ve Houwelingen, 1993). Belirli bir model için cvl, diğer birimler kullanılarak her birimin ne kadar iyi tahmin edilebileceğini ölçer ve böylece tahmin değerinin bir ölçüsü olarak değerlendirilir.

Log-olabilirlik ve cvl arasındaki fark, varyansı bilinen doğrusal modelde gösterilmiştir. Bir sabitin dışında, $-2\log(L(\hat{\beta}))$, $AKT = \sum(y_i - X_i\hat{\beta})^2$ artık kareler toplamına eşittir, cvl ise $PRESS = \sum(y_i - X_i\hat{\beta}_{(-i)})^2$ karelerinin tahmin edilen toplamına eşittir (Verweij ve Houwelingen, 1993).

Cox regresyon modelinde $l_i(\beta)$ aşağıdaki gibi türetilir. Birimlerdeki herhangi bir başarısızlık süresi aynı olmadığında kısmi olabilirlik,

$$L(\beta) = \prod_{j=1}^{n} \left(\frac{w_j}{\sum_{k \in R_j} w_k} \right)^{d_j} \tag{3.9}$$

ile verilmektedir. i birim dışarıda bırakıldığında, i faktörü düşer ve i birimi t_i zamanından önce tüm risk setlerinden çıkarılır. t'ler $j < i$ için $t_j < t_i$ olacak şekilde sıralanırsa,

$$L_{(-i)}(\beta) = \prod_{j<i} \left(\frac{w_j}{\sum_{k \in R_j} w_k - w_i} \right)^{d_j} \prod_{j>i} \left(\frac{w_j}{\sum_{k \in R_j} w_k} \right)^{d_j} \tag{3.10}$$

olur. i biriminin $L_i(\beta)$ kısmi olabilirliğe katkısı $\frac{L(\beta)}{L_{(-i)}(\beta)}$'ye eşittir, bu da

$$L_i(\beta) = \prod_{j<i} (1 - p_{ij})^{d_j} {p_{ii}}^{d_i}$$

ile $p_{ij} = \frac{w_i}{\sum_{k \in R_j} w_k}$'ye sebep olur.

p_{ij}, risk grupları ve hayatta kalma göz önüne alındığında i. birimin t_j zamanında ölme olasılığıdır. Böylece $L_i(\beta)$, i biriminin t_{i-1}'e kadar hayatta kalmasını ifade eder ve eğer $d_i = 1$ ise, t_i zamanında başarısız olmanın koşullu olasılığıdır. Log-olabilirliğine $l_i(\beta)$ katkısı ise,

$$l_i(\beta) = \sum_{j<i} d_j \ln(1 - p_{ij}) + d_i \ln(p_{ii}) \tag{3.11}$$

biçimindedir. Açıklayıcı değişkenleri olmayan boş modelde, tüm i için $p_{ij} = 1/(n - j + 1)$ ve durdurma yoksa tüm i için $l_i(0) = -l_i(n)$ olur (Verweij ve Houwelingen, 1993).

Doğrulama Kümesi Yaklaşımı

Mevcut veriler eğitim kümesi ve doğrulama (test) kümesi olmak üzere iki gruba ayrılır. Modeli oluşturmak için eğitim kümesi kullanılır. Daha sonra bu modelden doğrulama kümesindeki veriler için çıktı değerlerini tahmin etmesi beklenir. Bunu yaparken oluşan

hataların toplamı veya ortalaması daha sonra modeli değerlendirmek için kullanılır. Modellerin performansı artık onu oluşturmak için kullanılandan başka bir veri üzerinde değerlendirildiğinden, aşırı uyum riski azaltılır (Ekman, 2017). Test hatasının doğrudan bir tahminini sağladığı ve gerçek temel model hakkında daha az varsayım yaptığından dolayı AIC, BIC'e göre daha avantajlıdır. Ayrıca model serbestlik derecelerini belirlemenin zor olduğu (örneğin, modeldeki tahmin edicilerin sayısı) veya hata varyansını, σ^2, tahmin etmenin zor olduğu durumlarda dahi daha geniş bir model seçiminde kullanılabilir (James ve ark., 2017).

Bu doğrulama kümesi yaklaşımının iki potansiyel dezavantajı vardır. İlk olarak, daha kötü performansa neden olabilecek modele uyması için kullanılan veri miktarının azaltılması; ikincisi, sırasıyla doğrulama ve eğitim kümesine hangi birimlerin dahil edildiğine bağlı olarak sonucun oldukça değişken olabilmesidir. Çapraz doğrulama, doğrulama kümesi yaklaşımıyla yakından ilgilidir ancak aynı zamanda bu dezavantajları gidermeye çalışır (Ekman, 2017).

Bir Gözlem Dışarıda Bırakılarak Çapraz Doğrulama

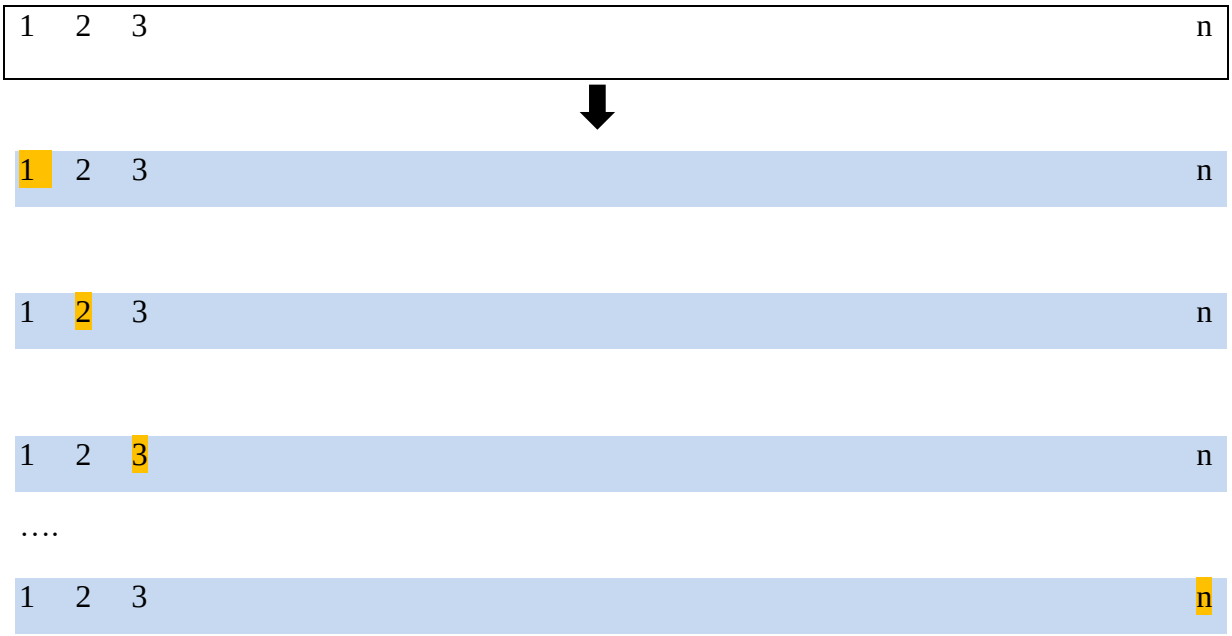
Bir gözlem dışarıda bırakılarak çapraz doğrulama (leave-one-out cross validation, LOOCV), doğrulama kümesi yaklaşımıyla yakından ilişkilidir, ancak bu yöntemin dezavantajlarını gidermeye çalışır (James ve ark., 2017). Doğrulama kümesi yaklaşımında olduğu gibi LOOCV birim kümesi iki kısma ayrılır. Ayrıca karşılaştırılabilir boyutta iki alt küme oluşturmak yerine doğrulama kümesi olarak tek bir birim (x_1, y_1) ve geri kalan birimler $\{(x_2, y_2), (x_3, y_3), \dots, (x_n, y_n)\}$ eğitim kümesi olarak kullanılır. İstatistiksel öğrenme yöntemi $n-1$ eğitim birimine uydurulur ve hariç tutulan birim için x_1 değeri kullanılarak bir y_1 tahmini yapılır. Uydurma (fitting) işleminde (x_1, y_1) kullanılmadığından, $HKO_1 = (y_1 - \hat{y}_1)^2$, test hatası için yaklaşık yansız bir tahmin sağlar. Ancak HKO_1 test hatası için yansız olsa da, tek bir birime(x_1, y_1) dayandığından büyük ölçüde değişken olduğu için zayıf bir tahmindir (James ve ark., 2017).

Doğrulama verileri için (x_2, y_2)'yi seçerek, istatistiksel öğrenme yöntemini $n-1$ birim $\{(x_1, y_1), (x_3, y_3), \dots, (x_n, y_n)\}$, üzerinde eğiterek yöntemi tekrarlayabiliriz ve $HKO_2 =$

$(y_2 - \hat{y}_2)^2$ hesaplamanır. Bu yaklaşımı n kez tekrarlamak, n kareli hata üretir, $HKO_1, HKO_2, \ldots, HKO_n$. HKO testi için LOOCV tahmini;

$$CV_{(n)} = \frac{1}{n}\sum_{i=1}^{n} HKO_i \qquad (3.12)$$

biçimindedir ve bu n test hatası tahmininin ortalamasıdır (James ve ark., 2017).



Şekil 3.1. LOOCV'nin şematik bir görüntüsü.

Şekil 3.1'de eğitim kümesi maviyle ve doğrulama kümesi turuncuyla gösterilmiştir. Bir dizi n veri noktası, biri hariç tüm birimleri içeren bir eğitim kümesine ve yalnızca bu birimi içeren bir doğrulama kümesine tekrar tekrar bölünür. Test hatası daha sonra n sonuçtaki HKO'ların ortalaması alınarak tahmin edilir. Örneğin; ilk eğitim kümesi birim 1 hariç hepsini içerir, ikinci eğitim kümesi birim 2 hariç hepsini içerir (James ve ark., 2017).

LOOCV'nin doğrulama kümesi yaklaşımına göre birkaç önemli avantajı vardır. İlk olarak, çok daha az yanlılığa sahiptir. LOOCV'de, neredeyse tüm veri kümesindeki kadar çok n - 1 birim içeren eğitim kümelerini kullanarak istatistiksel öğrenme yöntemini tekrar tekrar uygulamaktadır. Bu, eğitim kümesini tipik olarak orijinal veri kümesini yaklaşık yarısı kadar

olduğu doğrulama kümesi yaklaşımının tersidir. Sonuç olarak, LOOCV yaklaşımı, doğrulama kümesi yaklaşımının yaptığı kadar test hata oranını abartmama eğilimindedir. İkincisi, eğitim kümesi ve doğrulama kümesi bölmelerindeki rastgelelik nedeniyle tekrar tekrar uygulandığında farklı sonuçlar veren doğrulama yaklaşımının aksine, birden çok kez LOOCV gerçekleştirmek her zaman aynı sonuçları vermektedir. LOOCV'de eğitim kümesi ve doğrulama kümesi bölmelerinde rastgelelik yoktur (James ve ark., 2017).

k-Katlı Çapraz Doğrulama

LOOCV'ye bir alternatif, k-katlı (k-fold) çapraz doğrulamadır. Birim kümesi yaklaşık olarak eşit büyüklükte k gruba ayrılır. İlk kat bir doğrulama kümesi olarak ele alınır ve yöntem, kalan $k-1$ kata sığdırılır. Ortalama kare hatası HKO_1, daha sonra uzatılan katlamadaki birimler üzerinde hesaplanır. Bu yöntem k kez tekrarlanır; her seferinde farklı bir birimin grubu bir doğrulama kümesi olarak ele alınır. Bu süreç, test hatası $HKO_1, HKO_2, \ldots HKO_k$ için k tahminle sonuçlanır. k-katlı çapraz doğrulama tahmini, bu değerlerin ortalaması alınarak hesaplanır;

$$CV_{(k)} = \frac{1}{k}\sum_{i=1}^{k} HKO_i \qquad (3.13)$$

biçimindedir. Eşitlik 3.13, k-katlı çapraz doğrulama yaklaşımını göstermektedir.

