ONDOKUZ MAYIS ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

KAYIP VERİLİ COX REGRESYON YÖNTEMİNE BAYESCİ BİR YAKLAŞIM

DOKTORA TEZİ

Nesrin ALKAN

İstatistik Anabilim Dalı

EYLÜL 2012
SAMSUN

ONDOKUZ MAYIS ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

İSTATİSTİK ANABİLİM DALI

KAYIP VERİLİ COX REGRESYON YÖNTEMİNE BAYESCİ BİR YAKLAŞIM

DOKTORA TEZİ

Nesrin ALKAN
(08210563)

Tezin Savunma Tarihi : 13 Eylül 2012

Tez Danışmanı : Doç. Dr. Yüksel TERZİ

Ondokuz Mayıs Üniversitesi Fen Bilimleri Enstitüsü

İstatistik Anabilim Dalında

Nesrin ALKAN Tarafından Hazırlanan

KAYIP VERİLİ COX REGRESYON YÖNTEMİNE BAYESCİ BİR YAKLAŞIM

başlıklı bu çalışma jürimiz tarafından 13/09/2012 tarihinde yapılan sınav ile DOKTORA tezi olarak kabul edilmiştir.

Başkan	:	**Prof. Dr. Yüksel BEK** Ondokuz Mayıs Üniversitesi
Jüri Üyeleri	:	**Doç. Dr. Yüksel TERZİ** Ondokuz Mayıs Üniversitesi
		Doç. Dr. M.Ali CENGİZ Ondokuz Mayıs Üniversitesi
		Doç. Dr. Türkan ERBAY DALKILIÇ Karadeniz Teknik Üniversitesi
		Doç. Dr. Vedat SAĞLAM Ondokuz Mayıs Üniversitesi

..../..../2012

Prof. Dr. Recep TAPRAMAZ

Enstitü Müdürü

KAYIP VERİLİ COX REGRESYON YÖNTEMİNE BAYESCİ BİR YAKLAŞIM

ÖZ

Bu çalışmada pek çok araştırmada karşılaşılan kayıp değer probleminin Cox regresyonda çözümüne yönelik yöntemler incelendi. En çok kullanılan kayıp veri analiz yöntemlerinden en iyi olanı belirlemek amacıyla farklı kayıp oranlı ve farklı örnek genişlikli veri setlerindeki performansları Cox regresyon analizi uygulanarak incelendi. Bu amaçla 50, 100, 200 birimlik sağkalım verisi kullanılarak her bir veri setinde kayıp oranları %5, %10, %20 ve %40 olacak şekilde oluşturulan veri setlerine Cox regresyon analizi uygulandı.

Kayıp değerli veri setinde Bayesci Cox regresyon ve Cox regresyon yöntemleri incelendi ve performansları farklı durumlar için karşılaştırıldı. Öncelikle bu iki yöntem gözlemlerinde kayıp değer olmayan (tam) veri setindeki performansını değerlendirmek için Ondokuz Mayıs Üniversitesi Tıp Fakültesi'nden alınan akciğer kanserli hastaların verisine ayrı ayrı uygulandı ve hastaların sağkalım süresini etkileyen risk faktörlerine ilişkin sonuçlar incelendi.

Bayesci ve klasik Cox regresyon analizinin kayıp değerli veri setindeki performansını değerlendirmek amacıyla, akciğer kanserli hastaların verisi %20 kayıp değerli olacak şekilde MAR (Tesadüfi kayıp (Missing at random)) varsayımına uygun olarak silindi ve kayıp değerli veri seti elde edildi. Bu kayıp veri setine eksiksiz veri analizi (CCA) uygulanarak kayıp değerler giderildi ve daha sonra Bayesci ve klasik Cox regresyon analizlerinin performansları değerlendirildi.

Son olarak kayıp değerli verilerin giderilmesinde üstün bir performans gösteren çoklu değer atama yöntemi ile kayıp değer problemi giderilerek, elde edilen verilere Bayesci ve klasik Cox regresyon analizi uygulandı.

Bu çalışmada Bayesci Cox regresyon analizinin iki farklı önsel dağılım kullanılarak performansı değerlendirildi. Bu önsel dağılımlar bilgilendirici ve bilgilendirici olmayan önsel dağılımdır. Bilgilendirici önseller daha önce yapılmış benzer çalışmalardan elde edildi.

Anahtar Kelimeler: Sağkalım analizi, Cox regresyon analizi, Bayesci yaklaşım, Kayıp değer

A BAYESIAN APPROACH FOR COX REGRESSION WITH MISSING DATA

ABSTRACT

This study examined the methods that resolves missing value problem encountered in many research data sets. The most commonly used missing data analysis methods examined for different missing rates and different sample size with Cox regression analysis. For this purpose 50, 100, 200 survival data were used and data sets with 5%, 10%, 20% and 40% missing rates for each data set was obtained and Cox regression analysis was used for the data sets.

Bayesian Cox regression analysis and classical Cox regression methods examined in data set with missing value and their performance was compared for different situations. The both methods was applied to patients with lung cancer data set without missing value which were obtained from Ondokuz Mayıs University, Medical Faculty.

Performance of Bayesian and classic Cox regression analysis in order to evaluate in the data sets with missing value 20% of patients with lung cancer to be deleted in accordance with the assumption of MAR (Missing at random) and so that data set with missing value was obtained. Missing data problem is resolved with complete case analysis in lung canser data with missing value. Performances of Bayesian and classic Cox regression analysis were evaluated.

Finally missing data problem in lung cancer data with missing value is resolved with multiple imputation which shows a high performance and then Bayesian and classic Cox regression analysis of the data obtained was performed separately.

Informative and noninformative priors were used separately in Bayesian Cox regression for survival data with missing value. Informative priors were obtained from a similar previous study.

Keywords: Survival analysis, Cox regression analysis, Bayesian approach, missing value

TEŞEKKÜR

Tez çalışmamın her aşamasında değerli katkı ve eleştirileriyle bana yol gösteren danışmanım Sayın Doç. Dr. Yüksel TERZİ'ye, çalışmalarım boyunca yardımlarını esirgemeyen Sayın Doç. Dr. Mehmet Ali CENGİZ'e ve Sayın Prof. Dr. Yüksel BEK'e teşekkür ederim. Ayrıca tez çalışmam sırasında göstermiş oldukları anlayış, sabır ve desteklerinden dolayı eşim, oğlum ve ailem'e teşekkür ederim.

İÇİNDEKİLER

SİMGE VE KISALTMALAR LİSTESİ

$\hat{\mu}$	: Örneklem ortalama vektörü
W^{-1}	: Ters Wishart dağılımı
$\hat{\Lambda}$	: Kareler ve çarpımlar toplamı matrisi
Σ^{*}	: Simüle edilmiş kovaryans matrisi
$\hat{\lambda}$	: Kayıp bilgi oranı
S(t)	: Sağkalım fonksiyonu
f(t)	: Olasılık yoğunluk fonksiyonu
h(t)	: Hazard fonksiyonu
H(t)	: Birikimli hazard fonksiyonu
$S_0(t)$	: Temel sağkalım fonksiyonu
$\hat{\beta}$	: En çok olabilirlik tahmini
W	: Wald testi istatistiği
S	: Skor testi istatistiği
$\pi(\theta)$	: Bağımsız değişken sayısı
$P(\theta /y)$	: Sonsal dağılım
P(y)	: Sonsal dağılımın normallik sabiti
$L(\theta)$	: Olabilirlik fonksiyonu
$I(\theta)$	: Fisher bilgi matrisi
$L_p(y/\beta)$	: β parametreli kısmi olabilirlik fonksiyonu
$\hat{\rho}_k$	: Örneklem otokorelasyonu
$\hat{\gamma}(k)$	: Örnek otokovaryans fonksiyonu
$\hat{\theta}$	: En çok olabilirlik tahmin edicisi
$\overline{\theta}$	: Sonsal ortalama
pD	: Parametrelerin etki derecesi
$Y_{göz}$	: Verinin gözlenen kısmı
Y_{kay}	: Verinin kayıp kısmı
R	: Kayıplık matrisi
$P(R / Y_{göz}, Y_{kay}, \varphi)$	: Kayıp verinin dağılımı

m	:Atanmış veri seti sayısı
z_i	:Rasgele artıklar
AIC	:Akaike bilgi kriteri
BCR	:Bayesci Cox regresyon
BÖBCR	:Bilgilendirici önselli Bayesci Cox regresyon
BOÖBCR	:Bilgilendirici olmayan önselli Bayesci Cox regresyon
BIC	:Bayesci bilgi kriteri
CCA	:Eksiksiz veri analizi (Complete Case Analysis)
CR	:Cox regresyon
DIC	:Sapma bilgi kriteri
EM	:Beklenen maksimizasyon (Expectation Maksimizasyon)
HPD	:En yüksek sonsal yoğunluk (Highest posterior density)
LR	:Olabilirlik oran
MAR	:Tesadüfi kayıp (Missing at random)
MCAR	:Tam tesadüfi kayıp (Missing completely at random)
MCMC	:Markov zinciri Monte Carlo (Markov Chain Monte Carlo)
MEAN	:Ortalama değer atama (Mean Imputation)
MI	:Çoklu değer atama (Multiple Imputation)
MNAR	:Tesadüfi olmayan kayıp (Missing not at random)
NE	:Nispi etki
REG	:Regresyon değer atama (Regression Imputation)
RR	:Nispi risk (relative risk)
SHAO	:Standart hatadaki artış oranı

ŞEKİLLER LİSTESİ

ÇİZELGELER LİSTESİ

1. GİRİŞ

Klinik ve epidemiyolojik çalışmalarda araştırmacılar sık sık farklı tedavi gruplarının karşılaştırılması ile ilgilenirler. Gruplardaki bireyler birbirleri ile ilişkili bir çok özelliğe sahiptirler. Örneğin bireyler demografik değişkenlere (yaş, cinsiyet vb), fizyolojik değişkenlere (kandaki glikoz düzeyleri, kan basıncı vb), davranışsal değişkenler (diyet, sigara içme durumu vb) gibi pek çok özelliklere sahiptirler. Böyle değişkenler bağımsız değişken veya ortak değişken (covariate) olarak adlandırılır ve bağımlı (sonuç-yanıt) değişkeni açıklamak için kullanılır. Sansürlü verilerin yer aldığı bu tür verinin modellenmesi için en çok kullanılan yöntem Cox regresyon analizidir. Bilgisayar teknolojisinin gelişmesiyle birlikte Bayesci yaklaşımlara ilgi artmış ve Bayesci Cox regresyon yöntemi geliştirilmiştir.

Bayesci sağkalım analizini Wong ve ark. (2005) Çin'deki okul öncesi çocukların diş çürümesini durdurma ile ilgili sağkalım verisinin analizinde kullanmışlardır. Calle ve ark. (2006) yağsız ve tam yağlı yoğurt ürünlerinin raf ömürleri ile ilgili çalışmada Bayesci yaklaşımdan yararlanmışlardır. Yin ve Ibrahim (2006) cilt kanseri ile ilgili sağkalım verisinin analizinde Bayesci sağkalım analizi kullanmıştır (Kurt, 2008). Bu çalışmalarda kayıp değer içermeyen sağkalım verileri kullanılarak analiz yapılmıştır.

Bayesci yaklaşımda geçmiş zamanda yapılan benzer çalışmalardan veya uzman görüşünden elde edilen önsel bilgi ve verilerden elde edilen bilginin birleştirilmesi ile sonsal dağılım elde edilir. Bu sonsal dağılımdan yararlanarak parametreler hakkında çıkarsamalarda bulunulur.

Araştırmaların pek çoğunda veri hangi yöntemle toplanırsa toplansın kayıp değerlerle karşılaşılabilir. Kayıp değerli veri araştırmayı yapan kişi için önemli bir sorundur. Çünkü istatistiksel analizler ve bilgisayar programları verinin gözlem değerlerinin tümünün ölçülebildiğini varsayar. Bu nedenle kayıp değer probleminin giderilmesi gerekir.

Kayıp değerli veri probleminin giderilmesi için bilim adamları yaklaşık yüzyıldır çalışmasına rağmen büyük gelişim 1977 yılında Dempster, Laird ve Rubin'in önerdiği EM algoritması ile olmuş ve Rubin'in 1987 yılında önerdiği çoklu değer atama yöntemleriyle kayıp değerli veri problemlerine çözümler geliştirilmiştir.

Kayıp veri analizi, verideki kayıp değerlerin giderilerek istatistiksel analizlerin yapılabileceği durumu sağlayan yöntemleri içerir. Bu nedenle kayıp veri analizi, istatistiksel analizlerin uygulanabilmesi için uygun bir zemin hazırlaması nedeniyle ön

analiz olarak da adlandırılır. Kayıp veri analizi için geliştirilmiş pek çok yöntem vardır. Bunlar kayıp değer içeren gözlemlerin silinmesini içeren yöntemler ve kayıp değerler yerine uygun değer atayan yöntemler olmak üzere iki sınıfta toplanabilir. Kayıp değerli deneği silme yoluna başvuran yöntemler liste bazlı silme (Listwise deletion) olarak da adlandırılan eksiksiz veri analizi (Complete Case Analysis) ve çiftler bazında silme (pairwise deletion) yöntemleridir. Kayıp değer yerine uygun değer atayan ve en çok kullanılan yöntemler ise ortalama değer atama (Mean Imputation), regresyonla değer atama (Regression Imputation), EM algoritması (EM Algorithm) ve çoklu değer atama (Multiple Imputation) yöntemleridir.

Chen ve ark. (2002) yarı parametrik sağkalım modellerinde kayıp değerli değişkenler için Bayesci sonuç çıkarmak için modeller önermişlerdir. Melonom kanseri klinik çalışmasını içeren gerçek veri seti, metodun gösterilmesi için incelenmiştir. Ibrahim ve ark. (2008) kayıp değerli Cox orantılı hazard modelinde değişken seçimi için hesaplama algoritması ve Bayes Metodu geliştirmişlerdir. Hemming ve Hutton (2010) kayıp değerli sağkalım verilerinin analizi için Bayes yaklaşımı sunmuşlardır. Garcia ve ark. (2010) kayıp değerli Cox regresyon modelinde değişken seçimini incelemişlerdir.

Bu çalışmada kaynaklardaki diğer çalışmalardan farklı olarak kayıp değerli sağkalım verilerin analizinde klasik ve Bayesci Cox regresyon yöntemlerinin karşılaştırılması amaçlandı.

Ayrıca bu çalışmada pek çok araştırmada karşılaşılan kayıp değer problemini gideren yöntemler incelendi. En çok kullanılan kayıp veri analiz yöntemlerinden en iyi olanı belirlemek amacıyla farklı kayıp oranlı ve farklı örnek genişlikli veri setlerindeki performansları Cox regresyon analizi uygulanarak incelendi.

Klasik yaklaşımda iyi performans gösteren kayıp veri analiz yönteminin Bayesci yaklaşımda da performansının değerlendirilmesi amaçlandı. Bu amaçla akciğer kanserli hastaların sağkalım verisi kullanılarak verideki kayıp değerlere uygun yöntemle değer atayarak veri tamamlandığında Bayesci ve klasik Cox regresyon yöntemleri karşılaştırıldı.

Ayrıca kayıp değerli değişkene ait önsel bilgi olduğunda Bayesci yaklaşımın performansını değerlendirmek amacıyla akciğer kanseri verisindeki kayıp değerlere değer atanmadan Bayesci ve klasik Cox regresyon ayrı ayrı uygulanarak sonuçları karşılaştırıldı.

Bu çalışma beş bölüm ve kaynaklardan oluşmaktadır. Çalışmanın ikinci bölümünde, sağkalım analizinde kullanılan temel tanımlar, fonksiyonlar ve fonksiyonların birbirleriyle olan ilişkileri ve sağkalım analizinin en çok kullanılan yöntemlerinden Cox regresyon modeli, modeldeki parametrelerin tahmin yöntemi ve parametrelerin önemlilik testleri verildi.

Üçüncü bölüm olan metaryal ve yöntemlerde klasik yöntemlere bir alternatif sunan Bayesci yaklaşıma ele alındı. Bayesci yaklaşımın temel unsurlarından önsel dağılımlar anlatılarak Bayesci Cox regresyon için sonsal dağılım verildi ve bu bölümde klasik yöntemlerle Bayesci yöntemlerin karşılaştırılması yapıldı. Ayrıca bu bölümde kayıp verinin tanımı, kayıp veri mekanizması, kayıp veri analiz yöntemleri verildi.

Çalışmanın dördüncü bölümü olan bulgular ve tartışma kısmında farklı örnek genişlikli ve farklı kayıp oranlı veri setleri kullanılarak kayıp veri problemine çözüm getiren yöntemlerin performansları değerlendirildi. Daha sonra akciğer kanserli hastaların verileri kullanılarak Bayesci ve klasik Cox regresyon yöntemleri veride kayıp değer yokluğunda, varlığında ve kayıp değerler çoklu değer atama yöntemi ile tahmin edildiği durumlar için incelendi.

Çalışmanın son bölümünde ise elde edilen sonuçlar tartışıldı ve önerilerde bulunuldu.

2. GENEL BİLGİLER

2.1. Sağkalım Analizi

Sağkalım analizinde amaç, istenen olayın gerçekleşmesine kadar geçen sağkalım zamanının modellenmesidir. Ömürle ilgili çalışma içeren her türlü olayda benzer modelleme çalışmaları yapılabilir. Tıp alanında bir hastalığın gelişim zamanı, tedaviye cevabı ve ölümü sıklıkla incelenen bir olaydır. Sağkalım verisi; sağkalım zamanı, hastanın belirli özellikleri (cinsiyet, yaş, kan basıncı vb.), hastalık bilgisi, tedavi bilgisi, muayene verileri ve bunun gibi daha pek çok değerlendirmeye esas teşkil eden özelliklere sahiptir. Çoğu zaman sağkalım analizinde sağ kalma olasılığı ve sağkalım süresinin ortancası tahmin edilir. Ayrıca sağkalım dağılımları ve belirlenen risk faktörleri karşılaştırılır.