LOOCV, $k = n$ olduğunda k-katlı çapraz doğrulamanın özel bir durumudur. Genellikle k = 5 veya k = 10 kullanılmaktır. Ancak LOOCV yerine k-katlı çapraz doğrulama kullanmak aslında daha düşük varyanslı bir hata tahmini verir. Bunun nedeni, LOOCV'de her modelin, bize oldukça ilişkili sonuçlar veren neredeyse aynı birim setleri üzerinde eğitilmiş olmasıdır (Ekman, 2017). Yani istatistiksel öğrenme yöntemi, LOOCV'de n kez uygulanırken; k-katlı çapraz doğrulamada, örneğin k = 10 olduğunda, 10 kat çapraz doğrulama gerçekleştirdiği anlamına gelmektedir. Bu da n çok büyükse hesaplama sorunlarına yol açabileceğinden LOOCV yerine k-katlı çapraz doğrulama kullanmanın daha avantajlı olduğunu göstermektedir (James ve ark., 2017).

3.3. Küçültme Yöntemleri

Küçültme kelimesi, ingilizcede "bir şeyin boyutunu küçültme veya küçülme süreci" anlamına gelen *shrinkage* kelimesinden gelmektedir. "Küçültme" teriminin istatistikte iki anlamı vardır: Biri olgu olarak küçültme, bir modelden elde edilen tahminlerin fazla iyimser olduğu, yani gözlemlenen sonuçların genel ortalama sonuca tahminlerden daha yakın olduğu durumu tanımlar. Diğeri ise teknik olarak küçültme, bir küçültme tahmin edicisi küçültmeyi önceden tahmin eder ve tahmin edicileri buna göre ayarlar (Heinze, Wallisch ve Dunkler, 2017). Küçültme terimi, tüm p tahmin edicilerin dahil olduğu bir modeli içerir. Bununla birlikte, tahmin edilen katsayılar, en küçük kareler tahminlerine göre sıfıra doğru küçültme işlemine dayanır. Bu küçültme aynı zamanda düzenleme (regularization) olarak da bilinir ve varyansı azaltma etkisine sahiptir. Küçültmenin türüne bağlı olarak, bazı katsayıların tam olarak sıfır olduğunu tahmin edebilmektedir. Küçültme yöntemleri ile değişken seçimi yapılabilmektedir (James ve ark., 2017).

Adımsal seçim yöntemlerinin amacı, açıklayıcı değişkenlerin bir alt kümesini içeren bir model oluşturmaktadır. Dolayısıyla yorumlanması daha kolaydır ve daha iyi tahminler vermektedir. Ancak açıklayıcı değişkenler modelin içinde veya dışında olduğu için, tahmin edilen katsayılar son derece değişken olabilmektedir. Alternatif olarak, tüm açıklayıcı değişkenleri içeren bir model kurulabilir ve bazı regresyon katsayılarını sıfıra indiren bir teknik kullanılabilmektedir. Küçültme (shrinkage) yöntemleri, Alt bölüm 3.3.1-3.3.3'de ele alınmıştır.

3.3.1. Cezalandırılmış Olabilirlik

Cezalandırılmış olabilirlik, katsayıların tahmini için regresyon katsayılarının bir ceza fonksiyonunun (penalty function) log-olabilirlikten çıkarılmasına dayanır (Verweij ve Houwelingen, 1994). Bu, cezalandırılmış olabilirlik tahmini,

$$\hat{\beta} = \underset{\beta}{^{argmin}} \left[-\log\big(L(\beta)\big) + \lambda \sum_{i=1}^{p} |\beta_i|^q \right] \tag{3.14}$$

minimize edilmesiyle elde edilmektedir (Ekman, 2017). İlk terim negatif log-olabilirlik fonksiyonudur, bu yüzden bunu küçülterek verilere iyi uyan bir model elde edilmektedir.

İkinci terim olan küçültme cezası (shrinkage penalty), katsayıların boyutu sıfıra yakın olduğunda küçüktür. λ, cezanın etkisini kontrol eden negatif olmayan ($\lambda \geq 0$) bir düzeltme parametresidir. Düzeltme parametresi, literatürde *shrinkage parameter* veya *tuning parameter* olarak geçmektedir (Koole, 2017). λ değeri ne kadar büyük olursa, küçültme o kadar fazla olmaktadır. Çapraz doğrulama gibi en iyi alt küme seçimi yöntemleri düzeltme parametresi λ için en iyi değerin seçilmesinde de kullanılmaktadır (Ekman, 2017).

3.3.2. Ridge Regresyonu Yöntemi

Ridge regresyon, açıklayıcı değişkenler arasında çoklu bağlantı olduğu durumlarda, çoklu doğrusal regresyon modellerinde alternatif bir tahmin yöntemi olarak kullanılmıştır. Çoklu bağlantı ile, en çok olabilirlik tahmin edicisinden daha küçük bir toplam hata kareler ortalamasına (mean square error - HKO) sahiptir. Çoklu bağlantı büyük olduğunda, HKO'daki azalma oldukça büyük olabilir.

Ridge regresyon tahmin edicisi ilk olarak Hoerl ve Kennard (1970a, 1970b) tarafından doğrusal regresyon modelleri için önerilmiştir ve daha sonra lojistik regresyona genelleştirilmiştir (Schafer ve ark., 1984). Ridge regresyon yöntemi, ilişkili besinler ve kolon kanseri üzerine bir araştırmaya (Smith ve diğerleri 1991) ve ilişkili PCB türdeşlerinin meme kanseri riski üzerindeki ortak etkilerine ilişkin bir araştırmaya (Holford ve diğerleri 2000) uygulanmıştır. Barker ve Brown (2001), ikili tahmin değişkenleri yüksek oranda ilişkili olduğunda, lojistik regresyon analizi için Ridge tahmin edicilerini değerlendirmek üzere bir dizi simülasyon gerçekleştirmiştir. Bu yöntem, Xue, Kim ve Shore (2007) tarafından Cox regresyon modeli için geliştirilmiştir.

Cox regresyon modeli için, Lustbader (1986), kısmi olabilirlik fonksiyonunun, bağımsız olarak örneklenmiş Poisson rastgele değişkenlerinin olabilirlik fonksiyonuna eşdeğer olduğunu göstermiştir. Bu nedenle, kısmi olabilirlik fonksiyonunu maksimize ederek en çok olabilirlik tahmin edicisi (EÇO) olan $\hat{\beta}$ elde edilir (Xue, Kim ve Shore, 2007).

Eşitlik 3.14.'de $q = 2$ olarak alındığında Ridge regresyonu yöntemi,

$$\hat{\beta}_{ridge} = \underset{\beta}{^{argmin}} \left[-\log\big(L(\beta)\big) + \lambda \sum_{i=1}^{p} {\beta_i}^2 \right] \quad (3.15)$$

elde edilmektedir (Ekman, 2017).

Doğrusal veya genelleştirilmiş doğrusal regresyon modeli için EÇO'ların aksine, Cox regresyon modeli için maksimum kısmi olabilirlik tahmin edicileri asimptotik olarak yansızdır. Huang ve Harrington (2002), cezalandırılmış kısmi olabilirlik yaklaşımlarının aslında tüm açıklayıcı değişkenlerde doğrusallık olduğunda sonlu örneklemler için EÇO üzerinde bir miktar yanlılık düzeltmesi sağladığını göstermiştir.

Ridge regresyon tahmin edicisi $\hat{\beta}_{ridge}$, artık kareler toplamını en aza indirir (Gibbons, 1981). Schaefer ve ark. (1984) $\hat{\beta}_{ridge}$'in minimum uzunluk ile ağırlıklı hata kareler toplamını (AHKT) yaklaşık olarak en aza indirdiğini göstermiştir. EÇO'nun aksine, Ridge tahmin edicileri yansız değildir. Doğrusallık olduğunda, $\hat{\beta}_{ridge}$, yanlılığı bir miktar artırır ancak varyansı önemli ölçüde azaltır, bu nedenle EÇO'dan daha küçük bir HKO'ya sahiptir (Xue, Kim ve Shore, 2007).

3.3.3. LASSO Regresyon Yöntemi

Model seçimi, regresyon katsayılarının mutlak toplamının bir katını (λ) log-olabilirlikten çıkarmaya ve böylece bazı regresyon katsayılarını sıfır olarak belirlemeye dayanan LASSO cezaları uygulanarak da gerçekleştirilebilir (Tibshirani, 1996). Adımsal seçim yöntemlerine alternatif bir yöntem olarak, Tibshirani (1996) doğrusal regresyon modelleri ve genelleştirilmiş doğrusal modeller için LASSO değişken seçim yöntemini önermiştir ve Tibshirani (1997), LASSO regresyon yöntemini Cox regresyon modeli için genişletmiştir.

En küçük kareler (EKK) tahminleri, artık kare hatasının en aza indirilmesiyle elde edilir. EKK tahmininin araştırmacılar tarafından sıklıkla tercih edilmemesinin iki nedeni vardır. Birincisi, tahmin doğruluğudur: EKK tahminleri genellikle küçük yanlılığa ancak büyük varyansa sahiptir. Tahmin doğruluğu bazen bazı katsayıları küçülterek veya sıfıra ayarlayarak iyileştirilebilir. Bunu yaparak tahmin edilen değerlerin varyansını azaltmak için bir miktar yanlılık göz ardı edilir. Bu nedenle genel tahmin doğruluğu artırılabilir. İkincisi ise

yorumlamadır. Çok sayıda tahmin edici ile, genellikle en güçlü etkiye (effect) sahip daha küçük bir alt küme belirlemek istenir (Tibshirani, 1996).

EKK tahminlerini iyileştirmeye yönelik en iyi alt küme seçimi ve ridge regresyonu kullanılmaktadır. Ancak bu iki yöntemin de sakıncaları (drawback) vardır. En iyi alt küme seçimi yorumlanabilir modeller sağlar, ancak ayrı bir süreç olduğu için son derece değişken olabilir. Yani açıklayıcı değişkenler ya modelde tutulur ya da modelden çıkarılır. Verilerdeki küçük değişiklikler, çok farklı modellerin seçilmesine neden olabilir ve bu, tahmin doğruluğunu azaltabilir. Ridge regresyon, katsayıları küçülten ve dolayısıyla daha kararlı olan sürekli bir yöntemdir. Ancak, herhangi bir katsayıyı sıfıra ayarlamadığından kolayca yorumlanabilir bir model elde edilememektedir (Tibshirani, 1996).

LASSO regresyon yöntemi, katsayıların mutlak değerlerinin toplamının belirli bir sabitten küçük olmasına bağlı olarak artık kareler toplamını en aza indirir. Bu kısıtlama nedeniyle, bu yöntem bazı katsayıları küçültür ve diğer katsayıları sıfır olarak ayarlar. Dolayısıyla, hem alt küme seçiminin hem de küçültmenin özelliklerini kullanmaktadır (Tibshirani, 1997; Koole, 2017).

Eşitlik 3.14.'de $q = 1$ olarak alındığında LASSO regresyon yöntemi

$$\hat{\beta}_{lasso} = \underset{\beta}{argmin} \left[-\log\big(L(\beta)\big) + \lambda \sum_{i=1}^{p} |\beta_i| \right] \tag{3.16}$$

elde edilmektedir (Ekman, 2017).

Ridge regresyon yöntemi ve LASSO regresyon yöntemi arasındaki temel fark, Ridge regresyonu her zaman modeldeki tüm p değişkenlerini içerirken; LASSO bazı katsayıları sıfır olarak tahmin etmekte ve dolayısıyla karşılık gelen açıklayıcı değişkenleri dahil etmemektedir (Ekman, 2017).

LASSO regresyon yöntemi, aynı anda değişken seçimi ve küçültme gerçekleştirdiği için bir düzenleme yöntemi olarak oldukça kullanışlıdır. Tüm regresyon katsayılarını sıfıra doğru küçültür ve kullanılan düzenleme (regularization) miktarına bağlı olarak birçoğunu sıfır olarak ayarlamaktadır. Bu, özellikle birimlerden daha fazla regresyon katsayısının olduğu yüksek boyutlu verilerde faydalı olabilir. Bu durumda, yorumlanabilir bir tahmin kuralı elde etmek için güçlü değişken seçimi ve değişkenlerin aşırı uyumunu önlemek için küçültme istenmektedir (Koole, 2017).

LASSO regresyon modelleri, yüksek boyutlu model seçim problemlerinde açıklayıcı değişkenlerin sayısı k, örneklem büyüklüğünü n çok aştığında yaygın olarak kullanılmıştır. Örneklem büyüklüğü küçük olduğunda ($n<k$) araştırmacılar genellikle yorumlanabilir regresyon katsayılarıyla ilgilenirler ve bu amaçla LASSO regresyon modelleri, tahmin performanslarından çok daha az anlaşılır. LASSO regresyon yöntemi tarafından tahmin edilen regresyon katsayıları amaca göre saptırılır, ancak geleneksel tahminlerden daha küçük HKO'ya sahip olabilir. Yanlılık nedeniyle, açıklayıcı veya tanımlayıcı modellerde yorumlanması zordur. Taylor ve Tibshirani (2015), LASSO regresyon yöntemi tarafından model seçiminden sonra regresyon katsayıları üzerinde geçerli çıkarım yapma sorununu araştırmış ve bir çözüm önermiştir. Bununla birlikte, bu yöntemin performansı hakkında hala yeterli kanıt yoktur (Heinze, Wallisch ve Dunkler, 2017).