2.1.1. Sansürleme

Araştırma sonlandırıldığında bütün bireyler için "istenen olay" gerçekleşmemiş olabilir. Örneğin bir klinik çalışmada çalışma sona erdiğinde bazı hastalar hala yaşıyor olabilir. Başka sebeplerden ölmüş veya başka yere taşındığı için çalışmadan ayrılmış olabilir. Bu hastalar için sağkalım süresi kesin olarak bilinmeyeceğinden bu gözlemlere sansürlü veri adı verilir. Sansürleme üç şekilde oluşur (Anderson, 2007).

i) I. Tip Sansürleme

Bazı ilaçların öldürücü dozunun hayvanlara aynı anda verildiği bir çalışma düşünülsün. Sınırlı maliyet veya kısıtlı zaman nedeniyle araştırmacı bütün bireylerin ölmesini bekleyemez ve çalışmayı erken bir şekilde sonlandırabilir. Çalışma sona erdiği anda hala yaşayan hayvanlar sansürlüdür. Onların sağkalım zamanı, en az çalışma periyodunun uzunluğu kadardır. Yani sansürlü gözlemlerin sağkalım süresi, çalışma periyodunun uzunluğuna eşittir.

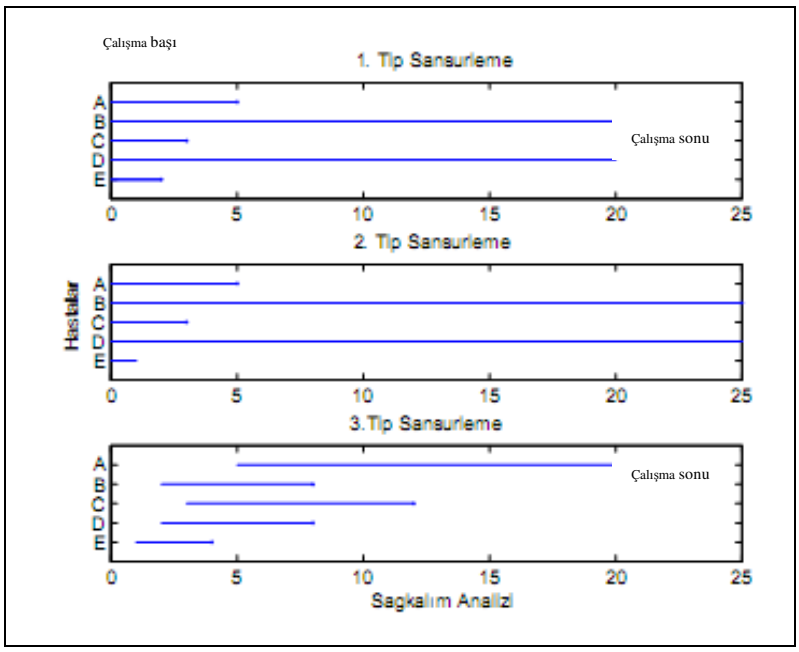
ii) II.Tip Sansürleme

Araştırmacı önceden belirlediği ölü hayvan sayısına ulaştığında çalışmayı sonlandırabilir. Böyle bir çalışmada sansürlü gözlemlerin sağkalım süresi en büyük sağkalım süreli sansürsüz gözlemin sağkalımına eşittir

iii) III. Tip sansürleme

İnsanları içeren birçok çalışmada çalışma zamanı sınırlıdır ve hastalar farklı zamanda çalışmaya girerler. Çalışma sonunda hala yaşayan sansürlü hastaların sağkalım zamanı çalışmaya girdikleri zamanla çalışma sona erdiği zaman arasındadır.

Bu bahsedilen sansürleme çeşitlerinin hepsi sağdan sansürlüdür. Birinci ve ikinci tip sansürlü gözlemler tek başına sansürlü gözlemler olarak adlandırılırken üçüncü tip sansürlü veriler kademeli veya sıralı sansürlü veri adını alırlar. Soldan sansürlü verilerde olay belli bir t zamanından önce tamamlanmıştır. Aralıklı sansürlemede ise olay t_1 ve t_2 gibi iki zaman dilimi arasında gerçekleşmektedir. Sansürleme çeşitleri Şekil 2.1'de verilmiştir.



Şekil 2.1. Sansürleme çeşitleri

2.1.2. Sağkalım Zamanının Dağılımı

Sağkalım zamanı belli bir olayın gerçekleşmesi anına kadar geçen zamanı ifade eden bir tesadüfi değişkendir. Sağkalım zamanının dağılımı, üç matematiksel fonksiyon ile tanımlanır. Bunlar sağkalım fonksiyonu, olasılık yoğunluk fonksiyonu ve hazard fonksiyonudur. Bu üç fonksiyondan biri biliniyor ise diğerleri de elde edilebilir.

2.1.2.1. Sağkalım Fonksiyonu

Bir bireyin sağkalım zamanı T ile gösterilsin. S(t), bir bireyin t zamanından daha uzun sağkalma olasılığı yani sağkalım fonksiyonudur.

$$S(t) = P(T > t) \ , \ 0 \le t < \infty \tag{2.1}$$

S(t) azalan bir fonksiyondur.

$$S(t) = \begin{cases} 1, & t = 0 \\ 0, & t = \infty \end{cases} \tag{2.2}$$

Yani zaman sıfır olduğunda sağkalım olasılığı 1'dir. Çünkü sıfır anında tüm bireyler sağdır. Ancak zaman ilerledikçe sağkalım olasılığı düşecektir, dolayısıyla zaman sonsuza doğru gittikçe artık bireyin yaşaması mümkün olmadığından sağkalım olasılığı sıfır olacaktır. Kümülatif dağılım fonksiyonu eşitlik (2.3)'de verildi.

$$F(t)=1-S(t) \tag{2.3}$$

2.1.2.2. Olasılık Yoğunluk Fonksiyonu

T sağkalım zamanının olasılık yoğunluk fonksiyonu f(t) ile gösterilir.

$$f(t) = \lim_{\Delta t \to 0} \frac{P(t < T \le t + \Delta t)}{\Delta t} \tag{2.4}$$

2.1.2.3. Hazard Fonksiyonu

Hazard fonksiyonu çok küçük zaman aralığı esnasında başarısızlığın olasılığı olarak tanımlanır. Birim zamandaki ölüm riskidir ve aşağıdaki formül ile gösterilir.

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \le T < (t + \Delta t) \mid T \ge t)}{\Delta t} \tag{2.5}$$

$$h(t) = \frac{f(t)}{1 - F(t)} \tag{2.6}$$

Birikimli hazard fonksiyonu,

$$H(t) = \int_0^t h(u)du \tag{2.7}$$

eşitliği ile bulunur.

2.1.2.4. Fonksiyonlar Arasındaki İlişki

Sağkalım fonksiyonu, olasılık yoğunluk fonksiyonu ve hazard fonksiyonları birbiriyle ilişkilidirler. Dolayısıyla bu üç fonksiyondan birisi biliniyorsa, diğerleri de bulunabilir. Bu fonksiyonlar arasındaki ilişkiler aşağıdaki gibi yazılabilir (Lee, 1992).

$$h(t) = \frac{f(t)}{S(t)} \tag{2.8}$$

Olasılık yoğunluk fonksiyonu birikimli dağılım fonksiyonunun türevi olarak tanımlanır.

$$f(t) = \frac{d}{dt}\left[1 - S(t)\right] = -S'(t) \tag{2.9}$$

Buna göre,

$$h(t) = -\frac{S'(t)}{S(t)} = -\frac{d}{dt}\log(S(t)) \tag{2.10}$$

S(0)=1 olması durumu kullanılarak 0'dan t'ye integral alınırsa,

$$-\int_0^t h(u)du = \log(S(t)) \tag{2.11}$$

elde edilir. Böylece birikimli hazard fonksiyonu,

$$H(t)=-\log(S(t)) \tag{2.12}$$

olarak bulunur.

$$S(t) = \exp[-H(t)] = \exp\left[-\int_0^t h(u)du\right] \tag{2.13}$$

$$f(t) = h(t)\exp[-H(t)] \tag{2.14}$$

Örneğin T sağkalım zamanı λ parametreli üstel dağılıma uyuyorsa olasılık yoğunluk fonksiyonu aşağıdaki gibi tanımlanır.

$$f(t) = \begin{cases} \lambda e^{-\lambda t}, t \geq 0, \lambda > 0 \\ 0 \quad , \qquad dd \end{cases}$$

birikimli dağılım fonksiyonu,

$F(t)=1-S(t)$

$$F(t) = \int_o^x \lambda e^{-\lambda t} dt = 1 - e^{-\lambda x}$$

Sağkalım fonksiyonu,

$$S(t) = e^{-\lambda t}$$

Hazard fonksiyonu,

$$h(t) = -\frac{S'(t)}{S(t)} = \lambda$$

olarak bulunur.

2.2. Cox Regresyon Modeli

Sansürlü verilerin yer aldığı sağkalım analizinde bağımlı değişken ile bağımsız değişkenler (covariate) arasındaki neden-sonuç bağıntısını ortaya koymak için yararlanılan regresyon yöntemine Cox regresyon analizi adı verilir. Sağlık alanında birçok kronik hastalıkta olduğu gibi bir hastalığın varlığı yada yokluğu tek başına sağkalım yada ölüm açısından belirleyici değildir. Hastalıktan hariç olarak bireylerin yaşı, hastalığın tipi, hastalık evresi, tedavi yöntemleri gibi. birçok faktör sağkalım süresini uzatabilir veya ölümü hızlandırabilir.

Hastalık tipi ve araştırma konusuna göre belirlenen faktörleri dikkate alarak çözüm aramak gerekir. Birçok faktörü birlikte değerlendirmede ilk akla gelen yöntem çoklu regresyon analizidir. Çoklu regresyon modelleri sonuç açısından olgular arasında zaman içinde oluşan farkı dikkate alarak değerlendirme yapma imkanına sahip değildir. Zaman içindeki değişimi dikkate alan ve sansürlü verilerle analiz yapmayı kolaylaştıran bir yönteme ihtiyaç duyulur. Bu yöntem Cox (1972) tarafından önerilen Cox regresyon modelidir. Modelde zaman ekseni boyunca her hangi iki bireyin hazard oranı (HR) sabittir. Bu yüzden Cox regresyon modeli orantılı hazard modeli (proportional hazards model) olarak da bilinir.

Bağımsız değişkenler vektörü **x** ve sağkalım süresi t olsun. Böylece bir bireyin bağımsız değişkenlere göre hazard (ölüm) fonksiyonu h(t;**x**) ile gösterilir. Bu durumda Cox Regresyon modelinin matematiksel ifadesi aşağıdaki gibidir.

$$h(t;\mathbf{x})=h_0(t)\exp(\boldsymbol{\beta}'\mathbf{x}) \qquad (2.15)$$

Burada h(t;x), hazard fonksiyonu **x**, px1 boyutlu bağımsız değişkenler vektörü $\boldsymbol{\beta}'$, 1xp boyutlu bilinmeyen regresyon katsayıları vektörü $h_0(t)$, temel hazard fonksiyonudur (baseline hazard function). Modelde yer alan $h_0(t)$, **x**'lerden bağımsız, zamana bağlı parametrik olmayan tahminleri veren bir fonksiyondur. Bu fonksiyona

temel denilmesinin nedeni **x** bağımsız değişkenleri sıfır değerini aldığında, Cox Regresyon Modeli ölüm riski olarak bilinen hazard fonksiyonuna eşit olmasıdır.
Çok değişkenli Cox regresyon modeli aşağıdaki gibidir.

$$h(t;\mathbf{x}) = h_0(t)\exp\{\sum_{i=1}^{p}\beta_i x_i\} \tag{2.16}$$

(2.16) eşitliğinde hazard fonksiyonunun, temel hazard fonksiyonuna oranı nispi risk (relative risk=RR) olarak adlandırılır.

$$RR = \frac{h(t)}{h_0(t)} = \exp\{\sum_{i=1}^{p}\beta_i x_i\} \tag{2.17}$$

(2.17) eşitliğinde $\exp\{\sum_{i=1}^{p}\beta_i x_i\}$, bağımsız değişkenlerdeki artışa karşılık, hazard oranındaki nispi olarak yüzde değişimi belirtir. Yani bağımsız değişkendeki bir birimlik artışa karşılık, hazard oranındaki yüzde değişimi ifade eder. Eşitlikte sağ tarafta t olmadığından, nispi risk tüm zaman değerleri için sabittir. Regresyon katsayıları sıfır olduğunda nispi risk bire eşit olur.
(2.17) eşitliğinde her iki tarafın logaritması alınırsa, sağ taraf lineer olacaktır.

$$\log\left[\frac{h(t)}{h_0(t)}\right] = \sum_{i=1}^{p}\beta_i x_i = \beta_1 x_1 + \ldots + \beta_p x_p \tag{2.18}$$

(2.16) eşitliğinden hareketle birikimli hazard fonksiyonu, birikimli sağkalım fonksiyonu ve olasılık yoğunluk fonksiyonu aşağıdaki gibi yazılabilir: Birikimli hazard fonksiyonu,

$$H(t;x) = \int_0^t h_0(u)\exp\left[\sum_{i=1}^{p}\beta_i x_i\right]du$$
$$= \exp\left[\sum_{i=1}^{p}\beta_i x_i\right]\int_0^t h_0(u)du \qquad (2.19)$$
$$= H_0(t)\exp\left[\sum_{i=1}^{p}\beta_i x_i\right]$$

ile verilir. Birikimli sağkalım fonksiyonu,

$$S(t;x) = \exp[-H(t,x)]$$
$$= \exp[-H_0(t)\exp(\sum\beta_i x_i)]$$
$$= [S_0(t)]^{\exp(\sum_{i=1}^{p}\beta_i\ x_i)} \qquad (2.20)$$

ile verilir. Olasılık yoğunluk fonksiyonu eşitlik (2.21)'de verildiği gibidir.

$$f(t;\mathbf{x}) = h_0(t)\exp\left(\sum_{i=1}^{p}\beta_i x_i\right)\exp\left[-H_0(t)\exp\left(\sum_{i=1}^{p}\beta_i x_i\right)\right] \quad . \qquad (2.21)$$

(2.20) eşitliğinde $S_0(t)$ temel sağkalım fonksiyonudur. Sağkalım fonksiyonu ile temel sağkalım fonksiyonu arasında üstel bir ilişki vardır.

2.2.1. Parametre Tahmini

En çok olabilirlik yöntemi parametre tahminlerinde sık kullanılan bir yöntemdir. En çok olabilirlik yöntemi olabilirlik fonksiyonunu maksimum yapan parametre tahminlerini belirler. Cox Regresyon modelinin parametre tahmininde olabilirlik fonksiyonu yerine kısmi olabilirlik fonksiyonu kullanılır. Kısmi olabilirlik kavramı, olabilirlik formülünde olasılıkların sadece tamamlanmış olan birimler için ele alınmasından kaynaklanmaktadır (Kleinbaum ve Klein, 1996).

k tane farklı başarısızlık (ölüm) sonucu $t_{(1)}<t_{(2)}<...<t_{(k)}$ olarak sıralanmış olsun. t_i zamanında risk altındaki birey sayısı $R(t_i)$ ile gösterilsin. i.nci birey için hazard fonksiyonu ile riskler seti için hazard fonksiyonların toplamının oranı riskler oranı olarak isimlendirilir ve (2.22) eşitliği ile hesaplanır.

$$\text{Riskler Oranı} = \frac{\exp(\beta' x_i)}{\sum_{j \in R(ti)} \exp(\beta' x_j)} \tag{2.22}$$

Kısmi olabilirlik fonksiyonu eşitlik (2.23) verildiği gibidir.

$$L(\beta) = \prod_{i=1}^{k} \frac{\exp(\beta' x_i)}{\sum_{j \in R(t_i)} \exp(\beta' x_j)} \tag{2.23}$$

Logaritmik kısmi olabilirlik fonksiyonu eşitlik (2.24)'de verilmiştir.

$$\ell(\beta) = \ln L(\beta) = \sum_{i=1}^{k} \beta' x_i - \sum_{i=1}^{k} \ln \left[\sum_{j \in R(t_i)} \exp(\beta' x_j) \right] \tag{2.24}$$

Bu ifadede β katsayılarına göre türev alınır. Olabilirlik fonksiyonunun birinci türevi ile eşitlik (2.25)'de verilen skor fonksiyonu elde edilir.

$$U(\beta) = \frac{\partial \ell(\beta)}{\partial \beta} = \sum_{i=1}^{k} x_i - \sum_{i=1}^{k} \frac{\sum_{j \in R(t_i)} \exp(\beta' x_j) x_j}{\sum_{j \in R(t_i)} \exp(\beta' x_j)} \tag{2.25}$$

$U(\beta) = 0$ alınarak, en çok olabilirlik tahminleri bulunabilir. Olabilirlik eşitliklerinin çözümlenerek parametre tahminlerinin belirlenmesinde iteratif yol izleyen Newton-Raphson algoritması kullanılır. Burada kullanılan kısmi olabilirlik fonksiyonu, veri setinde benzer süre gözlemleri olmadığı varsayımı altında parametre değerlerini tahmin etmektedir (Lee ve Wang, 2003). Olabilirlik fonksiyonunun ikinci türevi ise (2.26)'daki gibi bilgi matrisini verir (Kalbfleisch ve Prentice, 1980).

$$I(\beta) = -\frac{\partial^2 \ell(\beta)}{\partial \beta^2}$$

$$= \sum_{i=1}^{k} \frac{\sum_{j \in R(t_i)} \exp(\beta' x_j) x_j x_j'}{\sum_{j \in R(t_i)} \exp(\beta' x_j)} - \sum_{i=1}^{k} \frac{\left[\sum_{j \in R(t_i)} \exp(\beta' x_j) x_j\right]\left[\sum_{j \in R(t_i)} \exp(\beta' x_j) x_j'\right]}{\left[\sum_{j \in R(t_i)} \exp(\beta' x_j)\right]^2} \qquad (2.26)$$

Veri setinde benzer süreli gözlemler olduğunda logaritmik kısmi olabilirlik fonksiyonunun maksimum yapılmasında farklı yaklaşımlar önerilmiştir. Bunlar Breslow (1974) ve Efron (1975)'un önerdiği kısmi olabilirlik fonksiyonlarıdır. Bir çok uygulamada Breslow ve Efron yaklaşımlarından elde edilen parametre tahminleri arasında oldukça küçük ve önemsiz farklılıklar bulunmuştur (Lee ve Wang, 2003). Bu nedenle çalışmamızda bunlar arasında en yaygın bir yaklaşım olan Breslow kullanıldı.

Breslow'un kısmi olabilirliği eşitlik (2.27)'de verildiği gibidir.

$$L_B(\beta) = \prod_{i=1}^{k} \frac{\exp(\beta' x_i)}{\left[\sum_{j \in R(t_i)} \exp(\beta' x_j)\right]^{d_i}} \qquad (2.27)$$

Burada d_i, t_i zamanındaki ölü gözlem sayısını $R(t_i)$, t_i zamanından önce risk altındaki bireylerin sayısını göstermektedir.

β'ların en çok olabilirlik tahminleri $\hat{\beta}$ ile gösterilir. Elde edilen tahminlerin anlamlı olup olmadıklarının test edilmesi gerekmektedir.

2.2.2. Katsayıların Önemlilik Testleri

Cox Regresyon modelinde katsayılar tahmin edildikten sonra katsayıların önemliliği için $H_0 : \beta = 0$ hipotezi kısmi olabilirlik oran, Wald ve skor testleri ile test edilir (Hosmer ve Lemeshow, 1999).

2.2.2.1. Olabilirlik Oran Testi

Kurulan modelin istatistiksel olarak anlamlı olup olmadığı, olabilirlik oran testi ile test edilebilir. Bunun için her değişken için $H_0 : \beta = 0$ olan yokluk hipotezi test edilir. Test istatistiği eşitlik (2.28) de verildiği gibidir.

$$LR = -2\ln\left[\frac{L_0}{L_i}\right] \sim \chi^2_p \tag{2.28}$$

Burada L_0, ortak değişkensiz modelin olabilirlik değerini, L_i ise ortak değişkenli modelin olabilirlik değerini göstermektedir. Bulunan olabilirlik oran testi değeri (LR), p serbestlik dereceli ki-kare dağılımı gösterir. Eğer hipotez reddedilirse, en az bir regresyon katsayısı sıfırdan farklıdır yani kurulan model geçerlidir yorumu yapılır.

Modelin önemlilik testi yapıldıktan sonra, elde edilen olabilirlik fonksiyonuna en yüksek katkılı değişkenleri modele alarak yeni bir model kurmak (azaltılmış model) ve bu modeli tüm değişkenleri kapsayan tam modelle karşılaştırmak gerekir. Bu işlemler olabilirlik oran testi ile yapılır. Yokluk hipotezi ve test istatistiği aşağıdaki gibi kurulur:

H_0 : Modelden çıkarılan değişkenlere ait $\beta_i = 0$, i=1,2,…,p olmak üzere

$$LR = -2\ln\left[\frac{\text{tam model için L}}{\text{indirgenmiş model için L}}\right] \sim \chi^2_{(p-p_1)} \tag{2.29}$$

eşitliği yazılır. Burada LR, $(p\text{-}p_1)$ serbestlik dereceli ki-kare dağılımına sahiptir. Burada p tüm değişkenleri içeren modeldeki değişken sayısı, p_1 ise indirgenmiş modeldeki değişken sayısıdır. Eğer hipotez reddedilirse, değişken sayısı azaltılmış modelin, tam model kadar iyi olduğu söylenebilir. Bir başka deyişle indirgenmiş model dışında kalan değişkenlerin olabilirlik fonksiyonuna katkısı önemsizdir.

2.2.2.2. Wald Testi

Wald test istatistiği, tahmin edilen parametrenin standart hatasına oranlanması ile elde edilir ve eşitlik (2.30)'daki gibidir.

$$Z = \frac{\hat{\beta}}{SH_{\hat{\beta}}} \tag{2.30}$$

Wald istatistiği standart normal dağılım gösterir. Wald istatistiği büyük örnekler için bir serbestlik dereceli ki-kare dağılımına yaklaşır.

$$W = Z^2 = \left[\frac{\hat{\beta}}{SH_{\hat{\beta}}}\right]^2 \tag{2.31}$$

Bulunan W test istatistiği ile modele alınacak veya çıkarılacak değişkenlerin testi yapılabilir. Eğer test anlamlı çıkarsa, test edilen değişken modele alınır (Kleinbaum ve Klein, 1996).

2.2.2.3. Skor Testi

Değişkenlerin anlamlılığının test edilmesinde kullanılan diğer bir test ise skor testidir. Skor testi Wald testi ile benzer sonuçlar verir.

Skor testi, logaritmik olabilirlik istatistikleri kullanılarak bulunur. Test istatistiği,

$$S = \frac{\partial \ell(\beta) / \partial \beta}{\sqrt{I(\beta)}} \tag{2.32}$$

biçimindedir. Burada $\ell(\beta)$, (2.24) eşitliği ile verilen logaritmik kısmi olabilirlik, $I(\beta)$, (2.26) eşitliği ile verilen bilgi matrisidir. Skor istatistiği standart normal dağılım gösterir. Bu istatistiğin karesi, 1 serbestlik dereceli ki-kare dağılımı gösterir (Kleinbaum ve Klein, 1996).

MATERYAL VE YÖNTEMLER

3.1. Bayesci Yaklaşım

İstatistiksel yöntemlerden en çok kullanılan yöntemler frekansçı veya klasik yöntemler olarak bilinir. Bu yöntemler bilinmeyen parametreleri sabit varsayar ve parametreler hakkında olasılıksal ifadede bulunamazlar. Bayesci yöntemler klasik yöntemlere bir alternatif sunarlar. Bu yöntemler parametreleri rasgele değişkenler olarak ele alır ve olasılığı inanç derecesi (degrees of belief) olarak tanımlar. İnanç derecesinin anlamı, bir olayın olasılığı o olayın doğru olduğuna inanmanın derecesidir (Ibrahim ve ark. 2001). Bayesci yöntemlerde parametreler hakkında olasılıksal ifadelerde bulunabilirler. Bayesci yapının temelini Thomas Bayes'in 1763 yılında ortaya koyduğu basit Bayes teoremi oluşturmaktadır.

$P(y|\theta)$ yoğunluk fonksiyonu ile tanımlı istatistiksel bir model kullanarak $y=\{y_1,...,y_n\}$ verisinden θ tahmin edilmek istensin. Bayes felsefesi, θ parametresinin tam olarak tanımlanamayacağını ve parametre hakkındaki bu belirsizliği olasılıksal olarak ifade ederek dağılımlar üzerinden ele alır. Parametre ile ilgili belirsizliği tanımlayan en iyi dağılım normal dağılım ise θ'nın dağılımı 0 ortalamalı 1 varyanslı normal dağılımdır olarak ifade edilir. Bayesci yapının temel adımları aşağıda verildiği gibi tanımlanır.

1. θ'nın olasılık dağılımı $\pi(\theta)$ ile gösterilir. $\pi(\theta)$, önsel dağılım veya sadece önsel olarak bilinir. Önsel dağılım veriyle çalışmadan önce parametre hakkındaki araştırmacının bildiklerini ifade eder.

2. Veriler y ile gösterilir. θ bilindiği durumda y'nin dağılımını tanımlamak için $P(y/\theta)$ istatistiksel modeli seçilir.

3. $P(\theta/y)$ sonsal dağılımının hesaplanması ile veri ve önsel dağılımdan gelen bilginin birleştirilmesi ile araştırmacının θ hakkındaki bilgisi güncellenmiş olur (Congdon, 2003).

Üçüncü adım Bayes teoremin kullanılmasıyla elde edilir.

$$P(\theta / y) = \frac{P(\theta, y)}{P(y)} = \frac{P(y / \theta)\pi(\theta)}{P(y)} = \frac{P(y / \theta)\pi(\theta)}{\int P(y / \theta)\pi(\theta)d\theta} \tag{3.1}$$

Burada P(y) sonsal dağılımın normallik sabitidir. Ayrıca y'nin marjinal dağılımı olarak da tanımlanır. θ olabilirlik fonksiyonu $L(\theta)$, $P(y/\theta)$ fonksiyonu ile orantılıdır. Yani $L(\theta) \propto P(y / \theta)$ dır. Böylece Bayes teoreminin bir başka yazılımı eşitlik (3.2)'deki gibi olur.

$$P(\theta / y) = \frac{L(\theta)\pi(\theta)}{\int L(\theta)\pi(\theta)d\theta} \tag{3.2}$$

P(y), marjinal dağılımı bir integral hesabı sonucu bulunur. Bu integralin değeri sonsal dağılım hakkında herhangi ek bir bilgi sağlamaz. Böylece $P(\theta/y)$ sonsal dağılımı eşitlik (3.3)'de verildiği gibi orantılı bir form olarak ifade edilir.

$$P(\theta / y) \propto L(\theta)\pi(\theta) \tag{3.3}$$

Bayes teoremi, basit bir şekilde tanımlamak gerekirse, yeni bilgi ile mevcut bilginin nasıl güncellendiğini gösterir. Önce $\pi(\theta)$, önsel dağılımın belirlenmesi ile başlanır, daha sonra y verisinden bilgi elde edilerek θ hakkındaki bilgi güncellenir ve $P(\theta/y)$ elde edilir. Bunlar veri analizinde Bayesci yaklaşımın temel unsurlarıdır.

Teoride Bayesci yöntemler, istatistiksel çıkarımlar için klasik yöntemlere alternatif sunarlar. Bütün çıkarımlar $P(\theta/y)$ sonsal dağılımının kullanılmasıyla elde edilir. Uygulamada en ilkel problemlerde basit bir analitik çözüm ile sonsal dağılım elde edilebilirken, pek çok Bayesci analiz simülasyon yönteminin kullanımını içeren karmaşık hesaplamalar gerektirir. Sonsal dağılımdan örneklemler oluşturulur ve ilgilenilen parametrenin tahmin edilebilmesi için bu örneklemler kullanılır. Bu yöntemlerden en çok kullanılanı Gibbs örneklemedir.

Hem Bayesci hemde klasik yöntemlerin avantajları ve dezavantajları vardır. Pratik olarak yöntem seçimi, veri analizi ile gerçekleştirilmek istenen konuya bağlıdır.

Eğer önsel bilgi varsa (uzman görüşü veya geçmiş bilgi) ve bu bilgi analiz içine dahil edilmek istenirse Bayesci yöntemler kullanılır. Eğer olabilirliğe dayalı parametrelerin tahmini ile ilgilenilirse Newton Raphson gibi nümerik optimum yöntemler çok hassas tahminler verir. Böylece bu durumda Bayesci analizlerin kullanılmasına gerek yoktur.

3.1.1 Önsel Dağılımlar

Parametrenin önsel dağılımı, veri analiz edilmeden önce parametre hakkındaki belirsizliği ifade eden olasılık dağılımıdır. Önsel dağılım ve olabilirlik fonksiyonunun çarpılmasıyla parametrenin sonsal dağılımı elde edilir. Sonsal dağılım bütün çıkarımlar için kullanılır. Önsel dağılım olmadan hiçbir Bayesci çıkarım elde edilemez. Bayesci sonuç çıkarmada önselin seçimi çok önemlidir.

$\pi(\theta)$ önseli bilgilendirici olmayan önsel ise θ'nın sonsal dağılımına minimum etkide bulunur. Bilgilendirici olmayan önseller için diğer isimler dağınık (vague), belirsiz (diffuse) ve düz (flat) olarak verilir. Pek çok istatistikçi bilgilendirici olmayan önseli kullanır. Çünkü bu önseller daha objektif görünür. Fakat bilgilendirici olmayan önsellerin parametre hakkındaki belirsizliği temsil etmesini beklemek gerçekçi değildir. Bazı olaylarda bilgilendirici olmayan önseller uygun olmayan (improper) sonsallara yol açar ve uygun olmayan sonsal dağılım ile çıkarsama yapılamaz.

Sonsal dağılımın uygun olup olmadığını belirlemek için normalleştirme sabitinin bütün y'ler için sonlu (finite) olduğundan emin olmak gerekir. Bayesci çıkarımda uygun olmayan önseller sıklıkla kullanılır. Çünkü bilgilendirici olmayan önseller genellikle uygun (proper) sonsal dağılımlar oluşturur. Uygun olmayan önsel dağılımlar, uygun olmayan sonsal dağılımlara da neden olabilir. Eğer uygun olmayan önsel dağılım uygun olmayan sonsal dağılıma yol açıyorsa, uygun olmayan sonsal dağılıma dayalı çıkarımlar geçerli olmayacaktır.

Bayesci sağkalım analizinde uygun olmayan önsellerin kullanılması modelde herhangi bir uygun olmayan sonsal dağılıma yol açmaz (SAS, 2011).

Bilgilendirici (informative) önseller olabilirliğin egemenliği altında kalmayan önsellerdir ve sonsal dağılımda önemli bir etkiye sahiptir. Bu tip önsel dağılımlar gerçek uygulamalarda dikkatli bir şekilde belirlenmesi gerekir. Önsel dağılımın uygun kullanılması Bayesci yöntemin gücünü gösterir.

Bayesci sağkalım analizinde genellikle β için bilgilendirici önselin seçiminde normal önsel, bilgilendirici olmayan önselin seçiminde ise düzgün önsel kullanılır (Ibrahim ve ark.,2001).

3.1.1.1. Eşlenik (Conjugate) Önseller

Eğer önsel ve sonsal dağılımlar aynı dağılım ailesinden ise bu tür önsele eşlenik önsel denir. Örneğin eğer olabilirlik binom, eşlenik önsel beta dağılıyorsa θ'nın sonsal dağılımı da beta dağılımı olacaktır. Diğer sık kullanılan eşlenik/olabilirlik kombinasyonları Normal/Normal, Gamma/Poisson, Gamma/Gamma ve Gamma/Beta'dır. Eşlenik önsellerin kullanılması, sonsal dağılımların elde edilmesinde pratik bir yoldur.

3.1.1.2. Jeffreys'in Önseli

Jeffreys'in önseli çok kullanışlı bir önseldir. Yerel tekdüze (local uniformity) özelliğini sağlar. Yani aralık dışındaki büyük değerleri kabul etmeyen ve olabilirliğin önemli olduğu bölgede değişmeyen önsellerdir. Fisher'in bilgi matrisine dayanır. Jefreys'in önseli,

$$\pi(\theta) \propto \left|I(\theta)\right|^{1/2} \tag{3.4}$$

ile tanımlanır. Burada $\left|\ \right|$, determinantı gösterir ve $I(\theta)$, $P(y/\theta)$ olabilirlik fonksiyonuna dayalı Fisher bilgi matrisidir. Bu matris,

$$I(\theta) = -E\left[\frac{\partial^2 \log P(y/\theta)}{\partial \theta^2}\right] \tag{3.5}$$

eşitliği ile tanımlanır. Jeffreys'in önseli yerel olarak düzgündür (Locally uniform) yani bilgilendirici olmayan önseldir. Herhangi bir parametrik model $P(y/\theta)$ için bilgilendirici olmayan önsel bulmada otomatik bir düzenek sağlar. Jeffreys'in önselinin başka bir çekici özelliği, birebir dönüşümlerde değişmez olmasıdır. Değişmezlik

özelliğinin anlamı, θ için düzgün önsele sahipseniz ve $\phi(\theta)$, θ'nın birebir fonksiyonu ise

$$P(\phi(\theta)) = \pi(\theta)\left|\phi'(\theta)\right|^{-1} \tag{3.6}$$

Bu eşitlik $\phi(\theta)$'nın düzgün önselidir. Jefreys'in önseli bilgilendirici olmayan önsel için genel bir tarif sağlamasına rağmen bazı eksikliklere sahiptir. Önseller bazı modellerde uygun olmayandır ve uygun olmayan sonsal dağılımlara neden olabilir (Jeffreys, 1961).

3.1.2. Bayesci Çıkarsama

Parametre hakkındaki Bayesci çıkarsama parametrenin sonsal dağılımına dayanmaktadır. Sonsal dağılım, olabilirlik fonksiyonu ile önsel dağılımın çarpımından elde edilir. Bayesci sağkalım analizinde sonsal dağılım tahmin edilirken olabilirlik fonksiyonu olarak kısmi olabilirlik fonksiyonu kullanılarak ve parametrenin sonsal dağılımına dayanan çıkarsamada bulunulur.

Bayesci Cox modelde sonsal dağılım eşitlik (3.7)'de verildiği gibidir.

$$P(\beta / y) \propto L_p(y / \beta)P(\beta) \tag{3.7}$$

Burada $L_p(y / \beta)$, regresyon katsayısı β parametreli kısmi olabilirlik fonksiyonudur.

Klasik yöntemler parametrenin tahminini, momentler yöntemiyle veya en çok olabilirlik tahmincileriyle bulurlar. Buna karşın Bayesci yaklaşımlar genellikle parametre tahmininde sonsal ortalamayı kullanırlar. Sonsal ortalama,

$$E(\theta / y) = \int \theta P(\theta / y)d\theta \tag{3.8}$$

eşitliği ile verilir.

Eğer ilgilenilen sonsal dağılım biliniyor ise sonsal nokta tahmini bulunabilir. Modelin analitik olarak çözülmesi çok zor olduğu zaman MCMC gibi simülasyon algoritmaları kullanılmalıdır. MCMC yöntemi sonsal tahminleri elde etmek için kullanılan iteratif bir yöntemdir (Gilks ve ark, 1996).