LASSO regresyon tahmini ile ilgili bir diğer sorun, açıklayıcı değişkenlerin ölçeğine (scale) bağlı olmasıdır Bu nedenle, standart yazılımdaki LASSO regresyon yöntemi uygulamaları, regresyon katsayıları geri dönüştürülüp orijinal ölçekte rapor edildiğinden, yazılım uygulamalarından bazılarında kullanıcı tarafından görünmeyen birim varyansa yönelik dahili bir standardizasyon gerçekleştirir. Yine de, bu tür "herkese uyan tek boyut (one size fit all)" standardizasyonunun tüm modelleme amaçları için optimal olup olmadığı açık değildir; örneğin, sürekli ve ikili açıklayıcı değişkenlerin tipik bir karışımı düşünüldüğünde, sürekli açıklayıcı değişkenlerin farklı çarpıklığa ve ikili açıklayıcı değişkenlerin önemli ölçüde farklı denge derecelerine (degrees of balance) sahip olabileceği Porzelius, Schumacher ve Binder (2010) tarafından da ele alınmıştır (Heinze, Wallisch ve Dunkler, 2017).

4. UYGULAMA

1992-1997 yılları arasında metastaz yapmış böbrek kanseri (metastatik renal hücreli karsinom) olan 350 hasta, İngiltere'deki 31 merkezde interferon-α (IFN) ile medroksiprogesteron asetatı (MPA) tedavi türleri karşılaştırılmak üzere araştırmaya dahil edilmiştir. Fakat IFN yönteminin çeşitli etkileri ile randomize çalışmaya 347 hastaya ait veri kümesi ile devam edilmiştir Hastalara 12 hafta boyunca haftada 3 kez deri altı enjeksiyon yoluyla IFN (10MU) veya her gün ağızdan MPA (300mg) verilmiştir. Başarısızlık, ölüm olarak tanımlanmıştır. 347 hastadan 322'si (%92.79) başarısız ve 25'i (%7.21) durdurulmuştur (Royston, Sauerbrei ve Ritchie, 2004). Başarısızlık gerçekleşene kadar geçen süre, yaşam süresi olarak ele alınmıştır. Çalışmada yer alan analizler için Rstudio programı kullanılmıştır.

Yaş, cinsiyet, Dünya Sağlık Örgütü (DSÖ) performans ölçütü, teşhisten tedaviye kadar geçen süre, metastazdan tedaviye kadar geçen süre, böbrek alınması, metastatik yer, hemoglobin değerleri, beyaz kan hücresi değerleri ve tedavi türü değişkenleri incelenmiş; bu değişkenlere ait bilgiler Çizelge 4.1 ve Çizelge 4.2'de verilmiştir.

Çizelge 4.1. Kullanılan kategorik açıklayıcı değişkenler

Değişken	Değişken düzeyleri	n	Durdurulmuş Gözlemlerin Sayısı (%)	Başarısız Gözlemlerin Sayısı
Cinsiyet	0: Erkek	236	16 (%6.78)	220 (%93.22)
	1: Kadın	111	9 (%8.11)	102 (%91.89)
DSÖ performans ölçütü	0: Hastalık öncesi tüm performansı kısıtlama olmaksızın devam ya da fiziksel olarak yorucu faaliyetlerde kısıtlı	262	25 (%9.54)	238 (%90.46)
	1: Kendi kendine bakabilir ancak herhangi bir iş faaliyetinde bulunamaz	85	0 (%0.00)	85 (%100)
Böbrek alınması	0 : hayır	148	9 (%6.08)	139 (%93.92)
	1 : evet	199	16 (%8.04)	183 (%91.96)
Metastatik yer	0 : bir organda	57	4 (%7.02)	53 (%92.98)
	1 : birden fazla organda	290	21 (%7.24)	269 (%92.76)
Tedavi türü	0 : MPA	175	8 (%4.57)	167 (%95.43)
	1 : IFN	172	17 (%9.88)	155 (%90.12)

Çizelge 4.2. Kullanılan nicel açıklayıcı değişkenler

Değişkenler	**Yaş Ortalaması (±standart hata)**	**Ortanca Yaş Süresi**	**Min ve max değerleri**
Yaş (yıl)	58.62±0.54	59.75±10.12	28; 80
Teşhisten tedaviye kadar geçen süre (gün)	550.59±59.911	98±1116.02	0; 7683
Metastazdan tedaviye kadar geçen süre (gün)	130.08±22.58	36±420.68	0; 6405
Hemoglobin $(g\ dl^{-1})$	12.24±0.1	12.30±1.92	8; 18
Beyaz kan hücresi $(\times 10^{9} l^{-1})$	8.70±0.22	$8\pm4.05*10^{-9}$	3; 55

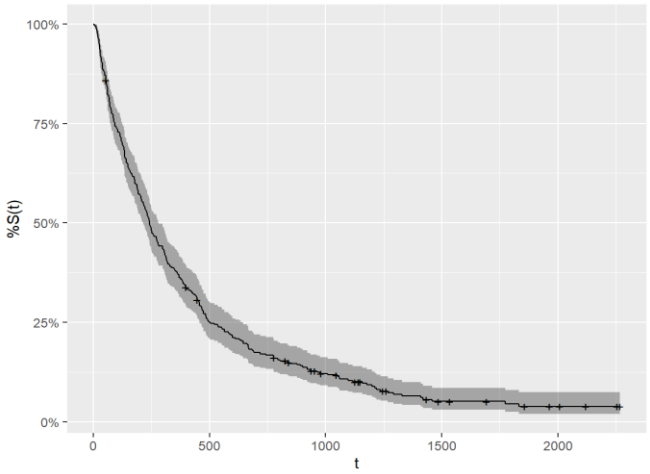
Bu çalışmada Cox regresyon modeline değişken seçim yöntemleri uygulanırken kullanılan model kurma aşamaları aşağıdaki gibi özetlenebilir:

1. Yaşam çözümlemesi için Rstudio'da *survival* paketi kullanılmıştır.
2. Böbrek kanseri veri kümesindeki değişkenlere ait Kaplan-Meier yaşam eğrileri grafikleri oluşturulmuş ve logrank sonuçları elde edilmiştir.
3. Modele tüm değişkenler dahil edilerek Cox regresyon çözümlemesi yapılmış ve orantılı tehlikeler varsayımı incelenmiştir.
4. Değişken seçim yöntemleri uygulanarak modele dahil edilecek değişkenler belirlenmiştir.
5. Kullanılan değişken seçim yöntemleri arasında bir karşılaştırma yapılarak hangi yöntemin böbrek veri kümesi için daha uygun olduğu verilmiştir.

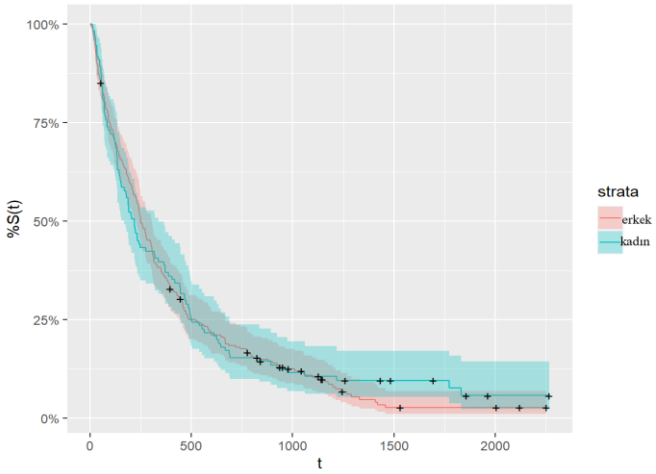
4.1. Parametrik Olmayan Yaşam Çözümlemesi Sonuçları

Böbrek kanseri veri kümesine ait K-M yaşam eğrileri elde edilmiş ve Şekil 4.1'de verilmiştir. Buna göre (a)'da K-M yaşam eğrisine bakıldığında yaşam olasılığının zaman ilerledikçe azaldığı ve (b)'de cinsiyetin yaşam olasılıkları üzerinde pek bir etkisinin olmadığı

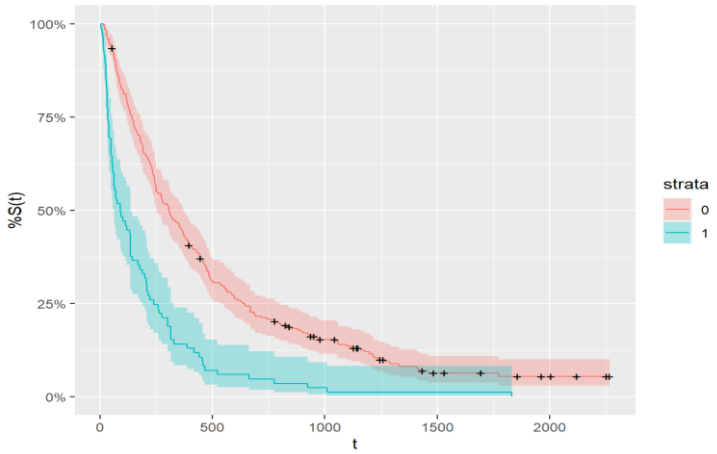
görülmektedir. (c)'de DSÖ performans ölçütü, hastalık öncesi tüm performansı kısıtlama olmaksızın devam ya da fiziksel olarak yorucu faaliyetlerde kısıtlı olan hastaların kendi kendine bakabilen ancak herhangi bir iş faaliyetinde bulunamayan hastalara göre yaşam olasılıklarından daha yüksek olduğu söylenebilir. (d)'de böbrekleri alınan hastaların yaşam olasılıklarının böbrekleri alınmayan hastalara göre daha düşük olduğu görülmektedir. (e)'de tek bir metastaza sahip hastaların yaşam olasılıklarının birden fazla metastazı olanlara göre daha yüksek olduğu söylenebilir. (f)'de tedavi türü IFN olan hastaların yaşam olasılıklarının tedavi türü MPA olan hastalara göre daha yüksek olduğu söylenebilir.



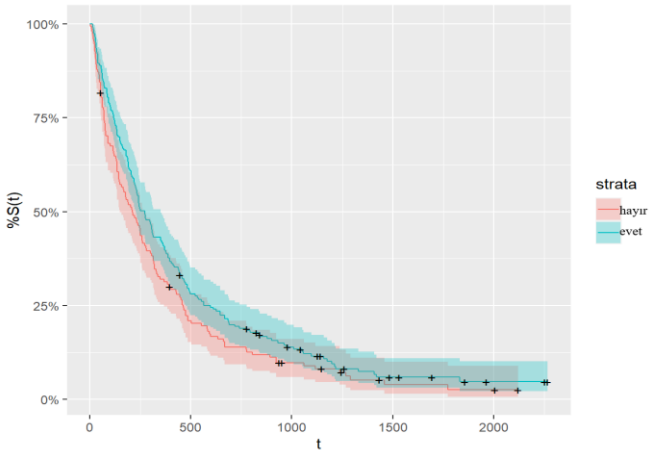
(a)