Bayesci yaklaşımda aralık tahminleri güvenilir (credible) kümeler olarak adlandırılır ve Bayesci güvenilir kümeler en yüksek sonsal yoğunluk (Highest posterior density-HPD) aralığı ile tanımlanır. Eğer

$$\int_{L}^{U} P(\theta / y)d\theta = 1-\alpha \tag{3.9}$$

ise θ için güvenilir küme (L,U) olur. Örneğin $\int_{L}^{U} P(\theta / y)d\theta = 0,95$ olan bir (L,U) aralığının bulunması ile θ için %95 güvenilir küme oluşturulmuş olur. Bayesci güvenilir kümenin seçiminde aralık uzunluğunun minimum olması istenir. $\%100(1-\alpha)$ HPD aralığı, sadece en büyük sonsal olasılık yoğunluk fonksiyonlu bu noktaları içerir (Dalpatadu ve ark., 2002).

3.1.2.1. Markov Zinciri Monte Carlo (MCMC)Yöntemi

MCMC yöntemi, ilgilenilen sonsal değeri hesaplayan sonsal dağılımdan örnekleme yapmak için kullanılan genel bir simülasyon yöntemidir. Bu yöntem hedef bir dağılımdan başarılı bir şekilde örnekleme yapar. Her bir örneklem bir öncekine bağlıdır ve böylece Markov zinciri oluşur. Monte Carlo, Markov zinciri örneklerinin kullanılması ile yaklaşık bir beklenen değer oluşturur. Monte Carlo'nun basit bir gösterimi eşitlik (3.10)'da verilmiştir (Walsh, 2002).

$$\int g(\theta)P(\theta)d\theta \cong \frac{1}{n}\sum_{i=1}^{n} g(\theta^{t}) \tag{3.10}$$

Burada g(.) ilgilenilen fonksiyondur. θ^{t} ise $P(\theta)$'dan elde edilen örneklemlerdir. (3.10) eşitliği ile verilen ifade $g(\theta)$'nın yaklaşık beklenen değeridir.

Markov Zinciri Monte Carlo yöntemi, modern Bayesci hesaplamalarda oldukça başarılıdır. Sadece en basit Bayesci modellerde sonsal dağılımın analitik formları bulunabilir ve çıkarımlar doğrudan yapılabilir. Karmaşık modellerde sonsal dağılım ile doğrudan çalışmak oldukça zordur. MCMC yöntemi ile keyfi bir sonsal yoğunluktan P(θ /y) örneklemler oluşturulur ve ilgilenilen değerin yaklaşık beklenen değerini bulmak için bu örneklemler kullanılır. Eğer simülasyon algoritması doğru bir şekilde uygulanırsa Markov zinciri oldukça geniş koşullar altında hedef dağılıma P(θ /y) yakınsamayı garanti eder. Başka bir deyişle bir Markov zinciri, simülasyonda her bir adımda gerçek dağılım için yaklaşımı iyileştirir (Gilks ve ark., 1996; Walsh, 2002).

MCMC yöntemini içeren örnekleme yöntemleri Gibbs örnekleme, Metropolis-Hasting örnekleme ve diğer hibrit algoritmalardan oluşmaktadır. Metropolis-Hasting algoritması değişkenlerin ortak dağılımından örneklemlerin bir dizisini oluşturur. Gibbs örnekleme ise normalleştirme sabitinin bilinmediği durumlarda P(θ /y) dağılımından örneklemler oluşturan güçlü bir algoritmadır. Genellikle çalışmalarda MCMC yöntemini kapsayan Gibbs örnekleme kullanılır(Congdon, 2006; Congdon, 2003; SAS Istitute, 2006).

3.1.2.2. Gibbs Örnekleme

Gibbs örnekleme modeldeki her bir parametre için tüm koşullu dağılımlarda ortak sonsal dağılımı ayrıştırmayı gerektirir ve daha sonra onlardan örnekler üretir. Her bir parametre birbirine yüksek derecede bağımlı değilse örneklemler etkilidir. Uygulaması kolay ve basit bir mantığa dayanan Gibbs örneklemenin oldukça yaygın kullanımı vardır. Gibbs örneklemesi, Geman ve Geman (1984) tarafından yapılan çalışmada tanımlanmıştır. Gelfand ve ark. (1990)'da Gibbs örneklemeyi ilk olarak Bayesci çıkarımda problem çözümü için kullanmışlardır.

Parametre vektörünün $\theta = (\theta_1, \theta_2, ..., \theta_k)'$ olabilirlik fonksiyonunun P(y/ θ), ve önsel dağılımının $\pi(\theta)$ olduğunu kabul edelim. $\pi(\theta_i / \theta_j, i \neq j, y)$ 'nin tam (full) sonsal koşullu dağılımları ortak sonsal yoğunluk ile orantılıdır ve eşitlik (3.11)'de verilmiştir. Markov zincirinin sonsal dağılıma yakınsaması için tam sonsal koşullu dağılımdan üretilen örneklemeler kullanılır (Gamerman,1997; Gelfand ve Smith, 1990).

$$\pi(\theta_i / \theta_j, i \neq j, y) \propto P(y/\theta)\pi(\theta) \qquad (3.11)$$

Gibbs örnekleme aşağıdaki adımları takip eder.

1. t=0 için keyfi bir başlangıç değeri seçilir. $\theta^{(0)} = \left\{\theta_1^{(0)}, \theta_2^{(0)}, ..., \theta_k^{(0)}\right\}$
2. θ'nın her bir bileşeni $\theta^{(t+1)} = \left\{\theta_1^{(t+1)}, \theta_2^{(t+1)}, ..., \theta_k^{(t+1)}\right\}$ olarak meydana getirmek için aşağıdaki yol izlenir.
    - $\pi(\theta_1 / \theta_2^{(t)} ... \theta_k^{(t)}, y)$'den $\theta_1^{(t+1)}$ çekilir.
    - $\pi(\theta_2 / \theta_1^{(t+1)}, \theta_3^{(t)}, ..., \theta_k^{(t)}, y)$'den $\theta_2^{(t+1)}$ çekilir.
    - ...
    - $\pi(\theta_k / \theta_1^{(t+1)}, ..., \theta_{k-1}^{(t+1)}, y)$'den $\theta_k^{(t+1)}$ çekilir.
3. t=t+1 alınır ve t<T ise 2. adıma gidilir. Burada T, iterasyon sayısı t ise her bir iterasyonu gösterir.

Gibbs örnekleme $\theta^{(t)}$'den $\theta^{(t+1)}$'e geçiş adımlarını sağlar ve parametreden parametreye güncellemeleri yapar. Yakınsama sağlandıktan sonra tüm simülasyon değerleri $\theta^{(t)}$'ler hedef sonsal dağılımdan elde edilmiş olur (Ibrahim ve ark.,2001).

Gibbs örneklemede dikkat edilmesi gereken bazı önemli noktalar vardır. Bunlar yakma (burn-in) peryodu, iterasyon sayısı ve zincirin başlangıç değerinin belirlenmesidir.

i) Yakma (Burn-in) Periyodu

Sonsal dağılımdan çıkarsama yapılırken başlangıç değerinin etkisini minimum yapmak için Markov zinciri örnekleminin başlangıç kısmının atılmasıdır. Teoride Markov zinciri sonsuz sayıda çalışırsa başlangıç değerinin etkisi sıfıra yaklaşır. Uygulamada sonsuz sayıda iterasyon mümkün olmadığından t iterasyondan sonra zincirin hedef dağılıma ulaştığı varsayılır. Ayrıca zincirin başlangıç kısmı çıkarılabilir ve sonsal çıkarım için iyi örneklemlerin kullanılabildiği varsayılır. Bu durumu sağlayan t değeri, yakma sayısıdır (SAS, 2011).

ii) İterasyon Sayısı

Pek çok uygulamada doğru sonsal çıkarımlar için birkaç bin iterasyon gereklidir.

iii) Başlangıç Değeri

Uygulamalarda genellikle Markov zinciri başlangıç değeri olarak en çok olabilirlik tahmin değeri kullanılır.

3.1.3 Bayesci İstatistiklerin Klasik İstatistiklerden Farkı

Bayesci ve klasik yaklaşımdaki en önemli fark parametrenin tanımıdır. Klasik yaklaşımda parametre kitlenin sabit bir özelliği olarak tanımlanır. Klasik analizlerde amaç parametrenin gerçek değerini tahmin etmek ve tahminler etrafındaki güven aralığını oluşturmaktır. Standart hata bu sürecin ayrılmaz bir parçasıdır ve tekrarlanan örneklere ait tahminlerin değişkenliğini temsil eder. Farkın daha iyi anlaşılması için %95 güven aralığı yorumu incelenecek olursa klasik yapıda %95 olasılıkla parametre A ile B arasına düşer yorumu yanlıştır. Çünkü herhangi bir örneklemden hesaplanan güven aralığı parametreyi ya içerir ya da içermez. Bu yorum yerine güven aralığı tekrarlanan örnekler üzerinde aralığın beklenen performansı olarak tanımlanır. Örneğin kitleden 100 örnek çekilerek ve her bir örnekten parametre tahmini etrafında %95 güven aralığı oluşturulsun. Bu durumda aralıkların 95 tanesinin kitle parametresini içermesi beklenir. Klasik yapıda olasılık ifadesi veriler için geçerlidir parametre için değil.

Bayesci yapıda ise parametre bir dağılıma sahip rasgele bir değişken olarak görülür. Bayesci analizlerin amaçlarından biri, bu dağılımın şeklinin belirlenmesidir. Örneğin ortalama ve standart sapma sırasıyla dağılımın merkezi ve yayılımı hakkında bilgi verir. Bayesci belirsizlik kavramı, tekrarlanan örnekleri içermez. Klasik yapının tersine parametre rasgele bir değişken olarak görülür. Örneğin Bayesci güven aralığında %95 olasılıkla ilgilenilen parametre A ve B arasına düşer şeklinde yorum yapılmasına izin verir. Bu yorum klasik güven aralığı yorumundan çok farklıdır. Çünkü buradaki olasılık ifadesi parametre içindir veri için değildir (Demirhan,2004).

3.1.4. Bayesci Analizlerin Avantajları ve Dezavantajları

Sağkalım analizinde Bayesci yaklaşımın avantajının ne olduğu ile ilgili çeşitli tespitler vardır. Birincisi sağkalım modellerinin oluşturulması, özellikle karmaşık sansürlü verinin varlığında genellikle zor olduğu iyi bilinir. Gibbs örneklemesi ve MCMC tekniklerinin kullanılmasıyla karmaşık sağkalım modelleri oldukça açık bir şekilde oluşturulabilir. İstatistiksel yazılımlarla uygulaması büyük ölçüde kolaylaşır.

Bayesci yaklaşım önsel bilginin katkısını sağlar. Oysa bu durum klasik yöntemlerde yoktur. Ayrıca klasik sonuç çıkarma, birçok modelde düzgün önsel gibi bilgilendirici olmayan önselin pek çok tipi ile Bayesci sonuç çıkarmanın özel bir durumu olarak elde edilebilir. Örneğin düzgün önselli sonsal model en çok olabilirlik tahmininin karşılığı olur. Bu durumda klasik yöntemlerle sonuç çıkarma Bayesci sonuç çıkarmanın özel bir durumu olarak görülebilir.

Bayesci yaklaşımın dezavantajları ise önsel dağılımın belirlenmesidir. Önsel bilginin seçilmesinde doğru olan kesin bir kural yoktur. Eğer önsel seçiminde dikkatli olunmazsa yanlış sonuçlar elde edilir (SAS, 2011).

Parametre hakkındaki kesin olmayan önsel bilgilerin önsel dağılıma dönüştürülmesinde bazı sıkıntılar ortaya çıkar. Özellikle parametreler arasında önsel ilişkiler varsa bilgi içeren önsel dağılımın belirlenmesi zor olur. Bu durumda da sonsal dağılımın elde edilmesi zorlaşır.

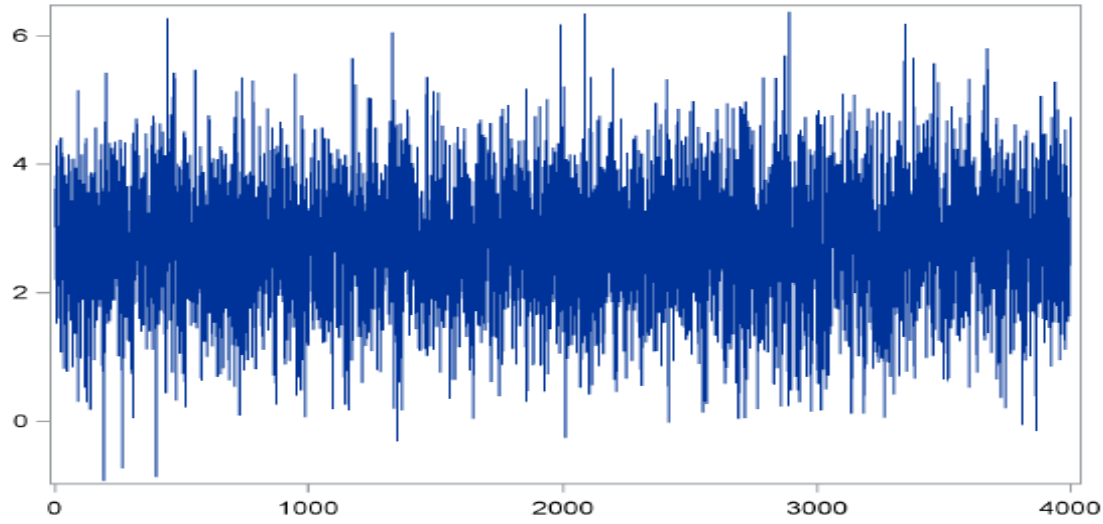
3.1.5. Markov Zincirinin Yakınsamasının Değerlendirilmesi

Markov zinciri Monte Carlo yöntemlerinde zincirin yakınsamasının değerlendirilmesi zincirin durağan hale gelmesi veya sonsal dağılıma ulaşıp ulaşmadığının karar verilmesidir. Yakınsama belirleyicileri bu sorunu çözmede yardımcı olur. Yakınsama değerlendirilirken sadece ilgilenilen parametrenin değil bütün parametrelerin yakınsamasının değerlendirilmesi gerekir.

Bayesci çıkarsamada iz (trace) grafikleri ile görsel olarak yakınsama değerlendirilmesinin yapılmasını sağlar. Ayrıca yakınsamanın değerlendirilmesinde otokorelasyon ve Geweke testi de kullanılabilir.

3.1.5.1. İz Grafikleri İle Görsel Analiz

İz grafikleri, yakınsamanın değerlendirilmesinde çok kullanışlıdır. İz grafikleri hedef dağılıma yakınsama olup olmadığını gösterir. Zincirde ilerleme devam ederken noktaların dağılımı değişmiyorsa zincir hedef dağılıma ulaşmış sayılır. Grafikte zincirin merkezi, bir değer üzerinde görülüyorsa ve merkez etrafında dalgalanmalar az ise bu durum zincirin hedef dağılıma ulaştığını gösterir. Yakınsama sağlamış bir zincirin iz grafiği ile gösterimi şekil 3.1'de verilmiştir.



Şekil 3.1. Yakınsama sağlamış bir zincirin iz grafiği

3.1.5.2. Geweke Testi

Geweke tarafından 1992'de bulunan bu test, yakınsamanın başarısızlığını tespit etmek için Markov zincirinin ilk kısmındaki değerlerle ikinci kısımdaki değerleri karşılaştırır. $\theta^{(t)}$ Markov zinciri, $(\theta_1^{(t)} : t = 1,2,...,n_1)$ ve $(\theta_2^{(t)} : t = n_a,...,n)$ olarak tanımlanan iki alt diziye ayrılır. Burada $1 < n_1 < n_a < n$ ve $n_2 = n - n_a + 1$ olarak tanımlanırsa sonsal ortalamalar aşağıdaki gibi verilir (Geweke, 1992).

$$\bar{\theta}_1 = \frac{1}{n_1}\sum_{t=1}^{n_1}\theta^t \quad \text{ve} \quad \bar{\theta}_2 = \frac{1}{n_2}\sum_{t=n_a}^{n}\theta^t \tag{3.12}$$

Eğer n_1/n ve n_2/n oranları sabitse $(n_1+n_2)/n<1$ olur ve zincir durağan hale gelir ve test istatistiği $n \to \infty$ iken standart normal dağılıma yakınsar.

$$z_n = \frac{\bar{\theta}_1 - \bar{\theta}_2}{\sqrt{\frac{\hat{s}_1(0)}{n_1} + \frac{\hat{s}_2(0)}{n_2}}} \tag{3.13}$$

Test istatistiği sonucunda p>0,05 ise Markov zincirinin hedef sonsal dağılıma yakınsadığı söylenir. Burada $\hat{s}_i(0)$, i=1,2 sıfır frekansta spectral yoğunluk tahminidir. $\frac{\hat{s}_1(0)}{n_1}$, $\bar{\theta}_1$'in ve $\frac{\hat{s}_2(0)}{n_2}$, $\bar{\theta}_2$'nin asimtotik varyansıdır. Bu asimtotik varyansın karekökü $\bar{\theta}_1$ ve $\bar{\theta}_2$ için standart hatayı verir (Sahlin, 2011).

3.1.5.3. Otokorelasyon

Gecikme zamanı k adet olan bir zincirde otokorelasyon, zincirde k iterasyon ile ayrılmış parametre değerlerinin kümesi arasındaki korelasyondur. Gecikme zamanı (lag) k için tanımlanan örneklem otokorelasyonu, örneklem otokovaryans fonksiyonuna dayanarak tanımlanır.

$$\hat{\rho}_k = \frac{\hat{\gamma}(k)}{\hat{\gamma}(0)}, \quad |h| < n \tag{3.14}$$

$\theta_j^{(t)}$'nin k.ncı gecikme zamanının örnek otokovaryans fonksiyonu eşitlik 3.15'da verildiği gibidir.

$$\hat{\gamma}(k) = \frac{1}{n-k}\sum_{t=1}^{n-k}(\theta_i^{(t+k)} - \bar{\theta}_i)(\theta_i^{t} - \bar{\theta}_i), \quad 0 \le k < n \tag{3.15}$$

Uzun gecikme zamanları arasındaki yüksek korelasyonların varlığı Markov zincirinin yavaş yakınsadığını gösterirler. Zincirin yavaş yakınsaması parametrelerin yüksek derecede ilişkili olmasından kaynaklanabilir.

3.1.6. Model Uyum İstatistikleri

İstatistiksel modellerin uygunluğunun değerlendirilmesi için en çok kullanılan kriterler Schwarz bilgi kriteri olarak da bilinen Bayesci bilgi kriteri, Akaike bilgi kriteri ve sapma bilgi kriterleridir. Daha küçük kritere sahip model daha uygun model olarak belirlenir.

3.1.6.1. Bayesci Bilgi Kriteri (BIC)

Model seçim kriterlerinden biri olan Schwarz kriteri olarak da bilinen Bayesci bilgi kriteri, istatistikte yaygın olarak kullanılmaktadır. Akaike bilgi kriterine benzer şekilde olabilirlik fonksiyonuna dayalı bir kriterdir. Bayesci bilgi kriteri,

$$\mathrm{BIC} = -2\log(\mathrm{P}(\mathrm{y}/\hat{\theta})) + \mathrm{p}\log(\mathrm{n}) \tag{3.16}$$

eşitliği ile bulunur. Burada $\hat{\theta}$, en çok olabilirlik tahmin edicisi ve p parametre sayısı ve n örnek genişliğidir.

3.1.6.2 Akaike Bilgi Kriteri (AIC)

Akaike bilgi kriteri(AIC) istatistiksel modelin uyum iyiliği ölçüsüdür. Hirotsugu Akaike tarafından 1973 yılında bulunmuştur. AIC,

$$\begin{aligned}\mathrm{AIC} &= \mathrm{D}(\hat{\theta}) + 2\mathrm{p} \\ &= -2\log(\mathrm{P}(\mathrm{y}/\hat{\theta}) + 2\mathrm{p}\end{aligned} \tag{3.17}$$

eşitliği ile verilir. Burada $\hat{\theta}$, en çok olabilirlik tahmin edicisi ve p parametre sayısıdır.

3.1.6.3. Sapma bilgi kriteri (DIC)

Sapma bilgi kriteri (DIC) (Spiegelhalter ve ark. 2002), bir model değerlendirme aracıdır. Akaike bilgi kriteri(AIC) ve Bayesci bilgi kriterine(BIC), Bayesci bir

alternatiftir. DIC, sonsal yoğunlukları kullanır. Yani önsel bilgiyi dikkate alır. Kriter, iç içe olmayan modellere ve özdeş ve bağımsız olmayan veriye sahip modellere uygulanabilir. MCMC'de DIC hesaplanması açıktır. DIC; AIC ve BIC gibi parametre uzayı üzerinde maksimizasyonu gerektirmez. Daha küçük DIC, veri setinin daha iyi uyumuna işaret eder. Modelin parametreleri θ ile gösterildiğinde, sapma bilgi formülü

$$\begin{aligned} DIC &= \overline{D} + pD \\ &= D(\overline{\theta}) + 2pD \end{aligned} \qquad (3.18)$$

biçimindedir. Burada,

$D(\theta) = 2\left(\log(f(y)) - \log(p(y \mid \theta))\right)$: sapma

$p(y \mid \theta)$: normalleştirilmiş sabitli olabilirlik fonksiyonu

$f(y)$: sadece verinin fonksiyonu olan standartlaştırma terimidir. Bu terim parametreye göre sabittir ve aynı olabilirlik fonksiyonuna sahip olan farklı modeller karşılaştırıldığı zaman ilgisizdir. DIC karşılaştırmalarında terim iptal edildiği için hesaplanması genellikle ihmal edilir. Sonsal ortalama,

$$\overline{\theta}: \quad \frac{1}{n}\sum_{t=1}^{n}\theta^t$$

eşitliği ile verilir. Sapmanın sonsal ortalaması,

$$\begin{aligned} \overline{D} &= E_{\theta/y}(D(\theta)) \\ &= E_{\theta/y}(-2\log P(y/\theta)) \\ &= \frac{1}{n}\sum_{t=1}^{n} D(\theta^t) \end{aligned} \qquad (3.19)$$

eşitliği ile verilir.

(3.19) eşitliği ile verilen sapmanın sonsal ortalaması, Gibbs örneklemesinin bir iterasyonunun sonunda hesaplanmış log olabilirliklerinin ortalamasıdır. Beklenen sapma, modelin veriye uyumunun ne kadar iyi olduğunu ölçer. $D(\bar{\theta})$, $\bar{\theta}$ 'da değerlendirilmiş sapmadır ve

$$D(\bar{\theta}) = D(E_{\theta/y}[\theta]) \\ = -2\log(P(y \mid \bar{\theta})) \tag{3.20}$$

şeklinde ifade edilir. En iyi sonsal tahminde değerlendirilen sapmadır.

pD : Parametrelerin etki derecesidir. Uyum ölçüsü ve tahminlerdeki sapma arasındaki fark olup,

$$p_D = \overline{D} - D(\bar{\theta}). \tag{3.21}$$

ile ifade edilir.

3.1.6.4. DIC ve AIC Arasındaki İlişki

Spiegelhalter ve ark. (1998) DIC değerini iyi bilinen AIC değerinin doğal bir genelleştirmesi olarak tanımlamışlardır. AIC değeri,

$$AIC = D(\hat{\theta}) + 2p \tag{3.22}$$

eşitliği ile bulunur. Burada $\hat{\theta}$ parametre vektörünün en çok olabilirlik tahmini ve p parametre sayısıdır. Ayrıca Spiegelhalter ve ark. (1998) hiyerarşik olmayan modeller için

$$p \approx p_D,\ \hat{\theta} \approx \bar{\theta} \text{ ve } DIC \approx AIC \tag{3.23}$$

olduğuna çalışmalarında yer vermişlerdir. Böylece

$$AIC \approx \overline{D} + P \tag{3.24}$$

eşitliğinin yazılabildiğini göstermişlerdir.

Ayrıca Ellison (2004) çalışmasında DIC eşitliğinin AIC'in hesaplamasına çok benzediğini ve bilgilendirici önselin olmaması durumunda bu iki kriterin eşit olmasının beklendiğini ifade etmiştir.

Spiegelhalter ve ark. (2002)'de DIC değeri AIC'in Bayesci bir eşdeğeri olarak tanımlamışlardır.

3.2.Kayıp Veri

Araştırmaların çoğunda verilerde gözlenemeyen (kayıp) değerlerle sıklıkla karşılaşılmaktadır. Kayıp verinin problem olarak görülmesinin nedeni istatistiksel analizlerin ve bilgisayar programlarının verilerin tümünün var olduğu durumlar için geliştirilmesinden kaynaklanmaktadır. Hemen hemen bütün istatistiksel yazılımlarda kayıp değerli denek yok sayılarak, varsayılan analiz uygulanmaktadır. Araştırmacıların büyük paralar ve zaman harcayarak elde ettikleri veriler birkaç gözlemin eksikliği nedeniyle heba olmaktadır. Verilerdeki bu silme işlemiyle örnek genişliği küçülmekte ve bu durum istatistiksel gücün azalmasına neden olmaktadır. Tüm bu nedenlerden dolayı kayıp veri probleminin giderilmesi önem kazanmaktadır.

3.2.1. Kayıp Veri Oluşum Mekanizması

Rubin (1976), Little ve Rubin (2002) kayıp veri probleminin çözümü için öncelikle kayıp veri oluşumunu sınıflandırmışlardır. Bu sınıflandırmada kayıp değerin veri ile nasıl bir ilişki içinde olduğu önem kazanır. Genel olarak kaynaklarda üç tip kayıp veri oluşumundan söz edilmektedir. Bunlar tesadüfi kayıplar (Missing at random-MAR), tam tesadüfi kayıplar (Missing completely at random-MCAR) ve tesadüfi olmayan kayıplar (Missing not at random-MNAR) şeklinde sınıflanmıştır.

3.2.1.1. Tesadüfi Kayıp (MAR)

Y değişkenindeki kayıp verinin olasılığı analiz edilen modeldeki diğer değişken veya değişkenlerin ölçümlerine bağlı, fakat Y değişkeninin kendisine bağlı değilse, verideki kayıp veri oluşum mekanizması tesadüfi olarak tanımlanır. Tesadüfi olarak kayıp terimi verilerin gelişigüzel kayıp anlamına geldiği düşünüldüğü için biraz yanıltıcı olabilir. Fakat tesadüfi olarak kayıp teriminin gerçek anlamı, kayıp verinin olasılığı ve ölçülen bir veya daha fazla değişkenin arasında sistematik bir ilişkinin varlığını işaret eder.

Örneğin bir eğitim araştırmacısının okuma başarısı ile ilgili yaptığı araştırmada, okuma başarısı değişkeninde İspanyol öğrencilerin, Kafkas öğrencilerine göre daha fazla kayıp değere sahip olduğu sonucunu buluyor. İkinci olarak kanser hastalarının yaşam kalitesi ile ilgili araştırma anketindeki sorulara yaşlı ve eğitim düzeyi düşük olan hastaların cevaplamayı reddetmesi örneği verilebilir. Bu örneklerdeki kayıp veri oluşum mekanizması tesadüfi kayıp (MAR) olarak kabul edilir. Çünkü kayıp değerler değişkenin kendisinden değil, diğer değişkenlerin etkisinden kaynaklanmaktadır. Örnekte verilen yaşam kalitesi değişkenindeki yanıtsızlık, yaş ve eğitim durumu ile ilgili iken yaşam kalitesi değişkeni ile ilgili değildir (Enders, 2010).

3.2.1.2. Tam Tesadüfi Kayıp (MCAR)

Y değişkenindeki kayıp veri olasılığı ne ölçülen diğer değişkenlere nede Y'nin kendisine bağlı değildir (Enders, 2010). Tam tesadüfi kayıp oluşum mekanizması, tesadüfi kayıp oluşum mekanizmasına göre daha kısıtlayıcı koşulları vardır. Çünkü burada kayıplık durumunun veriden tamamen bağımsız oluştuğu varsayılır. Daha önce verilen eğitim örneğinde, kişinin kişisel olaylar (hastalık, cenaze, ailesel problemler veya okul değişikliği), zamanlama ile ilgili zorluklar (araştırmacılar okulu ziyaret ettiklerinde okulun gezide olması) gibi nedenlere bağlı olarak okuma başarısı değişkeni kayıp değerlere sahip olabilir. Bu örneklerdeki kayıp veri oluşum mekanizması tam tesadüfi kayıp meydana getirir. Çünkü okuma başarısındaki kayıp değer olasılığı ne kendisine nede ölçülen diğer değişkenlere bağlıdır (Enders, 2010).

3.2.1.3. Tesadüfi Olmayan Kayıp (MNAR)

Y değişkenindeki kayıp veri olasılığı Y'nin kendisi ile ilgili ancak diğer değişkenlerle ilgili değilse, kayıp veri oluşum mekanizması tesadüfi olmayan kayıptır. Daha önce verilen eğitim örneğinde okuma başarısı düşük olan öğrenciler soruyu anlamakta zorlanıp soruyu yanıtsız bırakmışlarsa bu kayıp tesadüfi olmayan kayıptır. Benzer şekilde kanser hastalarının çok hasta olması o andaki yaşam kalitesini düşük olması nedeniyle bu soruyu yanıtsız bırakması da bu tip kayıba örnek olabilir. Bu örneklerdeki kayıp veri oluşum mekanizması tesadüfi olmayan kayıptır. Çünkü kayıp değer olasılığı kayıp değerli değişkenin kendisine bağlıdır (Enders, 2010).

Kayıp veri mekanizmasının tam olarak anlaşılması için Çizelge 3.1'de verilen örnek incelenebilir. Bu örnekte, bir iş için başvuruda bulunan 20 kişi IQ testine tabi tutulmuş ve daha sonra 6 aylık bir deneme süresi sonunda işverenler, potansiyel çalışanların iş performanslarını değerlendirerek veri kümesini oluşturmuş olsunlar.

Çizelge 3.1'de verilen iş performansı değerlendirme tablosunun MAR sütunu incelendiğinde IQ skoru küçük olan başvurular işe alınmamıştır. Böylece iş performans değerlerindeki kayıplık olasılığı kişilerin performansları ile ilgili değil IQ puanları ile ilgilidir (Enders, 2010) .

Çizelge 3.1. Farklı kayıp veri oluşum mekanizmalı iş performans örnek verisi

		İş Performans Değerlendirme			
IQ	Cinsiyet	Tam veri	MAR	MCAR	MNAR
78	B	9	-	-	9
84	E	13	-	13	13
84	B	10	-	-	10
85	E	8	-	8	-
87	B	7	-	7	-
91	E	7	7	7	-
92	B	9	9	9	9
94	B	9	9	9	9
94	E	11	11	11	11
96	B	7	7	-	-
99	E	7	7	7	-
105	E	10	10	10	10
105	E	11	11	11	11
106	E	15	15	15	15
108	B	10	10	10	10
112	B	10	10	-	10
113	B	12	12	12	12
115	E	14	14	14	14
118	E	16	16	16	16
134	B	12	12	12	12

Kayıp veri mekanizmasının MCAR olması durumunda kayıp olma olasılığı veriden tamamen bağımsızdır. Örnekte MCAR sütunu incelendiğinde kayıp değerler ne IQ ile nede iş performansı ile ilgilidir. Kayıp olma durumu başka nedenlerden kaynaklanmaktadır. Örneğin başvuruda bulunan bayan elemansa 6 aylık deneme sürecinde doğum iznine ayrılmış olabilir veya eşinin başka bir şehirde iş bulması nedeniyle ayrılmak zorunda kalmış olabilir veya o kişiyi değerlendirecek olan sorumlu danışman şirketin başka bir bölümüne atanmış olması gibi.

MNAR kayıp veri durumunda ise başvuran 20 kişinin 6 aylık deneme süresi sonundaki iş performansları incelenir ve kötü performanslı (performans puanı 8 ve altında olanlar) kişilerin işlerine son verilir. Sonuç olarak iş performansındaki kayıp olasılığı kişinin iş performans değerlerine bağlıdır, IQ değerine bağlı değildir.

3.2.1.4. Kayıp Veri Oluşum Mekanizmasının Matematiksel Gösterimi

$Y = \{Y_{göz}, Y_{kay}\}$, bazı elemanlarının kayıp olduğu değişken matrisini göstersin. Burada $Y_{göz}$, Y'nin gözlenen kısmını, Y_{kay} ise kayıp veri kısmını gösterir. R, Y ile aynı boyutta olan 0 ve 1'lerden oluşan kayıplık matrisidir. R'nin herbir elemanı Y'nin elemanlarının kayıp olup (0) olmadığını (1) gösterir.

$$R = \begin{cases} R_{ij} = 0 \rightarrow Y_{ij} \text{ kayıp} \\ R_{ij} = 1 \rightarrow Y_{ij} \text{ gözlenen} \end{cases}$$

MNAR kayıp veri mekanizması olduğunda kayıp verinin dağılımı

$$P(R / Y_{göz}, Y_{kay}, \phi) \tag{3.25}$$

şeklinde olur. Burada ϕ bilinmeyen parametredir (Enders 2010). Söz konusu kayıp veri mekanizması MAR olduğunda Y değişkenindeki kayıp veri olasılığı analiz edilen modeldeki ölçülen başka bir değişkenle ilgilidir. Fakat Y'nin kendi değerleri ile ilgili değildir. Yani R $Y_{göz}$'ye bağlı fakat Y_{kay}'ye bağlı değildir. Bu durumda kayıp verinin dağılımı,

$$P(R / Y_{göz}, \phi) \tag{3.26}$$

olur. Kayıp veri oluşum mekanizması MCAR olduğunda R, hem $Y_{göz}$ ile hemde Y_{kay} ile ilişkili değildir. Bu durumda kayıp verinin dağılımı

$$P(R / \phi) \tag{3.27}$$

ile verilir.

3.2.2. Kayıp Veri Problemi ve Analizi

Kayıp veri probleminin giderilmesi kayıp veri analizi olarak adlandırılır. Kayıp veri analizi, kayıp değerli denekleri silme (Case Deletion) ve kayıp veri için değer atama (Imputation Methods) yöntemleri ile iki şekilde yapılmaktadır (Schafer, 1997).

Kayıp veri analizinde en çok kullanılan yöntem kayıp değerli deneklerin silindiği yöntemlerdir. Bunlar Liste Bazında Silme (Listwise Deletion) ve Çiftler Bazında Silme (Pairwise Deletion) yöntemleridir.

Kayıp değerli veride denek veya değişkenlerin bilinen değerleri kullanılarak, kayıp değerlerin tahmin edilmesi değer atama olarak adlandırılır. Değer atama işleminden sonra kayıp değerler tamamlanarak yeni kayıp değer içermeyen veri seti oluşturulmuş olur. Oluşturulan bu yeni tamamlanmış veriye istenilen klasik istatistiksel analizler uygulanabilir. Verideki kayıp değerleri tahmin eden pek çok yöntem geliştirilmiştir. Bu yöntemlerden en çok kullanılanları Ortalama Değer Atama (Mean Imputation), Regresyon Ataması (Regression Imputation), Beklenti Maksimizasyonu Algoritması (Expectation Maximization - EM Algorithm), Çoklu Değer Atama (Multiple Imputation) yöntemleridir.