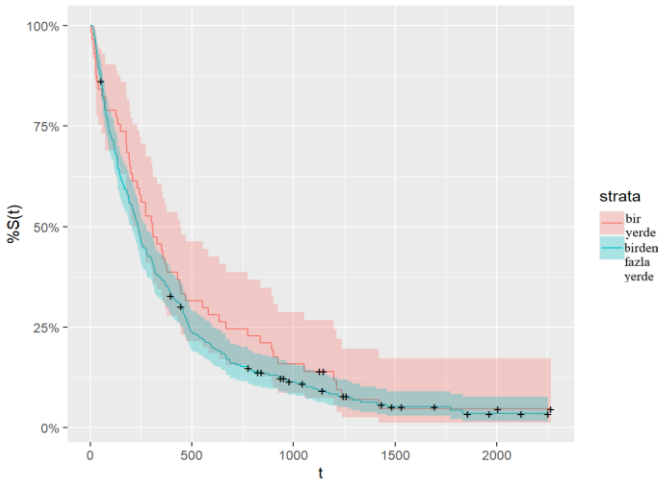
(b) Cinsiyet



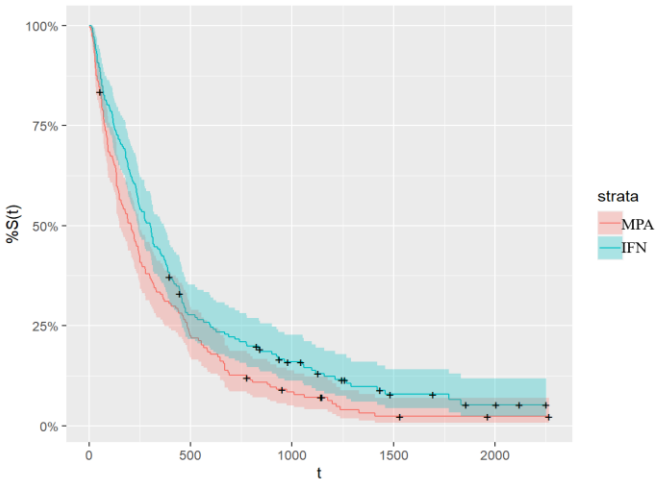
(c) DSÖ performans ölçütü



(d) Böbrek alınma durumu



(e) Metastatik yer



(f) Tedavi türü

Şekil 4.1. K-M yaşam eğrileri

Logrank testini kullanarak yaşam olasılıkları arasındaki farklılıklarını incelemek için böbrek veri kümesine ait nicel açıklayıcı değişkenler, ortanca değerleri kullanılarak kategorik değişkenlere dönüştürülmüştür. Logrank testi sonuçları Çizelge 4.2'de verilmiştir. Bu sonuçlara göre, %95 güven düzeyinde yaş (p=0.80), cinsiyet (p = 0.80) ve metastatik yer (p=0.30) değişkenleri hariç diğer tüm açıklayıcı değişkenlerin düzeyleri arasında yaşam olasılıkları açısından %95 güven düzeyinde istatistiksel olarak önemli bir fark olduğu görülmektedir.

Çizelge 4.3. Log-rank testi sonuçları

Değişken	**Değişken düzeyleri**	**Ortanca yaşam süresi (±standart hata)**	**p değeri**
Yaş	≤ 59.75	233±20.075	0.80
	> 59.75	251±28.706	
Cinsiyet	Erkek	251±20.524	0.80
	Kadın	218 ± 24.717	
DSÖ performans ölçütü	0	307 ± 32.302	<0.001
	1	90 ± 27.657	
Teşhisten tedaviye kadar geçen süre	≤ 98.60	134 ± 28.053	0.02
	> 98.60	273 ± 24.545	
Metastazdan tedaviye kadar geçen süre	≤ 35.60	210 ± 19.35	<0.001
	> 35.60	364 ± 52.833	
Böbrek alınması	Hayır	210 ± 33.488	0.06
	Evet	271 ± 25.188	
Metastatik yer	Bir yerde	305 ± 54.458	0.30
	Birden fazla yerde	234 ± 16.140	
Hemoglobin	≤ 12.50	135 ± 85.445	0.05
	> 12.50	243 ± 17.138	
Beyaz kan hücresi	≤ 8.00	315 ± 26.457	<0.001
	> 8.00	157 ± 20.266	
Tedavi türü	MPA	207 ± 27.080	0.01
	IFN	299 ± 27.323	

4.2. Cox Regresyon Modeli Sonuçları

Tüm değişkenler modelde iken CRM, böbrek kanseri veri kümesine uygulanmış ve oluşturulan modelin %95 güven düzeyinde anlamlı olduğu görülmüştür. Modele ait sonuçlar Çizelge 4.4'de verilmiştir.

Çizelge 4.4. CRM sonuçları

Değişken	$\widehat{\beta}$	Exp($\widehat{\beta}$)	Standart hata	Exp($\widehat{\beta}$) için güven aralığı	p değeri
Yaş	0.0006	1.0006	0.0057	0.9894 ; 1.0119	0.9184
Cinsiyet	-0.1152	0.8912	0.1262	0.6958 ; 1.1413	0.3613
DSÖ performans ölçütü	0.6577	1.9304	0.1881	1.4725 ; 2.5306	<0.001 *
Teşhisten tedaviye kadar geçen süre	-0.0001	0.9999	0.0001	0.9998 ; 1.0001	0.2901
Metastazdan tedaviye kadar geçen süre	-0.0004	0.9996	0.0002	0.9992 ; 1.0001	0.0992
Böbrek alınması	0.1159	1.1229	0.1340	0.8635 ; 1.4602	0.3872
Metastatik yer	-0.0272	0.9732	0.1603	0.7108 ; 1.3324	0.8652
Hemoglobin	-0.2299	0.7946	0.0359	0.7406 ; 0.8526	<0.001 *
Beyaz kan hücresi	0.0668	1.0691	0.0136	1.0410 ; 1.0980	<0.001 *
Tedavi türü	-0.3399	0.7118	0.1131	0.5704 ; 0.8884	0.0026 *

* : 0.05 anlamlılık düzeyinde istatistiksel olarak anlamlı bulunan değişkenler

Çizelge 4.4'de DSÖ performans ölçütü (p = <0.001), hemoglobin (p =< 0.001), beyaz kan hücresi (p=<0.001) ve tedavi türü (p=0.003) değişkenlerinin %95 güven düzeyinde istatistiksel olarak anlamlı olduğu görülmektedir.

CRM sonuçlarına göre, DSÖ performans ölçütü hastalık öncesi tüm performansı kısıtlama olmaksızın devam etmekte olan veya fiziksel olarak yorucu faaliyetlerde kısıtlı olan hastaların, kendi kendine bakabilen ancak herhangi bir iş faaliyetinde bulunamayan hastalara

göre 1.9304 kat daha fazla riskli olduğu söylenebilmektedir. Beyaz kan hücresi değerindeki bir birimlik artış başarısızlık riskini 1.0691 kat arttırmaktadır. Hemoglobin değerindeki bir birimlik azalış başarısızlık riskini 1.2585 kat arttırmaktadır. MPA tedavisi görenler IFN tedavisi görenlere göre 1/0.7118 = 1.4049 kat daha fazla risklidir.

CRM'nin geçerli olabilmesi için orantılı tehlikeler varsayımının incelenmesi gerekir. Bu varsayımı incelerken sıklıkla Schoenfeld artıkları yöntemi kullanılmaktadır. Bu yönteme ait sonuçlar Çizelge 4.5'de verilmiştir.

Çizelge 4.5. Schoenfeld artıkları yöntemi sonuçları

Değişken	p değeri
Yaş	0.3870
Cinsiyet	0.5687
DSÖ performans ölçütü	0.2774
Teşhisten tedaviye kadar geçen süre	0.1509
Metastazdan tedaviye kadar geçen süre	0.7540
Böbrek alınması	0.7911
Metastatik yer	0.6887
Hemoglobin	0.1105
Beyaz kan hücresi	0.1827
Tedavi türü	0.5302

Çizelge 4.5'de tüm açıklayıcı değişkenler için p değerleri 0.05 anlamlılık düzeyinden büyük olduğu için orantılı tehlikeler varsayımı sağlanmaktadır. Diğer bir deyişle, böbrek kanseri verisi için CRM'nin kullanılabileceği sonucuna ulaşılmıştır.

4.3. Değişken Seçim Yöntemlerinin Sonuçları

Böbrek kanseri veri kümesi için adımsal seçim ve en iyi alt küme seçim yöntemleri uygulanırken kullanılacak model kurma aşamaları aşağıdaki gibi özetlenebilir:

1. Adımsal seçim yöntemleri, *MASS* paketi ve en iyi alt küme seçim yöntemleri, *BeSS* paketi kullanılarak Rstudio'da uygulanmıştır.

2. Adımsal seçim yöntemlerinden AIC elde edilerek hangi değişkenlerin modele dahil edileceği belirtilmiş ve bu sonuçlar Çizelge 4.6, Çizelge 4.8, Çizelge 4.10 ve Çizelge 4.12'de verilmiştir.
3. Ayrıca her adımsal yöntem için CRM sonuçları bulunmuş ve her yöntem için aynı sonuçlar elde edilmiştir.

4.3.1. Adımsal Seçim Yöntemleri Sonuçları

Bir model oluştururken hangi açıklayıcı değişkenlerin modele dahil edilip edilmeyeceği önemli bir sorundur. Bu bölümde modelde yer alması anlamlı olan değişkenlere karar verebilmek için adımsal seçim yöntemlerinden ileriye doğru seçim, geriye doğru eleme, adımsal geriye doğru eleme ve geliştirilmiş geriye doğru eleme yöntemleri için sonuçlar elde edilmiştir.

AIC kriteri, o adımdaki modele ait AIC değerini ifade etmekte ve herhangi bir adımda elde edilen değişkenlere ait AIC değerleriyle kıyaslayarak hangi değişkenin modele dahil edileceğine karar vermektedir. Çizelge 4.6'da ileriye doğru seçim yöntemine ilişkin sonuçlar verilmiştir. AIC kriterine göre en küçük olan AIC değerine sahip olan değişken modele alınır. Bu şekilde her adım için elde edilen modeller aşağıdaki gibidir:

$$\begin{aligned}
\text{CRM} = {} & \text{yaş} + \text{cinsiyet} + \text{DSÖ performans ölçütü} \\
& + \text{teşhisten tedaviye kadar geçen süre} \\
& + \text{metastazdan tedaviye kadar geçen süre} + \text{böbrek alınması} \\
& + \text{metastatik yer} + \text{hemoglobin} + \text{beyaz kan hücresi} + \text{tedavi türü} \\
\text{Model F1} = {} & \text{hemoglobin} \\
\text{Model F2} = {} & \text{hemoglobin} + \text{beyaz kan hücresi} \\
\text{Model F3} = {} & \text{hemoglobin} + \text{beyaz kan hücresi} + \text{DSÖ performans ölçütü} \\
\text{Model F4} = {} & \text{hemoglobin} + \text{beyaz kan hücresi} + \text{DSÖ performans ölçütü} \\
& + \text{tedavi türü} \\
\text{Model F5} = {} & \text{hemoglobin} + \text{beyaz kan hücresi} + \text{DSÖ performans ölçütü} \\
& + \text{tedavi türü} + \text{metastazdan tedaviye kadar geçen süre}
\end{aligned}$$

Çizelge 4.6. İleriye doğru seçim yöntemi sonuçları

	Adım 1	**Adım 2**	**Adım 3**	**Adım 4**	**Adım 5**	**Adım 6**
AIC kriteri	3226.14	3171.80	3145.80	3125.30	3120.00	3116.20
Değişkenler	**AIC**	**AIC**	**AIC**	**AIC**	**AIC**	**AIC**
Yaş	3227.9	3173.7	3147.7	3127.3	3122.0	3118.2
Cinsiyet	3228.0	3170.9	3146.6	3126.7	3121.1	3117.6
DSÖ performans ölçütü	3184.0	3146.1	3125.3*			
Teşhisten tedaviye kadar geçen süre	3221.0	3168.0	3144.2	3124.8	3119.1	3117.5
Metastazdan tedaviye kadar geçen süre	3220.0	3166.9	3142.5	3123.1	3116.2*	
Böbrek alınması	3224.7	3173.5	3147.7	3127.3	3122.0	3118.1
Metastatik yer	3226.8	3173.6	3147.8	3127.2	3121.8	3118.2
Hemoglobin	3171.8*					
Beyaz kan hücresi	3198.9	3145.8*				
Tedavi türü	3221.3	3164.8	3141.1	3120.0*		

*: Her adımda modele dahil edilen değişkenler

Çizelge 4.6'da verilen ileriye doğru seçim yöntemi için Adım 6'da elde edilen en uygun model, Model F5 olarak adlandırılmıştır, en küçük AIC değerine sahip olduğundan en iyi modelin Model F5 olduğu söylenebilmektedir. Buna göre modelde hemoglobin, beyaz kan hücresi, DSÖ performans ölçütü, tedavi türü ve metastazdan tedaviye kadar geçen süre açıklayıcı değişkenleri yer almaktadır. Model sonuçları Çizelge 4.7'de verilmiştir.