3.2.2.1. Liste Bazında Silme (Listwise Deletion)

Liste bazında silme yöntemi, eksiksiz veri analizi (complete case analysis) olarak da bilinir. Bir yada daha fazla kayıp değere sahip herhangi bir deneğin veriden çıkartılması öngörülür. Liste bazında silme yönteminin en önemli faydası kolay bir şekilde uygulanmasıdır. Karmaşık kayıp veri analizlerine ve özel yazılımlara olan ihtiyacı ortadan kaldırır. Liste bazında silme yönteminde, kayıp veri oluşum mekanizmasının MCAR olması varsayımı vardır. Bu varsayım sağlanmadığı zaman yanlı parametre tahminleri üretir (Enders, 2010). Liste bazında silme yöntemi, kayıp değerli denekleri yani verinin büyük bir kısmını attığı için analize dahil eden verinin standart hatasının yükselmesine, daha geniş güven aralıklarına ve hipotez testinde güç kaybına neden olmaktadır (Allison, 2000).

Liste bazında silme yöntemi ile, regresyon analizi ve diğer çeşitlerinde (Lineer, Lojistik, Poisson, Cox) MCAR varsayımı ihlal edilse de yansız parametre tahminleri elde edilir (Little, 1992).

3.2.2.2. Çiftler Bazında Silme (Pairwise Deletion)

Doğrusal modeller için liste bazında silme yöntemine alternatif olarak sıklıkla kullanılan yöntem çiftler bazında silme yöntemidir. Bu yöntemde her bir değişken çifti için gözlemleri mevcut olan deneklerin tamamı kullanılır. Başka bir değişken çifti için analize dahil olan denekler değişebilir. Bu nedenle bu analize verisi mevcut denekler analizi (Available Case Analysis) denir (Allison, 2000). Örneğin iki değişken arasında korelasyon veya kovaryans hesabı yapılırken kayıp değere sahip denek silinerek tahmin yapılır. Her bir değişken çifti için silinen örnek değişebilir (Allison, 2000).

3.2.2.3. Ortalama Değer Atama (Mean Imputation)

Kayıp değerli veride kayıp değer yerine kayıp değerli değişkenin var olan deneklere ait gözlemlerinin ortalaması atanır. Kolay uygulanabilir bir yöntem olmasından dolayı sıklıkla kullanılan bir yöntemdir. Fakat bu yöntem, parametrelerin tahmininde yanlılığa sebep olur (Haitovsky, 1968).

3.2.2.4. Regresyonla Değer Atama (Regression Imputation)

Regresyonla değer atama yöntemi kayıp değer içermeyen değişkenler üzerine regresyon denklemi kurarak, kayıp değerlere atama yapan bir yöntemdir. Ortalama değer atama yöntemi gibi uzun yıllardır kullanılan bir yöntemdir. Regresyon değer atama yönteminin ilk adımında kayıp değer içermeyen değişkenlerden kayıp değerli değişkenleri tahmin eden regresyon eşitlikleri elde edilir. İkinci adımda kayıp değerli değişkenlerin değerlerinin tahminleri bulunur. Bu tahmini değerler, kayıp değerlerin yerine kullanılır. Böylece kayıp değerleri tamamlanmış bir veri seti oluşturulmuş olur.

3.2.2.5 Beklenti Maksimizasyonu- EM Algoritması (Expectation Maximization - EM Algorithm)

EM algoritması, kayıp değerli verilerde en çok olabilirlik tahminlerini kullanan genel bir yöntemdir. EM algoritması çok değişkenli normal dağılım varsayımı altında doğru tahminler veren iki adımlı iteratif bir yöntemdir. İlk adımı E adımı olarak adlandırılır ve beklenti adımıdır. Bu adımda gözlenen değerlerin kullanımıyla kayıp

değerli değişkenler üzerinden logaritmik olabilirliğin beklenen değeri hesaplanır. İkinci adım maksimizasyon adımı olup M adımı olarak adlandırılır. Bu adımda beklenen log-olabilirlik maksimize edilerek parametre tahminlerinin yeni değerleri elde edilir. Bu iki adım yakınsama sağlanıncaya kadar tekrar edilir.

EM algoritması kayıp değerli veri Y_{kay} ve θ parametresi arasındaki karşılıklı bağımlılıktan yararlanır. Diğer bir ifadeyle Y_{kay}, θ tahmini ile ilgili bilgi içerir ve θ, Y_{kay}'ın olası değerini bulmak için bize yardımcı olur. $Y_{göz}$ bilindiği durumda θ'nın tahmini için aşağıdaki yol izlenir.

- θ'nın başlangıç değerleri kullanılarak kayıp değerli Y_{kay} tamamlanır.
- θ, $Y_{göz}$ ve Y_{kay} tahmin değerleri kullanılarak tekrar tahmin edilir.

Bu iterasyon tahminler yakınsayıncaya kadar tekrarlanır.

Herhangi bir kayıp değerli veri probleminde bütün (complete) verinin ($Y = (Y_{göz}, Y_{kay})$) dağılımı,

$$P(Y/\theta) = P(Y_{göz}/\theta)P(Y_{kay}/Y_{göz},\theta) \tag{3.28}$$

ve bütün (complete) verinin log- olabilirliği,

$$L(\theta/Y) = \log(P(Y/\theta)$$
$$L(\theta/Y) = L(\theta/Y_{göz}) + \log(P(Y_{kay}/Y_{göz},\theta) + c \tag{3.29}$$

ile bulunur. Burada $L(\theta/Y) = \log(P(Y/\theta)$ bütün (complete) verinin log- olabilirliğini, $L(\theta/Y_{göz})$ gözlenen veri log olabilirliğini, c ise keyfi (arbitrary) bir sabiti göstermektedir. (3.29) eşitliğinde verilen $P(Y_{kay}/Y_{göz},\theta)$ terimi, θ ve mevcut değerler verildiğinde kayıp verinin tahmini dağılımıdır. Bu dağılım Y_{kay} ile θ arasındaki karşılıklı bağımlılığı ifade eder ve EM Algoritmasında temel bir rol oynar.

$\theta^{(t)}$ başlangıç tahmini bilindiğinde kayıp verinin tahmini olasılık dağılımı $P(Y_{kay}/Y_{göz},\theta^{(t)})$ üzerinden (3.29) eşitliğinin beklenen değeri elde edilir.

Bu beklenen değer eşitlik (3.30)'da verildiği gibidir.

$$E(\theta / \theta^{(t)}) = L(\theta / Y_{göz}) + H(\theta / \theta^{(t)}) + c \tag{3.30}$$

Burada

$$E(\theta / \theta^{(t)}) = \int L(\theta / Y) P(Y_{kay} / Y_{göz}, \theta^{(t)}) dY_{kay} \tag{3.31}$$

ve

$$H(\theta / \theta^{(t)}) = \int \log P(Y_{kay} / Y_{göz}, \theta) P(Y_{kay} / Y_{göz}, \theta^{(t)}) dY_{kay} \tag{3.32}$$

şeklinde ifade edilir. Dempster ve ark. (1977) çalışmasında kayıp değerli verinin parametre tahminlerinde $\theta^{(t+1)}$ tahminleri, $E(\theta / \theta^{(t)})$ fonksiyonunu maksimum yapıyorsa $\theta^{(t+1)}$ tahminleri $\theta^{(t)}$ parametre tahmini kadar ve ondan daha iyi tahmindir (Schafer, 1997).

Bu durum,

$$L(\theta^{(t+1)} / Y_{göz}) \geq L(\theta^{(t)} / Y_{göz}) \tag{3.33}$$

eşitliği ile ifade edilir. Bu ifadenin daha açık hali aşağıda verildiği gibidir.

$$\begin{aligned} &L(\theta^{(t+1)} / Y_{göz}) \geq L(\theta^{(t)} / Y_{göz}) \\ &= E(\theta^{(t+1)} / \theta^{(t)}) - E(\theta^{(t)} / \theta^{(t)}) + H(\theta^{(t)} / \theta^{(t)}) - H(\theta^{(t+1)} / \theta^{(t)}) \end{aligned} \tag{3.34}$$

Bu bilgiler ışığında EM algoritmasını iki adımlı iteratif bir yöntem olarak açıklamak daha doğru olur.

I. Beklenti (Expectation) Adımı (E Adımı)

$E(\theta / \theta^{(t)})$, $P(Y_{kay} / Y_{göz}, \theta^{(t)})$ üzerinden $L(\theta / Y)$'nin ortalaması ile hesaplanır.

II. Maksimizasyon (Maximization) Adımı (M Adımı)

$E(\theta / \theta^{(t)})$ fonksiyonunun maksimize edilmesiyle $\theta^{(t+1)}$ tahmini elde edilir.

EM Algoritmasında yakınsaklığı belirlemek için iterasyonlarda elde edilen parametre tahminlerinin değişimi incelenir.

Eğer $\left|\theta_j^{(t)} - \theta_j^{(t-1)}\right| \leq \varepsilon$ j=1,2,…,k ise ε'nun küçük bir değeri için (0,00001) için EM Algoritması t.nci iterasyonda yakınsamıştır denir.

3.2.2.6. Çoklu Değer Atama (Multiple Imputation-MI)

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

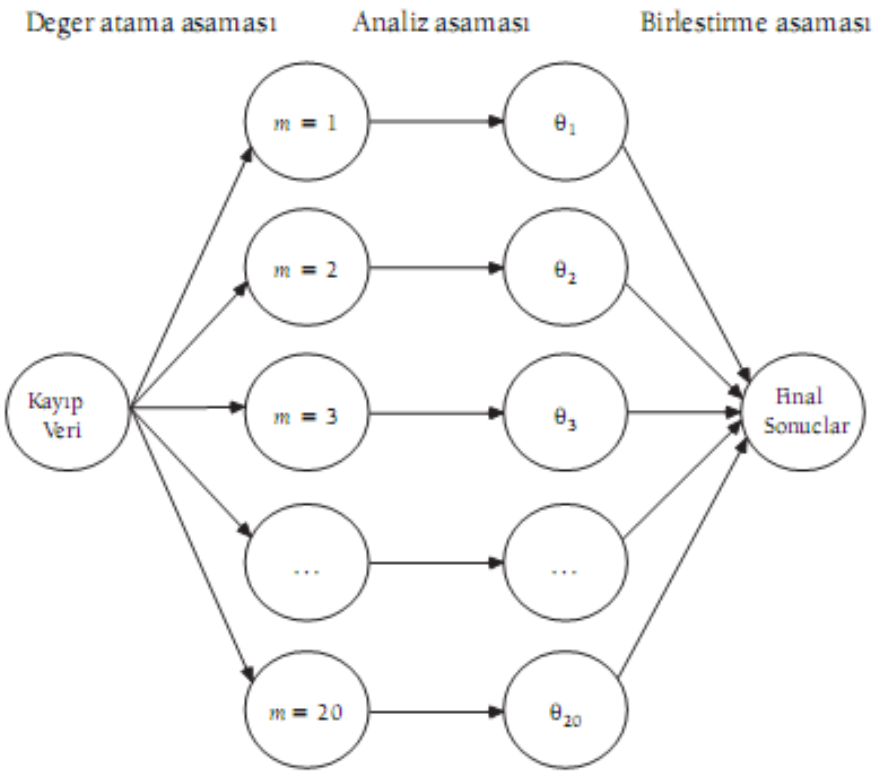
Kayıp değerler Markov zinciri Monte Carlo (MCMC) yöntemi ile tamamlanarak, kayıp değerleri tamamlanmış veri setleri oluşturulur. Kayıp değerleri tamamlanmış her bir veri seti eksiksiz veri gibi uygun bir istatistiksel teknikle analiz edilir. Daha sonra her bir veri setinden elde edilen çıkarımlar birleştirilerek tekbir çıkarım elde edilir.

Tek bir bağımlı değişken söz konusu olduğunda bağımsız değişkenler lineer regresyon modeli ile modellenir. Regresyon parametrelerinin sonsal dağılımları hesaplanır ve kayıp değerler tahmin edilerek tamamlanmış veri seti oluşturulur. Bu işlem *m* kez tekrarlanır.

Çok değişkenli durumlar için her bir kayıp değerli bağımsız değişken için koşullu bir dağılım varsayılır. Koşullu dağılımlardan çekim Gibbs örnekleyici kullanılarak oluşturulur.

Çoklu değer atama üç adımda gerçekleştirilir.

I. Uygun bir değer atama modelinden kayıp gözlemler için tahmin edilen makul değerlerden *m* elemanlı bir küme oluşturulur. Böylece m tane tam veri seti oluşturulur.
II. Eksiklikleri tamamlanmış veri seti belli bir yöntem veya model ile analiz edilir.
III. *m* tane analizin sonuçları birleştirilir.



Şekil 3.2. Çoklu değer atama aşamasının grafiksel gösterimi

Çoklu değer atamanın aşamaları Şekil 3.2'de verilen grafiksel gösterimde açıkça belirtilmiştir (Enders,2010).

SAS'da MI prosüdürü ile birden fazla bağımsız değişken için çoklu değer atama süreci gerçekleştirilir. Bu süreçte çok değişkenli kayıp değere sahip *p* boyutlu veri için kayıp değer yerine atama yapılmış veri setleri oluşturulur. Değer atanmış veri seti sayısı *m* ile gösterilir. Daha sonra bu *m* tam veri seti herhangi bir istatistiksel bir teknik ile analiz edilir. Son olarak MIANALYZE prosüdürü ile *m* tam veri setinden elde edilen sonuçların birleştirilmesi ile parametreler hakkında istatistiksel çıkarımlar elde edilir.

i) Değer Atama Aşaması

MCMC, istatistiksel uygulamalarda Markov zinciri yoluyla çok boyutlu olasılık dağılımlarından tesadüfi çıkarımlar üretmek için kullanılır. Markov zinciri, her bir elemanının dağılımı bir öncekine bağlı olan tesadüfi değişkenlerin bir dizisidir. MCMC elemanların dağılımının ortak bir dağılımda dengelemek için yeterli uzunlukta Markov

zinciri oluşturur. Bu sabit dağılım ilgilenilen dağılımdır ve bu dağılımdan simülasyonlar yapılır ve zincirin simülasyon adımları tekrarlanır.

Bayesci çıkarımda bilinmeyen parametreler hakkındaki bilgi sonsal olasılık dağılımı formunda açıklanır. MCMC, Bayesci çıkarımda sonsal dağılımları bulmak için kullanılan bir yöntemdir. Yani MCMC yoluyla bilinmeyen parametrelerin ortak dağılımı simüle edilebilir. Elde edilen simülasyon temelli sonsal parametrelerin tahminleri ile ilgilenilir.

Çoklu değer atama yönteminde verinin çok değişkenli normal dağıldığı varsayılır ve aşağıdaki adımlar izlenerek uygulanır.

I-ADIM (Değer Atama)

I adımı gözlenen değişkenlerden kayıp değerleri tahmin eden regresyon eşitlikleri oluşturmak için ortalama vektörünün ve kovaryans matrisinin elemanlarını kullanır. Veride kayıp değerli tek değişken olduğunda eşitlik (3.35) de verildiği gibi bir regresyon eşitliği kurulur.

$$Y^{*}_{i} = \hat{\beta}_0 + \hat{\beta}_1 x_i + z_i \qquad (3.35)$$

Burada Y^{*}_{i}, i.nci gözlemin kayıp değerli değişkeni için atanan değeri; $\hat{\beta}_0, \hat{\beta}_1$ regresyon katsayılarını z_i, ortalaması 0, varyansı regresyonun artık varyansına eşit olan normal dağılımlı rasgele artıklardır. Verilen regresyon eşitliği ile her bir kayıp değer için tahmin skorları ve bu skorlara artık terimin eklenmesiyle atanan değerler elde edilir. Her bir tahmin skoruna z_i artık teriminin eklenmesi, gözlenen veri bilindiğinde kayıp değer yerine konulacak mümkün değerlerin bir dağılımından rasgele bir çekimin olduğu anlamına gelir ve eşitlik (3.36)'da verildiği gibi gösterilir..

$$Y_{kay}^{(t+1)} \sim P(Y_{kay} / Y_{göz}, \theta^{(t)}) \qquad (3.36)$$

Yani t.nci iterasyonda geçerli parametre tahmini $\theta^{(t)}$ olmak üzere I-Adımında $P(Y_{kay} / Y_{göz}, \theta^{(t)})$ dağılımından kayıp veri için rasgele olarak bir örneklem tahmininde bulunulur. Elde edilen tahminler $Y_{kay}^{(t+1)}$ olarak gösterilir.

P-ADIM (Sonsal)

P adımı ortalama vektörü ve kovaryans matrisinin sonsal dağılımını tanımlayan bir Bayes analizidir. Bu adımda I adımında oluşturulan tamamlanmış veri kullanılarak ortalama vektörü ve kovaryans matrisi tahmin edilir. Bu tahmin vektörü ve matrisinin her bir elemanına artık terimin eklenmesiyle yeni bir parametre seti oluşturulur. Yani ortalama vektörü ve kovaryans matrisi onların sonsal dağılımından olarak çekilir. Bu tahminler simüle edilmiş değerlerdir. Çünkü Monte Carlo simülasyon tekniği ile türetilmişlerdir. Kovaryans matrisinin (Σ) sonsal dağılımı,

$$P(\Sigma / \hat{\mu}, Y) \sim W^{-1}(N-1, \hat{\Lambda}) \tag{3.37}$$

şeklinde gösterilir. Burada $P(\Sigma / \hat{\mu}, Y)$, sonsal dağılım, $\hat{\mu}$, örneklem ortalama vektörü, Y, I-adımında tamamlanmış veri matrisi, W^{-1}, ters Wishart dağılımı, N-1, serbestlik derecesini, $\hat{\Lambda}$, örneklem kareler ve çarpımlar toplamı matrisini gösterir ve dağılımın yayılımını tanımlar. Bu dağılım kullanılarak kovaryans matrisi türetilir.
Ortalama vektörünün ($\hat{\mu}$) sonsal dağılımı;

$$P(\hat{\mu} / Y, \Sigma) \sim MN(\hat{\mu}, N^{-1}\Sigma^{*}) \tag{3.38}$$

şeklinde gösterilir. Burada $P(\hat{\mu} / Y, \Sigma)$ sonsal dağılımı; MN, çok değişkenli normal dağılım, $\hat{\mu}$, örneklem ortalama vektörü ve Σ^{*}, simüle edilmiş kovaryans matrisini gösterir. Bu dağılım kullanılarak ortalama vektörü simüle edilir.

Sonsal dağılımlarından yeni parametreler çekilmesinden sonra I adımı güncellenmiş tahminleri kullanarak yeni regresyon katsayılar setini ve farklı atama değerlerini oluşturur. Bu yeni atamalar bir sonraki P adımına taşınır. Burada yeni parametre setleri oluşturulur (Enders, 2010).

P adımı eşitlik (3.39) ile matematiksel olarak özetlenebilir.

$$\theta^{(t+1)} \sim P(\theta / Y_{göz}, Y_{kay}^{(t+1)}) \tag{3.39}$$

Yani P-Adımında $Y_{kay}^{(t+1)}$ elde edildiğinde $P(\theta / Y_{göz}, Y_{kay}^{(t+1)})$ olasılık dağılımından $\theta^{(t+1)}$ ile gösterilen parametre tahmini elde edilir

Bu iki adım kayıp değerleri tahmin edilmiş güvenilir bir veri seti elde etmek için yeterince çok tekrarlanır (Schafer, 1997). Amaç iterasyonlar sonucu değişmeyen stabil bir dağılıma yakınsama yapmak ve daha sonra bu dağılımdan kayıp değerlerin yaklaşık bağımsız değerlerini elde etmektir.

Sonuçta bu tahminlerden oluşan ve $P(Y_{kay}, \theta / Y_{göz})$ dağılımına yakınsayan $(Y_{kay}^{(1)}, \theta^{(1)}), (Y_{kay}^{(2)}, \theta^{(2)}),\ldots\ldots$ markov zinciri oluşturulur.

ii) Atanmış Veri Setlerindeki Sonuçların Birleştirilmesi

Kayıp değerli verilerin analizinde kullanılan değer atama işlemi üç aşamada gerçekleşir. Birinci aşama değer atama aşaması, ikinci aşama parametre tahmini için analiz aşaması ve son aşama parametre tahminlerinin sonuçlarının birleştirildiği birleştirme aşamasıdır. Bu bölümde birleştirme aşaması anlatıldı. Birleştirme aşamasındaki amaç, m farklı veri kümesine uygulanan tekniğin sonuçlarının tek bir kümede birleştirilmesidir. Rubin (1987) birleştirilmiş parametre tahminleri ve standart hatalar için formülleri özetlemiştir. Örneğin birleştirilmiş parametre tahminleri analiz aşamasında elde edilen tahminlerin basit bir aritmetik ortalamasıdır. Standart hataların birleştirilmesi ise biraz daha karmaşık fakat yine aynı mantıkla elde edilmektedir.

Kayıp veri yapısı tesadüfi (MAR) yapıda olduğunda her bir parametre için m farklı yansız tahminler elde edilir. Bu m tane parametre tahmini tek bir nokta tahmininde birleştirilir. Rubin (1987), m tane tahminin aritmetik ortalaması olarak çoklu değer ataması sonucu elde edilen nokta tahmini olarak eşitlik (3.40)'da verildiği gibi tanımlanır.

$$\overline{Q} = \frac{1}{m}\sum_{i=1}^{m}\hat{Q}_i \qquad (3.40)$$

Burada $\hat{Q}_i$ i.nci kayıp değer atanmış veri setinden elde edilmiş parametre tahmini, $\overline{Q}$ ise birleştirilmiş nokta tahminidir. Çoklu değer atama işleminde varyans iki kaynaktan

hesaplanmaktadır. Bunlar atamalar içi ve atamalar arası varyansdır. Atamalar içi varyans, *m* tane örneklem varyansının aritmetik ortalamasıdır.

$$\bar{U} = \frac{1}{m}\sum_{t=1}^{m} SE_i^2 \tag{3.41}$$

Burada SE_i^2, i.nci veri setinden elde edilen standart hatanın karesidir. Yani örneklem varyansıdır. Atamalar içi varyans etkin bir şekilde örneklem değişkenliğini tahmin eder. Atamalar arası varyans, *m* veri setindeki parametre tahminlerinin değişkenliğinin miktarıdır.

$$B = \frac{1}{m-1}\sum_{i=1}^{m} (\hat{\theta}_i - \bar{\theta})^2 \tag{3.42}$$

Burada $\hat{\theta}_i$ i.nci atanmış veri setinden elde edilmiş parametre tahmini, $\bar{\theta}$ ise birleştirilmiş nokta tahminidir.

Toplam örneklem varyansı, atamalar içi varyans ve atamalar arası varyansı tek bir miktarda birleştirir.

$$T = \bar{U} + (1 + \frac{1}{m})B \tag{3.43}$$

Kayıp bilgi oranı (fraction of missing information), kayıp verinin parametre tahminin örneklem varyansı üzerine etkisinin miktarıdır. Kayıp bilgi oranı 0,715 olduğu zaman bunun anlamı, regresyon katsayısının örneklem varyansının %71,5'inin kayıp veriden kaynaklı olmasıdır. Kayıp bilgi oranı yüksek olan parametrelere yakınsama yavaş olur. Kayıp bilgi oranı ve varyansdaki nispi artış,

$$\hat{\lambda} = \frac{r + 2/(v_m + 3)}{r + 1}, \qquad r = \frac{(1 + \frac{1}{m})B}{\bar{U}} \tag{3.44}$$

(3.44) eşitliği ile elde edilir. Kayıp verinin neden olduğu örneklem varyansındaki artış oranı varyansdaki nispi artış verir. Parametrelerin standart hataları üzerinde kayıp verinin etkisi yoksa varyansdaki nispi artış (r) ve atamalar arası varyans sıfır olur. Buna karşın atamalar arası varyans ve atamalar içi varyans birbirine eşit olursa varyansdaki nispi artış 1 olur.

iii) Atanmış Veri Seti Sayısının Belirlenmesi

Çoklu değer atama analizlerinde en temel kararlardan biri, oluşturulacak veri seti sayısının belirlenmesidir. Kaynaklar incelendiğinde birkaç tane veri setinin yeterli olduğu özellikle de 3-5 arasında olması tavsiye edilir (Rubin 1987, Rubin 1996; Schafer 1997; Schafer & Olsen, 1998). Fakat daha fazla veri seti kullanmak için pek çok neden vardır. Bunların en önemlisi, veri seti sayısı büyük alındığında standart hatadaki azalmadır. Hatta sonsuz sayıda veri seti analiz edildiğinde mümkün olan en küçük standart hata elde edilecektir. m tane değer atamaya dayalı tahminlerin etkisi Rubin(1987) tarafından Nispi Etki (NE) olarak

$$\left(1+\frac{\hat{\lambda}}{m}\right)^{-1} \tag{3.45}$$

Eşitliği ile gösterilir. Burada $\hat{\lambda}$, kayıp bilgi oranıdır (fraction of missing information). Bu oran kayıp değerli verinin ne kadar hassas olduğunu rakamlarla gösterir (SAS'da son tabloda her bir parametre için otomatik olarak veriliyor). Örneğin m=5 ve belli bir parametre için kayıp bilgi oranı $\hat{\lambda}$=0,20 için nispi etki 0,96 olacaktır. Bunun anlamı, sonsuz veri seti sayısına bağlı olan örneklem varyansı, m=5 veri setine bağlı örneklem varyansının %96'sı olmasıdır. Pratik açıdan bakıldığında 5 veri setinin analiz edilmesi varsayılan minimum standart hata değerinden sadece $\sqrt{\left(1+\frac{0,20}{5}\right)}=1,02$ kez daha büyük bir standart hata oluşturur. Aşağıdaki tablo farklı kayıp bilgi oranları ve farklı veri seti sayıları için standart hatadaki artış oranını (SHAO) ve nispi etkiyi (NE) gösterir.

1. Standart hatadaki en büyük azalma veri seti sayısı 3-5 arasında olduğu zamanıdır. Veri seti sayısı 10 dan büyük olduğunda fayda çok az olur.
2. Kayıp bilgi oranı büyük olduğunda veri seti sayısının büyük tutulması gerekir.

Çizelge 3.2. Farklı kayıp bilgi oranları ve farklı veri seti sayıları için standart hatadaki artış oranı (SHAO) ve nispi etki (NE)

	m=3		m=5		m=10		m=20	
$\hat{\lambda}$	NE	SHAO	NE	SHAO	NE	SHAO	NE	SHAO
0,10	0,97	1,02	0,98	1,01	0,99	1,00	1,00	1,00
0,30	0,91	1,05	0,94	1,03	0,97	1,01	0,99	1,01
0,50	0,86	1,08	0,91	1,05	0,95	1,02	0,98	1,01
0,70	0,81	1,11	0,88	1,07	0,93	1,03	0,97	1,02

Çizelge 3.2 incelendiğinde $\hat{\lambda}$'nın %30 , m'in 5 olması durumunda %94'lik bir etkiye sahip olunur. m'nin değeri artırılıp m=10 alındığında etki 0,97 olacaktır. m'nin değeri ikiye katlanmasına rağmen etkide az bir kazanç görülmektedir. Çoğu durumda m sayısının büyük tutulması ve analiz edilmesinin çok az avantajı olmaktadır.

Çoklu değer atamanın avantajları şunlardır.

- Standart tam veri yöntemleri ve yazılımları ile birlikte çalışır**.**
- Çoklu değer atama kullanışlı bir yöntemdir çünkü m sayısı küçük olsa bile iyi sonuçlar elde edilir. Uygulamaların çoğunda 3-5 atanmış veri seti iyi sonuçlar elde etmek için yeterli olmaktadır.

4. BULGULAR VE TARTIŞMA

Bu çalışmada 4 farklı uygulama yapıldı. Kayıp veri değerlerinin farklı yöntemlerle hesaplanması için regresyon değer atama ve EM algoritması kullanımında SPSS paket programından, çoklu değer atama için SAS 9.3 programından ve ortalama değer atama için MS EXCEL programından yararlanıldı. Bayesci Cox Regresyon (BCR) ve Cox Regresyon (CR) analizleri için SAS 9.3 programı kullanıldı. Analiz sonuçlarının düzenlenmesi ve sonuçların grafiksel sunumu için MS EXCEL programı kullanıldı.

4.2.Uygulama I

Çalışmada ilk önce kayıp veri problemini en etkin ve doğru şekilde belirlemek için farklı örnek genişlikli verilerde ve farklı kayıp oranlarında kayıp veri yöntemlerinin performansları CR uygulanarak karşılaştırıldı.

4.1.1.Cox Regresyonu ile Kayıp Veri Analizi Yöntemlerinin Karşılaştırılması

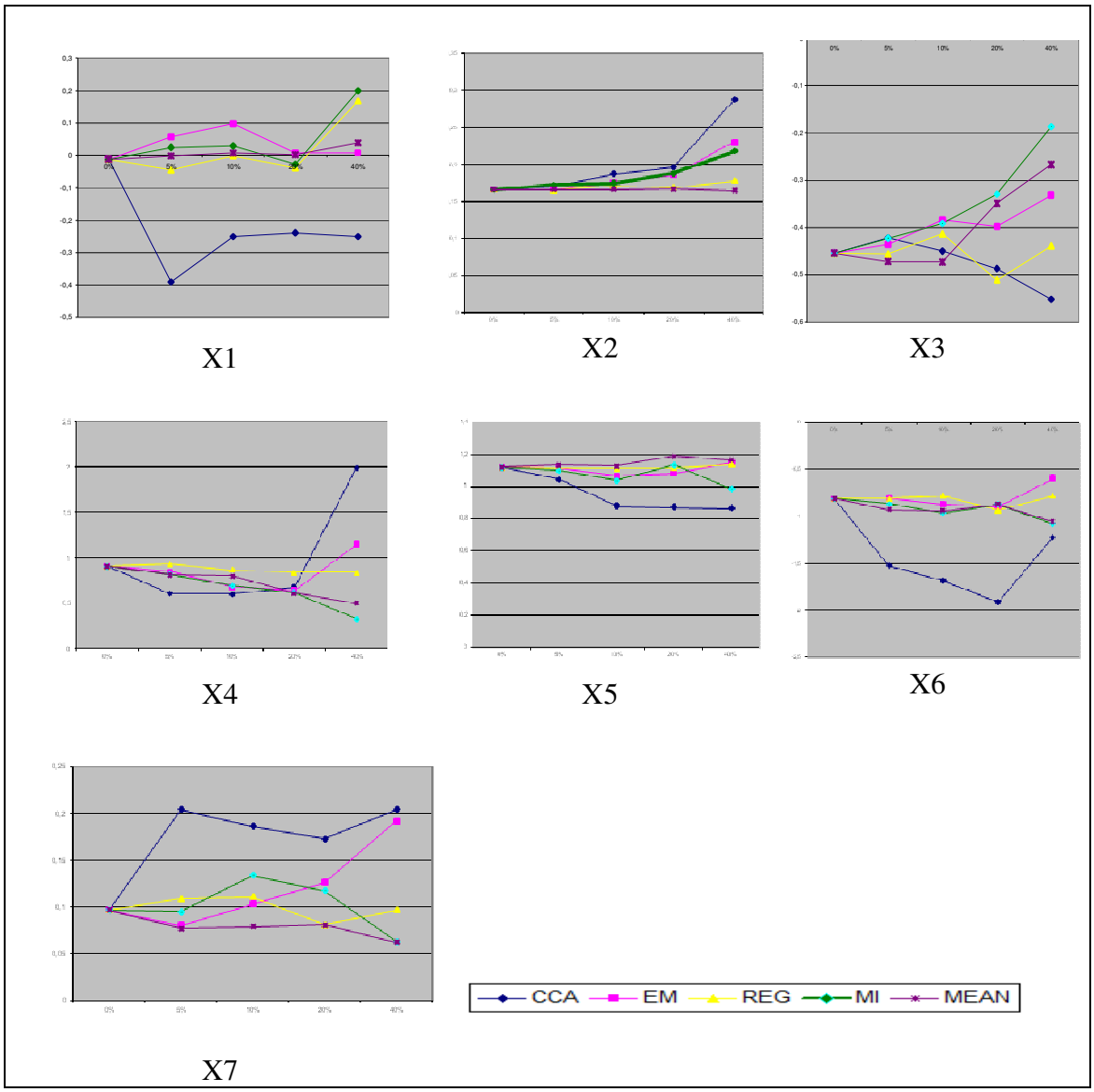
Örnek genişliği 50, 100 ve 200 olan yedi değişkenli sağkalım verisi kullanılarak ve her bir veride kayıp gözlem oranı %5, %10, %20 ve %40 olacak şekilde MAR varsayımına uygun bir şekilde değerler silinerek, kayıp değerli veriler elde edildi.

Farklı örnek genişlikli ve farklı kayıp oranlı veri setleri eksiksiz veri analizi (Complete Case Analysis-CCA), beklenti maksimizasyonu (Expectation Maximization-EM), regresyonla değer atama (Regression Imputation-REG), çoklu değer atama (Multiple Imputation-MI) ve ortalama değer atama (Mean Imputation) yöntemleri ile tamamlanıp, elde edilen verilere CR analizi uygulandı. Analiz sonucu elde edilen regresyon katsayıları ve standart hatalar kayıp değer içermeyen tam verinin CR analiz sonuçları ile karşılaştırıldı. Karşılaştırma yapılırken kayıp değer içermeyen tam verinin sonuçları gerçek sonuçlar olarak ele alınıp, kayıp değer tamamlama yöntemleri bu gerçek sonuçlara yakınlığı bakımından incelendi.

Farklı kayıp oranlı 50 örnek genişlikli veri için elde edilen regresyon katsayıları Çizelge 4.1'de verilmiştir. Sonuçların daha açık görülebilmesi için her bir değişkenin regresyon katsayısının aldığı değerler Şekil 4.1'de grafiksel olarak gösterildi.

Çizelge 4.1. Farklı kayıp oranlı 50 örnek genişlikli veri için regresyon katsayıları

Değişkenler	Kayıp oranı	N=50				
		CCA	**EM**	**REG**	**MI**	**MEAN**
X1	0%	-0,012	-0,012	-0,012	-0,012	-0,012
	5%	-0,391	0,057	-0,045	0,025	-0,002
	10%	-0,251	0,098	-0,003	0,029	0,008
	20%	-0,239	0,008	-0,039	-0,027	0,002
	40%	-0,250	0,008	0,168	0,2	0,039
X2	0%	0,177	0,177	0,177	0,177	0,177
	5%	0,203	0,193	0,168	0,196	0,207
	10%	0,241	0,202	0,172	0,187	0,214
	20%	0,319	0,222	0,22	0,182	0,218
	40%	0,045	-0,036	0,129	0,188	0,233
X3	0%	-0,454	-0,454	-0,454	-0,454	-0,454
	5%	-0,422	-0,436	-0,456	-0,423	-0,472
	10%	-0,45	-0,384	-0,413	-0,391	-0,473
	20%	-0,487	-0,398	-0,511	-0,33	-0,349
	40%	-0,552	-0,332	-0,439	-0,187	-0,267
X4	0%	0,903	0,903	0,903	0,903	0,903
	5%	0,603	0,846	0,93	0,815	0,809
	10%	0,599	0,678	0,861	0,69	0,799
	20%	0,672	0,638	0,831	0,616	0,607
	40%	1,989	1,148	0,833	0,323	0,498
X5	0%	1,119	1,119	1,119	1,119	1,119
	5%	1,043	1,116	1,114	1,1	1,135
	10%	0,879	1,064	1,114	1,035	1,127
	20%	0,874	1,081	1,118	1,136	1,189
	40%	0,867	1,148	1,137	0,986	1,164
X6	0%	-0,813	-0,813	-0,813	-0,813	-0,813
	5%	-1,524	-0,808	-0,811	-0,867	-0,932
	10%	-1,684	-0,873	-0,782	-0,963	-0,942
	20%	-1,915	-0,899	-0,938	-0,876	-0,876
	40%	-1,222	-0,6	-0,782	-1,081	-1,054
X7	0%	0,097	0,097	0,097	0,097	0,097
	5%	0,204	0,08	0,109	0,095	0,077
	10%	0,186	0,103	0,111	0,134	0,079
	20%	0,173	0,126	0,081	0,117	0,081
	40%	0,204	0,192	0,097	0,063	0,062



Şekil 4.1. Farklı kayıp oranlı 50 örnek genişlikli veri için regresyon katsayılarının grafiksel gösterimi

50 birimlik örnek genişliğinde farklı kayıp oranlı veriler için kayıp veri tamamlama yöntemleri uygulandı. Oluşturulan tamamlanmış verilere CR uygulanarak elde edilen sonuçlar kayıp değer içermeyen yani tam verinin sonuçlarıyla karşılaştırıldı.