Çizelge 4.7'de verilen CRM sonuçlarına göre, DSÖ performans ölçütü hastalık öncesi tüm performansında kısıtlama olmaksızın devam etmekte olan veya fiziksel olarak yorucu faaliyetlerde kısıtlı olan hastaların, kendi kendine bakabilen ancak herhangi bir iş faaliyetinde bulunamayan hastalara göre 1.9601 kat daha fazla riskli olduğu söylenebilmektedir. Beyaz kan hücresi değerindeki bir birimlik artış, başarısızlık riskini 1.0715 kat arttırmaktadır. Hemoglobin değerindeki bir birimlik azalış başarısızlık riskini 1/0.8041=1.2436 kat

arttırmaktadır. MPA tedavisi görenler IFN tedavisi görenlere göre 1/0.7145 = 1.3995 kat daha fazla risklidir. Metastazdan tedaviye kadar geçen süredeki bir birimlik azalış başarısızlık riskini 0.9996 kat arttırmaktadır.

Çizelge 4.7. İleriye doğru seçim yöntemi sonuçlarına göre oluşturulan CRM sonuçları

Değişken	$\widehat{\boldsymbol{\beta}}$	**Exp($\widehat{\boldsymbol{\beta}}$)**	**Standart hata**	**Exp($\widehat{\boldsymbol{\beta}}$) için güven aralığı**	**p değeri**
Hemoglobin	-0.2181	0.8041	0.0343	0.7518 ; 0.8599	<0.001
Beyaz kan hücresi	0.0690	1.0715	0.0133	1.0439 ; 1.0997	<0.001
DSÖ performans ölçütü	0.6730	1.9601	0.1365	1.5000 ; 2.5614	<0.001
Tedavi türü	-0.3362	0.7145	0.1129	0.5726 ; 0.8915	0.0029
Metastazdan tedaviye kadar geçen süre	-0.0004	0.9996	0.0002	0.9992 ; 1.0000	0.0422

Böbrek kanseri verisi için geriye doğru eleme yöntemi sonuçlarına göre elde edilen Çizelge 4.8'de AIC kriteri, herhangi bir adımda elde edilen değişkenlerin AIC değerleriyle kıyaslayarak hangi değişkenin modelden çıkarılmasına karar vermektedir. Tüm değişkenlerin bulunduğu bir model ile başlanır. AIC kriterine göre en küçük olan AIC değerine sahip olan değişken modelden çıkarılır. Bu şekilde her adım için elde edilen modeller aşağıdaki gibidir:

$$\begin{aligned}\text{CRM} = {} & \text{yaş} + \text{cinsiyet} + \text{DSÖ performans ölçütü} \\ & + \text{teşhisten tedaviye kadar geçen süre} \\ & + \text{metastazdan tedaviye kadar geçen süre} + \text{böbrek alınması} \\ & + \text{metastatik yer} + \text{hemoglobin} + \text{beyaz kan hücresi} + \text{tedavi türü}\end{aligned}$$

$$\begin{aligned}\text{Model B1} = {} & \text{cinsiyet} + \text{DSÖ performans ölçütü} + \text{teşhisten tedaviye kadar geçen süre} \\ & + \text{metastazdan tedaviye kadar geçen süreı} + \text{böbrek alınması} \\ & + \text{metastatik yer} + \text{hemoglobin} + \text{beyaz kan hücresi} + \text{tedavi türü}\end{aligned}$$

Model B2 = cinsiyet + DSÖ performans ölçütü + teşhisten tedaviye kadar geçen süre + metastazdan tedaviye kadar geçen süre + böbrek alınması + hemoglobin + beyaz kan hücresi + tedavi türü

Model B3 = cinsiyet + DSÖ performans ölçütü + teşhisten tedaviye kadar geçen süre + metastazdan tedaviye kadar geçen süre + hemoglobin + beyaz kan hücresi + tedavi türü

Model B4 = DSÖ performans ölçütü + teşhisten tedaviye kadar geçen süre + metastazdan tedaviye kadar geçen süre + hemoglobin + beyaz kan hücresi + tedavi türü

Model B5 = DSÖ performans ölçütü + metastazdan tedaviye kadar geçen süre + hemoglobin + beyaz kan hücresi + tedavi türü

Çizelge 4.8. Geriye doğru eleme yöntemi sonuçları

	Adım 1	**Adım 2**	**Adım 3**	**Adım 4**	**Adım 5**	**Adım 6**
AIC kriteri	3124.1	3122.10	3120.20	3118.90	3117.50	3116.20
Değişkenler	**AIC**	**AIC**	**AIC**	**AIC**	**AIC**	**AIC**
Yaş	3122.1*					
Cinsiyet	3122.9	3121.0	3119.0	3117.5*		
DSÖ performans ölçütü	3142.9	3141.0	3139.6	3138.1	3137.3	3136.3
Teşhisten tedaviye kadar geçen süre	3123.3	3121.3	3119.3	3117.6	3116.2*	
Metastazdan tedaviye kadar geçen süre	3125.5	3123.6	3121.6	3118.9	3119.1	3120.0
Böbrek alınması	3122.9	3120.9	3118.9*			
Metastatik yer	3122.1	3120.2*				
Hemoglobin	3163.2	3161.4	3159.7	3158.0	3156.2	3155.2
Beyaz kan hücresi	3139.7	3137.8	3135.8	3134.8	3134.4	3133.5
Tedavi türü	3131.1	3129.1	3127.2	3125.8	3124.2	3123.1

*: Her adımda modelden çıkarılan değişkenler

Çizelge 4.8'de verilen geriye doğru eleme yöntemi için Adım 6'da elde edilen en uygun model, Model B5 olarak adlandırılmıştır, en küçük AIC değerine sahip olduğundan en iyi modelin Model B5 olduğu söylenebilmektedir. Buna göre modelde DSÖ performans ölçütü, metastazdan tedaviye kadar geçen süre, hemoglobin, beyaz kan hücresi ve tedavi türü açıklayıcı değişkenleri yer almaktadır.

Böbrek kanseri veri kümesi için geriye doğru eleme yöntemi sonuçlarına göre oluşturulan CRM sonuçları Çizelge 4.9'da verilmiştir ve Çizelge 4.7'de elde edilen sonuçlar ile aynıdır.

Çizelge 4.9. Geriye doğru eleme yöntemi sonuçlarına göre oluşturulan CRM sonuçları

Değişken	$\widehat{\boldsymbol{\beta}}$	**Exp($\widehat{\boldsymbol{\beta}}$)**	**Standart hata**	**Exp($\widehat{\boldsymbol{\beta}}$) için güven aralığı**	**p değeri**
DSÖ performans ölçütü	0.6730	1.9601	0.1365	1,5000 ; 2,5614	0.00
Metastazdan tedaviye kadar geçen süre	-0.0004	0.9996	0.0002	0,9992 ; 1,0000	0.0422
Hemoglobin	-0.2181	0.8041	0.0343	0,7518 ; 0,8599	<0.001
Beyaz kan hücresi	0.0690	1.0715	0.0133	1,0439 ; 1,0997	<0.001
Tedavi türü	-0.3362	0.7145	0.1129	0,5726 ; 0,8915	0.0029

Böbrek kanseri verisi için adımsal geriye doğru eleme yöntemi sonuçlarına göre elde edilen Çizelge 4.10'da verilmiştir. AIC kriteri, herhangi bir adımda elde edilen değişkenlerin AIC değerleriyle kıyaslayarak hangi değişkenin modelden çıkarılmasına karar vermektedir. Tüm değişkenlerin bulunduğu bir model ile başlanır. AIC kriterine göre en küçük olan AIC değerine sahip olan değişken modelden çıkarılır. Fakat burada çıkarılmasına karar verilen değişkenler bir sonraki adımda yine '(+)' sembolü ile gösterilerek modeldeki etkisinin görülebilmesi için modele dahil edilmiştir ve bu değişkenlerin AIC değerleri verilmiştir. Bu şekilde her adım için elde edilen modeller aşağıdaki gibidir:

CRM = yaş + cinsiyet + DSÖ performans ölçütü
+ teşhisten hastalık tedaviye kadar geçen süre
+ metastazdan tedaviye kadar geçen süre + böbrek alınması
+ metastatik yer + hemoglobin + beyaz kan hücresi + tedavi türü

Model SB1 = cinsiyet + DSÖ performans ölçütü
+ teşhisten hastalık tedaviye kadar geçen süre
+ metastazdan tedaviye kadar geçen süre + böbrek alınması
+ metastatik yer + hemoglobin + beyaz kan hücresi + tedavi türü

Model SB2 = cinsiyet + DSÖ performans ölçütü
+ teşhisten hastalık tedaviye kadar geçen süre
+ metastazdan tedaviye kadar geçen süre + böbrek alınması
+ hemoglobin + beyaz kan hücresi + tedavi türü

Model SB3 = cinsiyet + DSÖ performans ölçütü
+ teşhisten hastalık tedaviye kadar geçen süre
+ metastazdan tedaviye kadar geçen süre + hemoglobin
+ beyaz kan hücresi + tedavi türü

Model SB4 = DSÖ performans ölçütü + teşhisten hastalık tedaviye kadar geçen süre
+ metastazdan tedaviye kadar geçen süre + hemoglobin
+ beyaz kan hücresi + tedavi türü

Model SB5 = DSÖ performans ölçütü + metastazdan tedaviye kadar geçen süre
+ hemoglobin + beyaz kan hücresi + tedavi türü

Çizelge 4.10. Adımsal geriye doğru eleme yöntemi sonuçları

	Adım 1	**Adım 2**	**Adım 3**	**Adım 4**	**Adım 5**	**Adım 6**
AIC kriteri	3124.1	3122.10	3120.20	3118.90	3117.50	3116.20
Değişkenler	**AIC**	**AIC**	**AIC**	**AIC**	**AIC**	**AIC**
Yaş	3122.1*	3124.1 **(+)**	3122.1 **(+)**	3120.8 **(+)**	3119.5 **(+)**	3118.2 **(+)**
Cinsiyet	3122.9	3121.0	3119.0	3117.5*	3118.9 **(+)**	3117.6 **(+)**
DSÖ performans ölçütü	3142.9	3141.0	3139.6	3138.1	3137.3	3136.3
Teşhisten tedaviye kadar geçen süre	3123.3	3121.3	3119.3	3117.6	3116.2*	3117.5 **(+)**
Metastazdan tedaviye kadar geçen süre	3125.5	3123.6	3121.6	3120.2	3119.1	3120.0
Böbrek alınması	3122.9	3120.9	3118.9*	3120.2 **(+)**	3119.0 **(+)**	3118.1 **(+)**
Metastatik yer	3122.1	3120.2*	3122.1 **(+)**	3120.9 **(+)**	3119.5 **(+)**	3118.2 **(+)**
Hemoglobin	3163.2	3161.4	3159.7	3158.0	3156.2	3155.2
Beyaz kan hücresi	3139.7	3137.8	3135.8	3134.8	3134.4	3133.5
Tedavi türü	3131.1	3129.1	3127.2	3125.8	3124.2	3123.1

*: Her adımda modelden çıkarılan değişkenler

Çizelge 4.10'da verilen adımsal geriye doğru eleme yöntemi için Adım 6'da elde edilen en uygun model, Model SB5 olarak adlandırılmıştır, en küçük AIC değerine sahip olduğundan en iyi modelin Model SB5 olduğu söylenebilmektedir. Buna göre modelde DSÖ performans ölçütü, metastazdan tedaviye kadar geçen süre, hemoglobin, beyaz kan hücresi ve tedavi türü açıklayıcı değişkenleri yer almaktadır.

Böbrek kanseri veri kümesi için adımsal geriye doğru seçim yöntemi sonuçlarına göre oluşturulan CRM sonuçları Çizelge 4.11'de verilmiştir. Elde edilen sonuçlar Çizelge 4.7'de elde edilen sonuçlar ile aynıdır.

Çizelge 4.11. Adımsal geriye doğru eleme yöntemi sonuçlarına göre oluşturulan CRM sonuçları

Değişken	$\widehat{\beta}$	Exp($\widehat{\beta}$)	Standart hata	Exp($\widehat{\beta}$) için güven aralığı	p değeri
Hemoglobin	-0.2174	0.8046	0.0276	0.7524 ; 0.8605	<0.001
Beyaz kan hücresi	0.0689	1.0714	0.0142	1.0439 ; 1.0997	<0.001
DSÖ performans ölçütü	0.6702	1.9545	0.2669	1.496 ; 2.5542	<0.001
Tedavi türü	-0.3348	0.7155	0.0808	0.5734 ; 0.8927	0.003
Metastazdan tedaviye kadar geçen süre	-0.0004	0.9996	0.0002	0.9992 ; 0.9999	0.043

Geliştirilmiş geriye doğru eleme yöntemi, geriye doğru eleme yönteminin standartlaştırılmış halidir. Böbrek kanseri veri kümesi için uygulanan yöntemin sonuçları Çizelge 4.12'de verilmiştir. Örneğin, Adım 1 için tüm değişkenlerin olduğu bir model oluşturularak başlanır. Yaş (p=0.9183), cinsiyet (p=0.3586), teşhisten tedaviye kadar geçen süre (p=0.2752), metastazdan tedaviye kadar geçen süre (p=0.067), böbrek alınması (p=0.3869) ve metastatik yer (p=0.8655) değişkenlerine ait kriter değerleri, 0.05 anlamlılık düzeyinden daha büyük oldukları için kara listeye (black list) alınmışlardır. Kara listeye alınan bu değerlerden en büyük olan çıkartılarak bir sonraki adıma geçilir. Bu süreç, değişkenler için kriter değerleri 0.05 anlamlılık düzeyinden küçük olduğunda sonlandırılır ve uygun model elde edilir. Bu şekilde her adım için elde edilen modeller aşağıdaki gibidir:

$$\begin{aligned}\text{CRM} = {} & \text{yaş} + \text{cinsiyet} + \text{DSÖ performans ölçütü} \\ & + \text{teşhisten hastalık tedaviye kadar geçen süre} \\ & + \text{metastazdan tedaviye kadar geçen süre} + \text{böbrek alınması} \\ & + \text{metastatik yer} + \text{hemoglobin} + \text{beyaz kan hücresi} + \text{tedavi türü}\end{aligned}$$

$$\begin{aligned}\text{Model A1} = {} & \text{cinsiyet} + \text{DSÖ performans ölçütü} \\ & + \text{teşhisten hastalık tedaviye kadar geçen süre} \\ & + \text{metastazdan tedaviye kadar geçen süre} + \text{böbrek alınması} \\ & + \text{metastatik yer} + \text{hemoglobin} + \text{beyaz kan hücresi} + \text{tedavi türü}\end{aligned}$$

Model A2 = cinsiyet + DSÖ performans ölçütü
+ teşhisten hastalık tedaviye kadar geçen süre
+ metastazdan tedaviye kadar geçen süre + böbrek alınması
+ hemoglobin + beyaz kan hücresi + tedavi türü

Model A3 = cinsiyet + DSÖ performans ölçütü
+ teşhisten hastalık tedaviye kadar geçen süre
+ metastazdan tedaviye kadar geçen süre + hemoglobin
+ beyaz kan hücresi + tedavi türü

Model A4 = DSÖ performans ölçütü + teşhisten hastalık tedaviye kadar geçen süre
+ metastazdan tedaviye kadar geçen süre + hemoglobin
+ beyaz kan hücresi + tedavi türü

Model A5 = DSÖ performans ölçütü + metastazdan tedaviye kadar geçen süre
+ hemoglobin + beyaz kan hücresi + tedavi türü

Çizelge 4.12'de verilen geliştirilmiş geriye doğru eleme yöntemi için Adım 6'da elde edilen en uygun model, Model A5 olarak adlandırılmıştır, en küçük AIC değerine sahip olduğundan en iyi modelin Model A5 olduğu söylenebilmektedir. Buna göre modelde hemoglobin, beyaz kan hücresi, DSÖ performans ölçütü, tedavi türü ve metastazdan tedaviye kadar geçen süre açıklayıcı değişkenleri yer almaktadır.

Çizelge 4.12. Geliştirilmiş geriye doğru eleme yöntemi sonuçları

	Adım 1		**Adım 2**		**Adım 3**		**Adım 4**		**Adım 5**		**Adım 6**
	Değişkenler için kriter (p değeri)	**Kara liste**	**p değeri**	**Kara liste**	**p değeri**	**Kara liste**	**p değeri**	**Kara liste**	**p değeri**	**Kara liste**	**p değeri**
Yaş	0.918	0.918*									
Cinsiyet	0.359	0.359	0.351	0.351	0.346	0.346	0.419	0.419*			
DSÖ performans ölçütü	0		0		0		0				
Teşhisten tedaviye kadar geçen süre	0.275	0.275	0.271	0.271	0.274	0.274	0.405	0.405	0.413	0.413*	
Metastazdan tedaviye kadar geçen süre	0.067	0.067	0.063	0.063	0.064	0.064	0.067	0.067	0.058	0.058	0.016
Böbrek alınması	0.387	0.387	0.386	0.386	0.396	0.396*					
Metastatik yer	0.866	0.866	0.859	0.859*							
Hemoglobin	0.00		0.00		0.00		0.00		0.00		0.00
Beyaz kan hücresi	0.00		0.00		0.00		0.00		0.00		0.00
Tedavi türü	0.003		0.003		0.003		0.003		0.003		0.003

*: Her adımda modelden çıkarılan değişkenler

Böbrek kanseri veri kümesi için geliştirilmiş geriye doğru seçim yöntemi sonuçlarına göre oluşturulan CRM sonuçları Çizelge 4.13'de verilmiştir. Elde edilen sonuçlar Çizelge 4.7'de elde edilen sonuçlar ile aynıdır.

Çizelge 4.13. Geliştirilmiş geriye doğru eleme yöntemi sonuçlarına göre oluşturulan CRM sonuçları

Değişken	$\widehat{\boldsymbol{\beta}}$	**Exp($\widehat{\boldsymbol{\beta}}$)**	**Standart hata**	**Exp($\widehat{\boldsymbol{\beta}}$) için güven aralığı**	**p değeri**
DSÖ performans ölçütü	0.6730	1.9601	0.1365	1.5000 ; 2.5614	<0.001
Metastazdan tedaviye kadar geçen süre	-0.0004	0.9996	0.0002	0.9992; 1.0000	0.0422
Hemoglobin	-0.2181	0.8041	0.0343	0.7518; 0.8599	<0.001
Beyaz kan hücresi	0.0690	1.0715	0.0133	1.0439; 1.0997	<0.001
Tedavi türü	-0.3362	0.7145	0.1129	0.5726; 0.8915	0.0029

4.3.2. Küçültme Yöntemleri Sonuçları

Küçültme yöntemleri, adımsal seçim yöntemlerine alternatif yöntemlerdir. Ridge ve LASSO regresyon modelleri kullanılarak değişkenlerin seçimi yapılır. Ridge ve LASSO regresyon modellerini oluştururken önemli bir sorun, en iyi model için düzeltme parametresi olan λ değerinin nasıl belirleneceğidir. Bunu yapmak için k-katlı çapraz doğrulama yöntemi kullanılmaktadır. Bu çalışma için k = 10 olarak belirlenmiştir (Ekman, 2017).

Böbrek kanseri veri kümesi için küçültme yöntemleri uygulanırken kullanılacak model kurma aşamaları aşağıdaki gibi özetlenebilir:

1. Küçültme yöntemleri, *glmnet* paketi kullanılarak Rstudio'da uygulanmıştır.
2. Ridge ve Lasso regresyon modelleri için 10^{-2} ile 10^{10} arasında 100 farklı λ değerleri oluşturulmuş (James ve ark., 2017) ve bu modelleri özetleyen grafik Şekil 4.2'de verilmiştir.

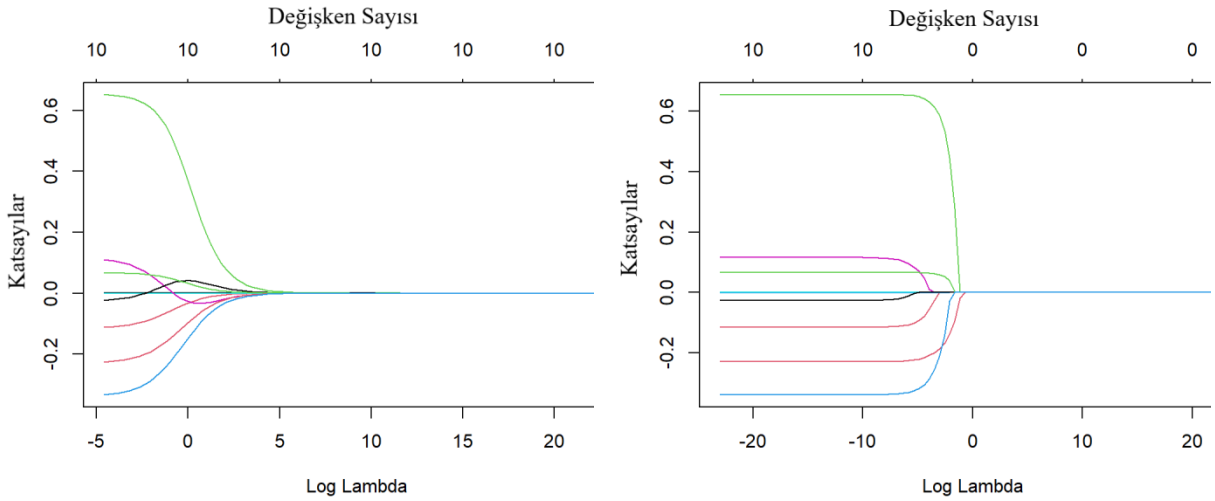
3. On-katlı çapraz doğrulama yöntemiyle λ değerleri içinden lambda.min (λ_{min}) ve lambda.1se (λ_{1se}) değerleri Çizelge 4.14'de verilmiş ve bu değerler için ideal en uygun Ridge ve Lasso modelleri oluşturulmuştur. Bu modellere ait şekiller Şekil 4.3'de verilmiştir.
4. Oluşturulan ideal Ridge ve LASSO modellerine ait katsayılar sırasıyla Çizelge 4.15 ve Çizelge 4.16'da verilmiş, hangi değişkenlerin modele dahil edilip edilmeyeceği yorumlanmıştır.
5. Oluşturulan en iyi modeller için model sonuçlarının karşılaştırılması Çizelge 4.17'de verilmiştir.

Çizelge 4.14. Ridge ve LASSO regresyon modellerine ilişkin lambda değerleri

	Ridge	**LASSO**
$\boldsymbol{\lambda_{min}}$	0.2154	0.0486
SH($\boldsymbol{\lambda_{min}}$)	0.1199	0.0991
Sıfır olmayan (Non zero) değişkenler	10	6
$\boldsymbol{\lambda_{1se}}$	2.0092	0.1233
SH($\boldsymbol{\lambda_{1se}}$)	0.1310	0.1069
Sıfır olmayan (Non zero) değişkenler	10	4

λ_{min} değeri, mininum çapraz doğrulama hatası veren ideal lambda değeridir. λ_{1se} değeri, minimum çapraz doğrulama hatasının tek-standart hata kuralına (one-standard error rule) göre en büyük lambda değeridir (Sill ve ark., 2014). Tek-standart hata kuralı, düşük hata ile en iyi modeli seçmek için farklı sayıda parametreye sahip modelleri karşılaştırmada kullanılmaktadır. Çizelge 4.14'de Ridge ve LASSO regresyon modelleri için elde edilen λ_{min} ve λ_{1se} değerleri verilmiştir. LASSO regresyon modeli için λ_{min} değeri 0.0486 bulunmuş ve bu değer ile LASSO regresyon modeli oluşturulduğunda, modelde 6 açıklayıcı değişken olduğu belirtilmiştir. Aynı şekilde λ_{1se} değeri 0.1233 bulunmuş ve bu değer ile LASSO regresyon modeli oluşturulduğunda ise modelde 4 açıklayıcı değişken olacağı belirtilmiştir. Aynı durum Ridge regresyon modeli için geçerli değildir. Ridge regresyon modelinde katsayılar sıfırlanmadığından her zaman tüm açıklayıcı değişkenler modeldedir.

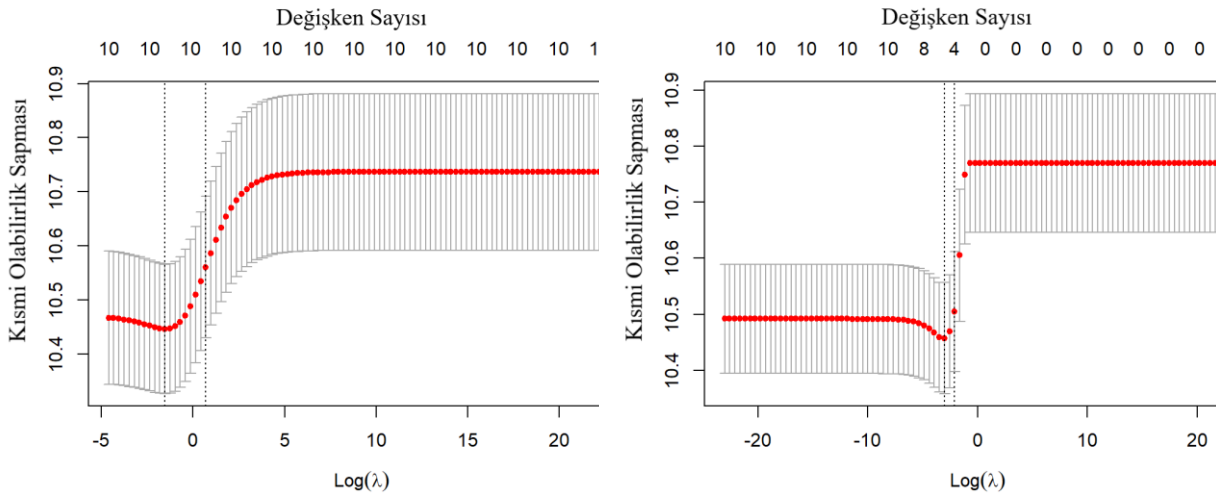
Şekil 4.2'de, (a)'da Ridge regresyon modeline ait katsayılar grafiği ve (b)'de LASSO regresyon modeline ait katsayılar grafiği verilmiştir. Bu grafikler, düzeltme parametresi λ'nın bir fonksiyonu olarak böbrek kanseri verisine ait 10 değişkene ilişkin katsayı değerlerini göstermektedir. Düzeltme parametresi λ, x ekseninde $\log\lambda$ olarak verilmiştir. Grafiklerin üzerindeki sayılar, λ değerine karşılık gelen modele dahil edilecek değişken sayısını göstermektedir. (b)'de λ değeri arttıkça bazı LASSO katsayıları 0'a doğru küçülmektedir. Böylece modele girecek değişken sayısı da azalmaktadır. Fakat (a)'da λ değeri arttıkça Ridge katsayıları sıfıra çok fazla yaklaşsa da sıfırlanmamaktadır. Bu da tüm açıklayıcı değişkenlerin her zaman modelde olduğunu göstermektedir.



(a) Ridge regresyon modeli

(b) LASSO regresyon modeli

Şekil 4.2. Küçültme yöntemlerine ait katsayılar grafikleri



(a) Ridge regresyon modeli (b) LASSO regresyon modeli

Şekil 4.3. Küçültme yöntemlerine ait kısmi olabilirlik sapması grafikleri

Şekil 4.3'de, (a)'da Ridge regresyon modeli için kısmi olabilirlik sapması (partial likelihood deviance) grafiği ve (b)'de LASSO regresyon modeli için kısmi olabilirlik sapması grafiği gösterilmektedir. Bu grafikler, model uygun olduğunda modeli değerlendirmede yardımcı olması için en uygun λ değerlerini ve çapraz doğrulanmış hatayı (cross validated error) göstermektedir. Grafiklerde soldaki dikey çizgi çapraz doğrulama hata eğrisinin minimum noktasına ulaştığı yerdir, yani λ_{min} değerini göstermektedir. Sağdaki dikey çizgi bir standart sapma içinde çapraz doğrulama hatası olan en düzenli modeli, yani λ_{1se} değerini göstermektedir (Simon ve ark., 2011).

Şekil 4.3(b)'de λ_{min} değeri minimum noktasındadır, yani modelde λ_{min} için en fazla 7 değişken ve λ_{1se} için en fazla 4 değişken olabileceğini göstermektedir. Yine de λ_{1se} değeri ise λ_{min} değerini gösteren noktadan oldukça yüksekte olduğundan model oluşturulurken genellikle λ_{min} değeri ile ilgilenilmektedir.

Çizelge 4.15. On-katlı çapraz doğrulama yönteminden elde edilen ideal lambda değeri (λ_{min}) için Ridge regresyon modeli sonuçları

Değişkenler	$\widehat{\beta}$	Exp($\widehat{\beta}$)
Yaş	0.00099	1.00099
Cinsiyet	-0.07407	0.92861
DSÖ performans ölçütü	0.57715	1.78096
Teşhisten tedaviye kadar geçen süre	-0.00025	0.99975
Metastazdan tedaviye kadar geçen süre	-0.00005	0.99995
Böbrek alınması	0.03545	1.36086
Metastatik yer	0.01896	1.01914
Hemoglobin	-0.17369	0.84056
Beyaz kan hücresi	0.05616	1.05777
Tedavi türü	-0.26169	0.76975

Çizelge 4.15'de λ_{min} değerinden elde edilen Ridge regresyon modeli katsayıları verilmektedir. Buna göre oluşturulan model,

$$\begin{aligned}\text{Ridge Model} = {} & (0.00099 * \text{yaş}) - (0.07407 * \text{cinsiyet}) \\ & + (0.57715 * \text{DSÖ performans ölçütü}) \\ & - (0.00025 * \text{metastazdan tedaviye kadar geçen süre}) \\ & - (0.00005 * \text{teşhisten tedaviye kadar geçen süre}) \\ & + (0.03545 * \text{böbrek alınması}) + (0.01896 * \text{metastatik yer}) \\ & - (0.17369 * \text{hemoglobin}) + (0.05616 * \text{beyaz kan hücresi}) - (0.26169 \\ & * \text{tedavi türü})\end{aligned}$$

biçiminde elde edilir.

Çizelge 4.15'de verilen CRM sonuçlarına göre, yaş, metastazdan tedaviye kadar geçen süre ve teşhisten tedaviye kadar geçen süre gibi açıklayıcı değişkenlerin katsayıları sıfıra çok yaklaşmıştır. Aslında bu değişkenler modelden atılmasa da etkisi oldukça azaltılmıştır. Adımsal seçim yöntemlerinde modelden çıkartılan değişkenlerle Ridge modeldeki etkisi azaltılan değişkenler benzerdir. Diğer açıklayıcı değişkenlerin sonuçlarına bakılırsa; DSÖ performans ölçütü hastalık öncesi tüm performansında kısıtlama olmaksızın devam etmekte

olan veya fiziksel olarak yorucu faaliyetlerde kısıtlı olan hastaların, kendi kendine bakabilen ancak herhangi bir iş faaliyetinde bulunamayan hastalara göre 1.78096 kat daha fazla riskli olduğu söylenebilmektedir. Beyaz kan hücresi değerindeki bir birimlik artış, başarısızlık riskini 1.05777 kat arttırmaktadır. Hemoglobin değerindeki bir birimlik azalış, başarısızlık riskini 1/0.84056=1.18968 kat arttırmaktadır. Tedavi türü MPA uygulanan hastaların başarısızlık riski, IFN uygulanan hastalara göre 1/0.76975=1.29912 kat daha fazladır.

Çizelge 4.16. On-katlı çapraz doğrulama yönteminden elde edilen ideal lambda değeri ($\boldsymbol{\lambda_{min}}$) için LASSO regresyon modeli sonuçları

Değişkenler	$\widehat{\boldsymbol{\beta}}$	**Exp($\widehat{\boldsymbol{\beta}}$)**
Yaş	.	.
Cinsiyet	.	.
DSÖ performans ölçütü	0.58589	1.79659
Teşhisten tedaviye kadar geçen süre	-0.00016	0.99984
Metastazdan tedaviye kadar geçen süre	-0.00002	0.99998
Böbrek alınması	.	.
Metastatik yer	.	.
Hemoglobin	-0.18599	0.83028
Beyaz kan hücresi	0.05884	1.06061
Tedavi türü	-0.20901	0.81139

LASSO regresyon modelini Ridge regresyon modelinden ayıran temel fark, bazı katsayıları sıfır olarak belirlediğinden karşılık gelen açıklayıcı değişkenleri modele dahil etmemektedir. Çizelge 4.16'da verilen λ_{min} değerinden elde edilen LASSO regresyon modeli verilmektedir. Buna göre yaş, cinsiyet, böbrek alınması ve metastatik yer değişkenlerine ait katsayılar sıfır olduğundan modele dahil edilmemektedirler. Buna göre oluşturulan model,

$$\text{LASSO Model} = (0.58589 * \text{DSÖ performans ölçütü}) - (0.00016 * \text{metastazdan tedaviye kadar geçen süre}) - (0.00002 * \text{teşhisten tedaviye kadar geçen süre}) - (0.18599 * \text{hemoglobin}) + (0.05884 * \text{beyaz kan hücresi}) - (0.20901 * \text{tedavi türü})$$

biçiminde elde edilir.

Çizelge 4.16'da verilen CRM sonuçlarına göre, DSÖ performans ölçütü hastalık öncesi tüm performansında kısıtlama olmaksızın devam etmekte olan veya fiziksel olarak yorucu faaliyetlerde kısıtlı olan hastaların, kendi kendine bakabilen ancak herhangi bir iş faaliyetinde bulunamayan hastalara göre 1.79659 kat daha fazla riskli olduğu söylenebilmektedir. Metastazdan tedaviye kadar geçen süredeki bir birimlik azalış başarısızlık riskini 0.99984 kat ve teşhisten tedaviye kadar geçen süredeki bir birimlik azalış başarısızlık riskini 0.99998 kat arttırmaktadır. Hemoglobin değerindeki bir birimlik azalış, başarısızlık riskini 1/0.83038=1.2043 kat arttırmaktadır. Beyaz kan hücresi değerindeki bir birimlik artış, başarısızlık riskini 1.06061 kat arttırmaktadır. Tedavi türü MPA uygulanan hastaların başarısızlık riski, IFN uygulanan hastalara göre 1/0.81139 = 1.2324 kat daha fazladır.

Çizelge 4.17'de oluşturulan nihai modellerin, değişken seçim yöntemlerinden elde edilen AIC, AIC_C , AIC_{SUR} , BIC ve BIC_C değerleri verilmiştir.

Çizelge 4.17. Değişken seçim yöntemleriyle elde edilen modellere ilişkin sonuçlar

Model	**AIC**	**AIC_C**	**AIC_{SUR}**	**BIC**	**BIC_C**
Adımsal Seçim Yöntemleri	3116.189	3099.512	3116.519	3135.062	3118.728
Ridge Model	2912.660	2904.054	2913.594	2918.063	2917.739
LASSO Model	2906.143	2900.985	2894.663	2906.569	2909.190

Adımsal seçim yöntemlerine göre beş değişken, Ridge regresyon modeline göre on değişken ve LASSO regresyon modeline altı değişken böbrek kanseri veri kümesi için modelde yer alan değişkenler olarak belirlenmiştir. Çoklu bağlantı olması durumunda değişken seçim yöntemi olarak küçültme yöntemlerinin kullanımı önerilebilir.

5. SONUÇ VE ÖNERİLER

Cox regresyon modeli yarı parametrik bir model olduğundan yaşam çözümlemesinde sıklıkla kullanılmaktadır. Yaşam verilerinin incelenmesinde model seçimine karar verirken, hangi açıklayıcı değişkenlerin kullanılacağına ve son modelin uygun olup olmadığına karar vermek oldukça önemlidir. Bu çalışmada Cox regresyon modelinde adımsal seçim yöntemi, en iyi alt küme seçim yöntemi ve küçültme yöntemleri gibi değişken seçim yöntemleri incelenerek elde edilen sonuçlara göre küçültme yöntemlerinin kullanılması önerilmiştir.

Bu tez çalışmasında 347 gözlem ve 10 açıklayıcı değişkenden oluşan böbrek kanseri veri kümesi kullanılmıştır. Veri kümesi için orantılı tehlikeler varsayımı sağlandığından CRM'nin uygulanabileceği sonucuna ulaşılmıştır. CRM sonuçlarına göre DSÖ performans ölçütü, hemoglobin değeri, beyaz kan hücresi değeri ve tedavi türü değişkenleri önemli değişkenler olarak belirlenmiştir. Adımsal seçim yöntemleri uygulanarak her yöntem için adımlar ve CRM sonuçları verilmiştir. DSÖ performans ölçütü, metastazdan tedaviye kadar geçen süre, hemoglobin, beyaz kan hücresi ve tedavi türü açıklayıcı değişkenleri modele dahil edilen değişkenler olarak belirlenmiştir ve benzer CRM sonuçlarına ulaşılmıştır. Dolayısıyla en iyi alt küme yöntemleriyle incelendiklerinde de aynı AIC ve BIC değerleri elde edilmiştir. Küçültme yöntemleri uygulanarak uygun modeller elde etmek için çapraz doğrulama yöntemi kullanılarak düzeltme parametresi λ belirlenmiştir. Bu değerler kullanılarak LASSO ve Ridge regresyon modelleri elde edilmiş ve bu modellerde hangi değişkenlerin yer alması gerektiği belirlenmiştir. Ridge regresyon modelinde bütün değişkenler modelde yer almaktadır ancak ancak yaş, metastazdan tedaviye kadar geçen süre ve teşhisten tedaviye kadar geçen süre olmak üzere üç değişkenin parametre katsayılarının sıfıra yaklaşma eğiliminde olduğu görülmüştür. LASSO regresyon modeline göre, DSÖ performans ölçütü, hemoglobin, beyaz kan hücresi, tedavi türü, metastazdan tedaviye kadar geçen süre ve teşhisten tedaviye kadar geçen süre değişkenleri istatistiksel olarak anlamlı bulunmuştur. Literatürde paket programlardaki kullanım kolaylığı nedeniyle yaygın olarak adımsal seçim yöntemleri kullanılmaktadır. Ancak çoklu bağlantı olması durumunda değişken seçim yöntemi olarak küçültme yöntemlerinin kullanılması uygun olacaktır.

Bu çalışmada doğrusal modeller için Rstudio programı kullanılarak geliştirilmiş olan küçültme yöntemleri Cox regresyon modeline adapte edilmiştir. Bundan sonraki çalışmalarda, Cox regresyon modeli için veri kümesi, eğitim kümesi ve doğrulama kümesi

biçiminde ikiye ayrılarak değişken seçim yöntemleri için makine öğrenmesi yaklaşımları kullanılabilir.

KAYNAKLAR

Akaike, H., A New Look at The Statistical Model Identification, Annals of The Institute of Statistical Mathematics, 19, 716-723, 1974.

Ata, N., Karasoy, D., Sözer, M.T., Orantılı Tehlike Varsayımının İncelenmesinde Kullanılan Yöntemler ve Bir Uygulama, Eskişehir Osmangazi Üniversitesi Mühendislik ve Mimarlık Fakültesi Dergisi, 20(1), ?, 2007.

Baker, S.G., Wax, Y., Patterson, B.H., Regression Analysis of Grouped Survival Data: Informative Censorning and Double Sampling, International Biometric Society, 49, 379-389, 1993.

Barker, L., Brown, C., Logistic Regression When Binary Predictor Variables are Highly Correlated, Stat Med, 20, 1431-1442, 2001.

Blagus, R., Babic, S., abe: Augmented Backward Elimination, R package version 3.0.1, URL : https://cran.r-project.org/web/packages/abe/abe.pdf (Erişim tarihi: 24.08.2022), 2017.

Box-Steffensmeier, J.M., Jones, B.S., Event History Modeling: A Guide for Social Scientists, Cambridge University Press, 2004.

Breslow, N.E., Covariance Analysis of Censored Survival Data, Biometrics, 30(1), 89-99, 1974.

Claeskens, G., Hjort, N.L., Model Selection and Model Averaging, Cambridge University Press, 2008.

Clark, T.G., Bradburn, M.J., Love, S.B., Altman, D.G., Survival Analysis Part I: Basic Concepts and First Analysis, British Journal of Cancer, 89, 232-238, 2003.

Collett, D., Modelling Survival Data in Medical Research, Taylor & Francis, 1994.

Cox, D. R., Regression Models and Life-Tables, Journal of the Royal Statistical Society, Series B, 34, 187-220, 1972.

Cox, D. R., Partial likelihood, Biometrika, 62, 269-276, 1975.

Dunkler, D., Plischke, M., Leffondre, K., Heinze, G., Augmented Backward Elimination: A Pragmatic and Purposeful Way to Develop Statistical Models, PloS one, 9(11), 2014.

Efron, B., The Efficiency of Cox's Likelihood Function for Censored Data, Journal of the American Statistical Association, 72(359), 312-319, 1977.

Ekman, A., Variable Selection for The Cox Proportional Hazards Model: A Simulation Study Comparing The Stepwise, Lasso and Bootstrap Approach, Yüksek lisans tezi, Umea University Department of Mathematics and Mathematical Statistics, İsveç, 2017.

Eröz, İ., Aralıklı Durdurulmuş Veriler ile Yaşam Çözümlemesi, Yüksek lisans tezi, Hacettepe Üniversitesi Fen Bilimleri Enstitüsü, Ankara, 2019.

Fan, J., Li, G., Li, R., An Overview on Variable Selection for Survival Analysis, Contemporary Multivariate Analysis and Experimental Designs- In Celebration of Professor Kai-Tai Fangs 65th Birthday, The World Scientific Publisher, 2005.

Faraooq, F.B., Karami, J.M., Model Selection Strategy for Cox Proportional Hazards Model, Dhaka University J. Sci., 67(2), 111-116, **2019.**

Gibbons, D.G., A Simulation Study of Some Ridge Estimators, J Am Stat Assoc, 76, 131-139, 1981.

Hastie, T., Tibshirani, R., Wainwright, M., Statistical Learning with Sparsity The Lasso and Generalizations, CRC Press A Chapman& Hall Book, 2015.

Hastie, T., Tibshirani, R., Friedman, R., Narasimhan, B., Tay, K., Simon, N., Qian, J., glmnet: Lasso and Elastic-Net Regularized Generalized Linear Models, R package version

4.1-2, URL: https://cran.r-project.org/web/packages/glmnet/glmnet.pdf (Erişim tarihi: 07.07.2022), 2021.

Hazra, A., Gogtay, N., Biostatistics Series Module 9: Survival Analysis, Indian Journal of Dermatology, 62(3), 251-257, 2017

Heinze, G., Wallisch, C., Dunkler, D., Variable Selection- A Review and Recommendations for The Practicing Statistician, Biometrical Journal, 60, 431-449, 2017.

Hoerl, A.E., Kennard, R.W., Ridge Regression: Biased Estimation for Nonorthogonal Problems, Technometrics, 12, 55-67, 1970a.

Hoerl, A.E., Kennard, R.W., Ridge Regression: Some Simolations, Commun Stat, 4(2), 1105-1123, 1970b.

Holford, T.R., Zheng, T., Mayne, S.H., Tessari, J.D., Boyle, P., Joint Effects of Nine Polychlorinated Biphenyl (PCB) Congoners on Breast Cancer Risk, Int J Epidemiol, 29, 975-982, 2000.

Houwelingen, H.C.V., Sauerbrei, W., Cross-validation, Shrinkage and Variable Selection in Linear Regression Revisited, Open Journal of Statistics, 3, 79-102, 2013.

Huang, J., Harrington, D., Penalized Partial Likelihood Regression for Right-Censored Data with Bootstrap Selection of the Penalty Parameter, Biometrics, 58(4), 781-791, **2002.**

Hurvich, C.M., Tsai, C.L., Regression and Time Series Model Selection in Small Samples, Biometrika, 76, 297-307, 1989.

Hurvich, C.M., Simonoff, J.S., Tsai, C.L. Smoothing Parameter Selection in Nonparametric Regression Using an Improved Akaike Information Criterion, J.R. Statistical Society B, 60(2), 271-293, 1998.

James, G., Witten, D., Hastie, T., Tibshirani, R., An Introduction to Statistical Learning with Applications in R, 8. Baskı, Springer, 2017.
Kaplan, E.L., Meier, P., Nonparametric Estimation from Incomplete Observations, Journal of The Statistical Association, 53(282), 457-481, **1958.**

Karasoy, D., Ata Tutkun, N., Yaşam Çözümlemesi, İlk Baskı, Nobel Yayıncılık, 2016.

Karasoy, D., Keskin Kaplan, S., Tied Survival Times in Survival Analysis, The Journal of Operations Research, Statistics, Econometrics and Management Information Systems, 5(1), 85-102, 2017.

Kleinbaum, D.G., Klein, M., Survival Analysis A Self-Learning Text, 3rd Edition, Springer, 2012.

Koole, R.L., Variable Selection and Shrinkage in The Cox Proportional Hazards Model, Bachelor of Science in Applied Mathematics, Delft University of Technology, Hollanda, 2017.

Lee, E.T., Wang, J.W., Statistical Methods for Survival Data Analysis, John Wiley & Sons, New Jersey, 2003.

Liang, H., Zou, G., Improved AIC Selection Strategy for Survival Analysis, Comput Stat Data Analysis, 52(5), 2538-2548, 2008.

Liu, X., Survival Analysis Models and Applications, 1st Edition, Wiley, 2012.

Machin, D., Cheung, Y.B., Parmar, M.K., Survival Analysis a Practical Approach, 2nd Edition, John Wiley & Sons, New Jersey, 2006.

Mukhopadhyay, K., Singh, R., Survival Analysis in Clinical Trials: Basics and Must Know Areas, Perspectives in Clinical Research, 2(4), 145-148, 2011.

Petersson, S., Sehlstedt, K., Variable Selection Techniques for The Cox Proportional Hazards Model: A Comparative Study, University of Gothenburg, School of Business, Economics and Law, 2018.

Porzelius, C., Schumacher, M., Binder, H.,Sparse Regression Techniques in Low-Dimensional Survival Data Settings, Statistical Computing, 20, 151-163, 2010.

Ripley, B., Venables, B., Bates, D.M., Hornik, K., Gebhardt, A., Firth, D., MASS: Support Functions and Datasets for Venables and Ripley's MASS, R package version 7.3.-58.1, URL: https://cran.r-project.org/web/packages/MASS/MASS.pdf (Erişim tarihi: 24.08.2022), 2022.

Royston, P., Sauerbrei, W., Ritchie, A., Is Treatment with Interferon-α Effective in All Patients with Metastatic Renal Carcinoma? A New Approach to the Investigation of Interactions, British Journal of Cancer, 90, 794-799, 2004.

Schafer, R.L., Roi, L.D., Wolfe, R.A., A Ridge Logistic Estimator, Commun Stat Theory Methods, 13(1), 99-113, 1984.

Schoenfeld, D., Partial Residuals for The Proportional Hazards Regression Model, Biometrika, 69(1), 239-241, 1982.

Schwarz, G., Estimating the Dimension of A Model, The Annals of Statistics, 6, 461-464, 1978.

Sill, M., Hielscher, T., Becker, N., Zucknick, M., c060: Extended Inference with Lasso and Elastic-Net Regularized Cox and Generalized Linear Models, Journal of Statistical Software, 62(5), 1-22, 2014.

Simon, N., Friedman, J., Hastie, T., Tibshirani, R., Coxnet: Regularized Cox Regression, 2011.

Smith, K.R., Slattery, M.L., French, T.K., Collinear Nutrients and The Risk of Colon Cancer, J Clin Epidemiol, 44, 715-723, 1991.

Taylor, J., Tibshirani, R.J., Statistical Learning and Selective Inference, Proceedings of The National Academy of Sciences of The United States of America, 112, 7629-7634, 2015.

Therneau, T.M., Grambsch, P.M., Penalized Survival Models and Frailty, Journal of Computational and Graphical Statistics, 12(1), 156-175, 2003.

Therneau, T.M., Lumbey, T., Elizabeth, A., Cynthia, C., survival: Survival Analysis, R package version 3.4-0, URL: https://cran.r-project.org/web/packages/survival/survival.html, (Erişim tarihi: 22.08.2022), 2022.

Tibshirani, R., Regression Shrinkage and Selection via The Lasso, J.R. Statistical Society B, 58(1), 267-288, 1996.

Tibshirani, R., The Lasso Method for Variable Selection in The Cox Model, Statistics in Medicine, 16, 385-395, 1997.

Verweij, P.J.M., Houwelingen, H.C.V., Cross-validation in Survival Analysis, Statistics in Medicine, 12, 2305-2314, 1993.

Verweij, P.J.M., Houwelingen, H.C.V., Penalized Likelihood in Cox Regression, Statistics in Medicine, 13, 2427-2436, 1994.

Volinsky, C.T, Raftery, A.E., Bayesian Information Criterion for Censored Survival Models, Biometrics, 56, 256-262, 2000.

Wen, C., Zhang, A., Quan, S., Wang, X., BeSS: Best Subset Selection in Linear, Logistic and CoxPH Models, R package version 2.0.3, URL: https://cran.r-project.org/web/packages/BeSS/BeSS.pdf (Erişim tarihi: 24.08.2022), 2021.

Xue, X., Kim, M.Y., Shore, R.E., Cox Regression Analysis in Presence of Collinearity: An Application to Assesment of Health Risks Associated with Occupational Radiation Exposure, Lifetime Data Anal, 13, 333-350, 2007.