Şekil 4.1'de CCA ile tamamlanmış veriden elde edilen regresyon katsayıları tam verinin katsayıları ile karşılaştırıldığında genellikle gerçek sonuçlara yaklaşılmadığı gözlendi.

MEAN yöntemi genel olarak gerçek parametre değerlerinden çok uzaklaşmamış fakat en iyi tahminleri de elde etmemiştir. MEAN yöntemi CCA ile karşılaştırıldığında daha iyi bir performans sergilemiştir.

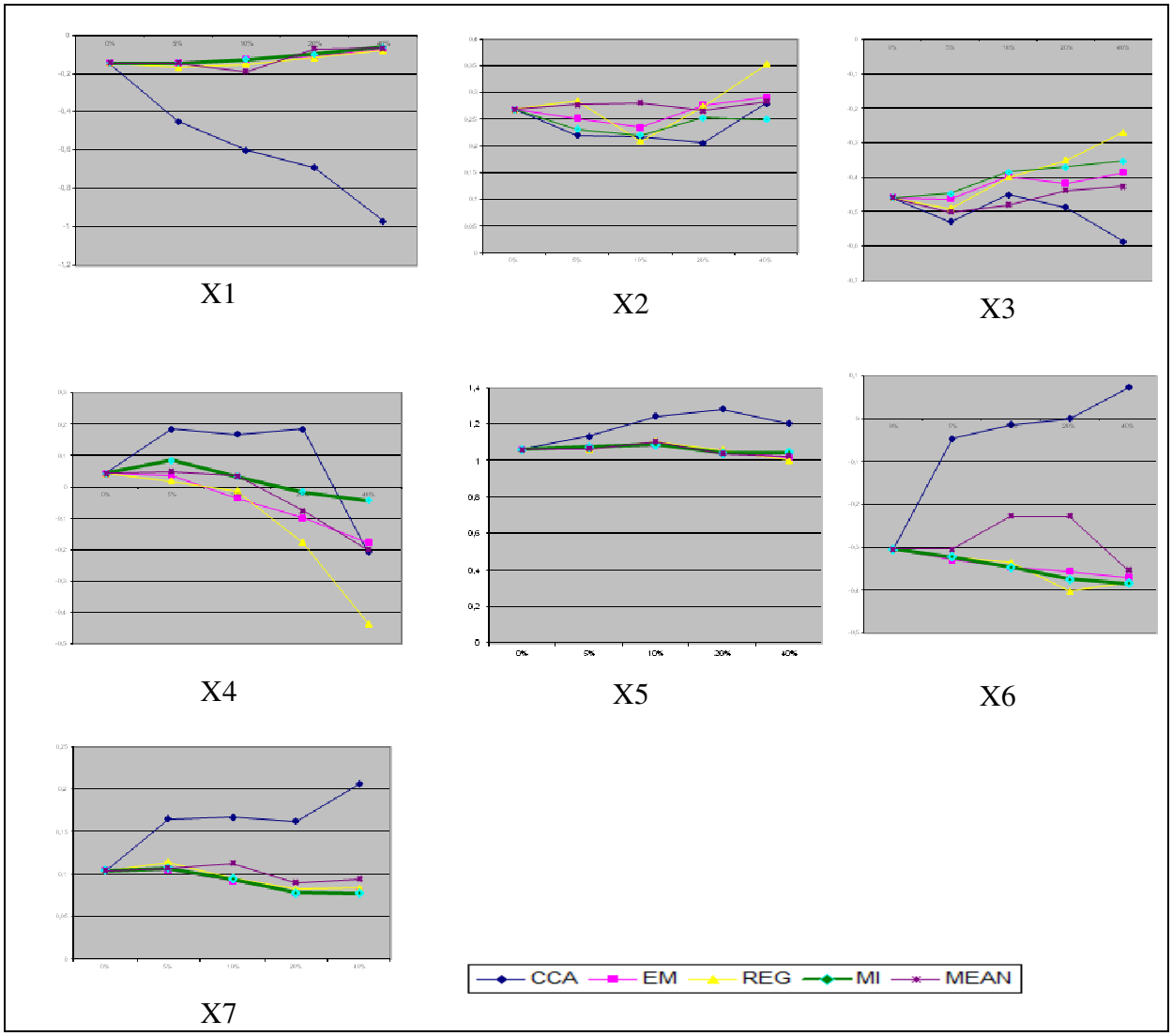
EM ve MI yöntemleri 50 birimlik örnek genişliği için %20 ve daha az kayıp oranı için tam veriden elde edilen regresyon katsayısına genel olarak yakın değerler elde ederken, kayıp oranı %20'den büyük olduğunda gerçek sonuçlardan biraz uzaklaşmıştır.

REG yöntemi ile tamamlanan verilerin CR sonucu elde edilen parametre değerleri neredeyse bütün kayıp oranlarında gerçek sonuca yakın sonuçlar elde etmiştir.

Farklı kayıp oranlı 100 örnek genişlikli veri için elde edilen regresyon katsayıları Çizelge 4.2'de verilmiştir. Sonuçlar ise Şekil 4.2'de grafiksel olarak gösterildi.

Çizelge 4.2. Farklı kayıp oranlı 100 örnek genişlikli veri için regresyon katsayıları

Değişkenler	Kayıp oranı	N=100				
		CCA	**EM**	**REG**	**MI**	**MEAN**
X1	0%	-0,144	-0,144	-0,144	-0,144	-0,144
	5%	-0,45	-0,143	-0,165	-0,145	-0,144
	10%	-0,598	-0,126	-0,151	-0,126	-0,189
	20%	-0,69	-0,106	-0,117	-0,099	-0,07
	40%	-0,97	-0,073	-0,08	-0,06	-0,065
X2	0%	0,269	0,269	0,269	0,269	0,269
	5%	0,22	0,252	0,284	0,231	0,277
	10%	0,217	0,235	0,209	0,22	0,281
	20%	0,206	0,276	0,274	0,253	0,266
	40%	0,28	0,291	0,353	0,249	0,283
X3	0%	-0,459	-0,459	-0,459	-0,459	-0,459
	5%	-0,529	-0,463	-0,492	-0,448	-0,502
	10%	-0,451	-0,397	-0,401	-0,384	-0,48
	20%	-0,488	-0,418	-0,351	-0,370	-0,439
	40%	-0,587	-0,386	-0,269	-0,354	-0,427
X4	0%	0,043	0,043	0,043	0,043	0,043
	5%	0,184	0,037	0,018	0,085	0,048
	10%	0,167	-0,036	-0,012	0,033	0,035
	20%	0,184	-0,098	-0,178	-0,016	-0,076
	40%	-0,207	-0,178	-0,435	-0,044	-0,201
X5	0%	1,061	1,061	1,061	1,061	1,061
	5%	1,133	1,074	1,06	1,075	1,064
	10%	1,243	1,085	1,102	1,087	1,104
	20%	1,283	1,048	1,057	1,04	1,039
	40%	1,202	1,018	1,003	1,044	1,024
X6	0%	-0,305	-0,305	-0,305	-0,305	-0,305
	5%	-0,047	-0,329	-0,323	-0,323	-0,305
	10%	-0,015	-0,347	-0,335	-0,346	-0,227
	20%	0	-0,356	-0,403	-0,374	-0,227
	40%	0,073	-0,37	-0,381	-0,384	-0,353
X7	0%	0,104	0,104	0,104	0,104	0,104
	5%	0,165	0,105	0,114	0,107	0,107
	10%	0,167	0,092	0,096	0,094	0,113
	20%	0,162	0,078	0,082	0,078	0,09
	40%	0,206	0,079	0,084	0,077	0,094



Şekil 4.2. Farklı kayıp oranlı 100 örnek genişlikli veri için regresyon katsayılarının grafiksel gösterimi

100 birimlik örnek genişliğinde CCA yönteminin tahminlerinin gerçek değere yaklaşamadığı bütün kayıp oranlarında gözlenmektedir. %5 kayıp oranında CCA dışındaki yöntemler gerçek değere yakın sonuçlar vermiştir. Bu kayıp oranında 50 birimlik örneğe göre yöntemler gerçek değere çok daha fazla yaklaşmışlardır.

MEAN yöntemi kayıp oranı %10 ve daha küçük olduğunda gerçek değere çok yakın değerler elde ederken, kayıp oranı %10'dan büyük olduğunda sonuçlar gerçek değerden uzak kalmıştır. MEAN yönteminde 50 birimlik örneğe göre 100 birimlik örnekteki parametre tahminleri gerçek parametre değerine daha çok yaklaşmıştır.

EM yönteminde 50 birimlik örnek genişliğine göre 100 birimlik örnek genişliğinde elde edilen parametrelerin gerçek değere daha çok yaklaştığı ve bütün

kayıp değer oranlarında gerçek parametre değerlerinden çok uzaklaşılmadığı gözlenmiştir. Ayrıca %5 ve %10'a göre %20 ve %40 kayıp oranlarında gerçek parametre değerlerine biraz daha uzak değerler elde edilmiştir.

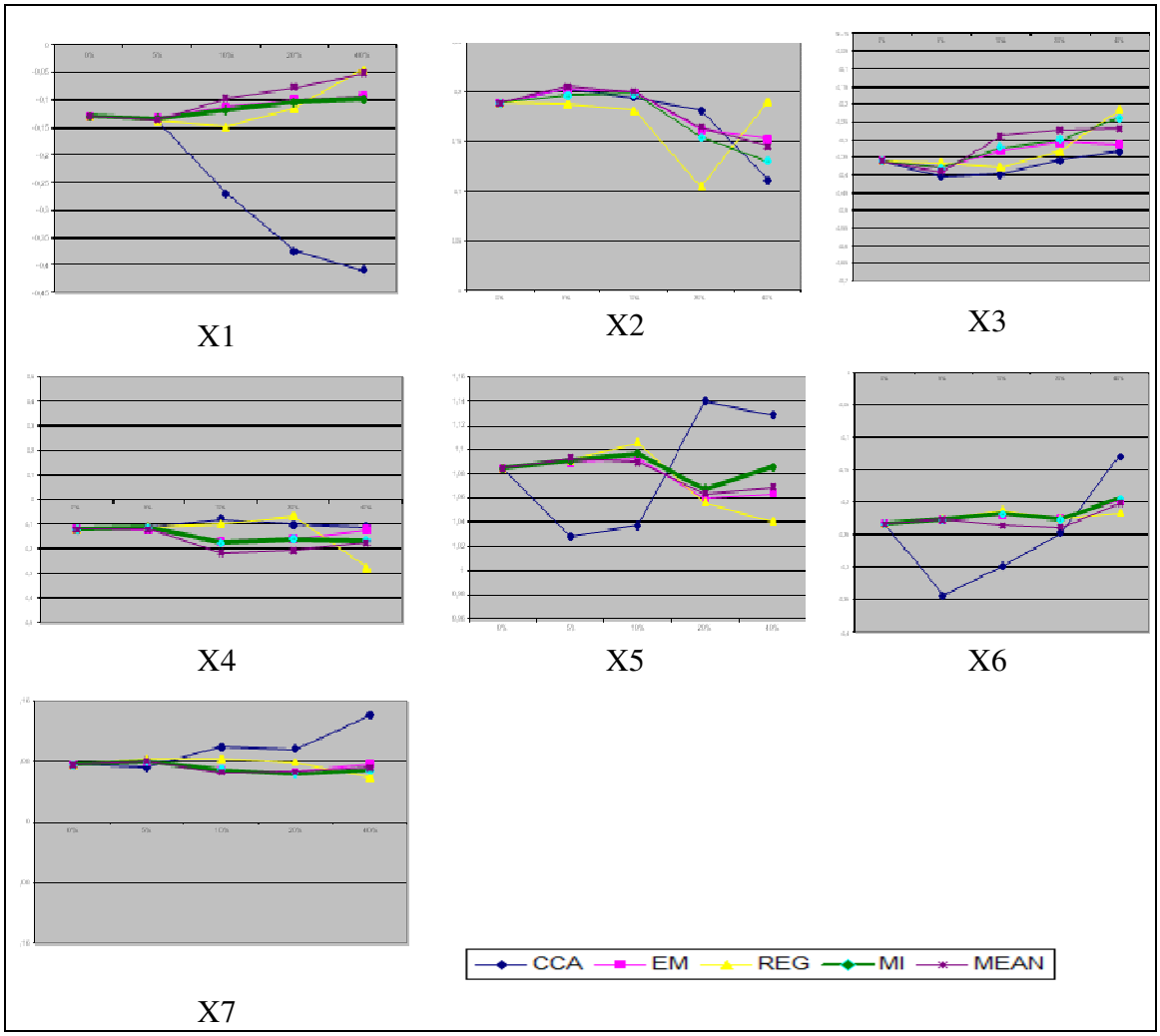
REG yönteminde kayıp değer oranı %5 ve %10 için gerçek değere yakın sonuçlar elde edilirken %20 ve %40'da sapmalar artmıştır.

MI yöntemi, EM değer atama yöntemine benzer şekilde örnek genişliğinin artması ile parametreye benzerlik açısından daha iyi bir performans göstermiştir. Bütün kayıp oranlarında gerçek parametre değerine yakın sonuçlar elde etmiştir.

Farklı kayıp oranlı 200 örnek genişlikli veri için elde edilen regresyon katsayıları Çizelge 4.3'de, grafikleri ise Şekil 4.3'de verildi.

Çizelge 4.3. Farklı kayıp oranlı 200 örnek genişlikli veri için regresyon katsayıları

Değişkenler	Kayıp oranı	N=200				
		CCA	**EM**	**REG**	**MI**	**MEAN**
X1	0%	-0,129	-0,129	-0,129	-0,129	-0,129
	5%	-0,137	-0,132	-0,138	-0,134	-0,136
	10%	-0,27	-0,112	-0,149	-0,118	-0,097
	20%	-0,374	-0,1	-0,115	-0,104	-0,077
	40%	-0,409	-0,095	-0,045	-0,098	-0,052
X2	0%	0,189	0,189	0,189	0,189	0,189
	5%	0,203	0,202	0,187	0,196	0,205
	10%	0,194	0,199	0,181	0,198	0,200
	20%	0,181	0,162	0,104	0,154	0,164
	40%	0,111	0,153	0,19	0,13	0,145
X3	0%	-0,361	-0,361	-0,361	-0,361	-0,361
	5%	-0,403	-0,38	-0,366	-0,381	-0,39
	10%	-0,397	-0,33	-0,377	-0,322	-0,291
	20%	-0,358	-0,307	-0,332	-0,3	-0,274
	40%	-0,332	-0,313	-0,216	-0,239	-0,268
X4	0%	-0,119	-0,119	-0,119	-0,119	-0,119
	5%	-0,11	-0,122	-0,11	-0,114	-0,118
	10%	-0,077	-0,17	-0,097	-0,175	-0,217
	20%	-0,101	-0,16	-0,067	-0,159	-0,207
	40%	-0,111	-0,121	-0,274	-0,165	-0,175
X5	0%	1,085	1,085	1,085	1,085	1,085
	5%	1,028	1,09	1,091	1,091	1,093
	10%	1,038	1,093	1,106	1,097	1,09
	20%	1,14	1,06	1,056	1,068	1,064
	40%	1,129	1,064	1,041	1,086	1,069
X6	0%	-0,233	-0,233	-0,233	-0,233	-0,233
	5%	-0,343	-0,227	-0,226	-0,227	-0,227
	10%	-0,299	-0,22	-0,213	-0,218	-0,235
	20%	-0,247	-0,224	-0,226	-0,227	-0,238
	40%	-0,129	-0,198	-0,217	-0,194	-0,202
X7	0%	0,077	0,077	0,077	0,077	0,077
	5%	0,073	0,08	0,084	0,08	0,081
	10%	0,1	0,068	0,084	0,07	0,066
	20%	0,097	0,067	0,079	0,064	0,067
	40%	0,141	0,078	0,059	0,069	0,073



Şekil 4.3. Farklı kayıp oranlı 200 örnek genişlikli veri için regresyon katsayılarının grafiksel gösterimi

Farklı kayıp oranlı 200 birimlik örnek genişlikli veri, önce kayıp veri giderme yöntemleri ile tamamlandı. Daha sonra CR analizi uygulandığında her bir yöntem sonucu ile elde edilen regresyon katsayıları ve gerçek parametre değerleri Çizelge 4.3'de verilmiştir. Yöntemlerden elde edilen sonuçların gerçek parametre değerine benzerliğinin açıkça görülmesi için sonuçlar grafiksel olarak Şekil 4.3'de gösterilmiştir.

Çizelge 4.3.'deki değerler ve sonuçlar incelendiğinde CCA yöntemi genel olarak %5 kayıp oranında gerçek sonuca çok yakın değerler vermiştir. Ayrıca CCA yöntemi %10, %20 ve %40 kayıp oranlarında ise gerçek değere diğer yöntemler kadar yakın olmasa da diğer 50 ve 100 birimlik örnek genişliklerine göre daha iyi bir performans sergilemiştir.

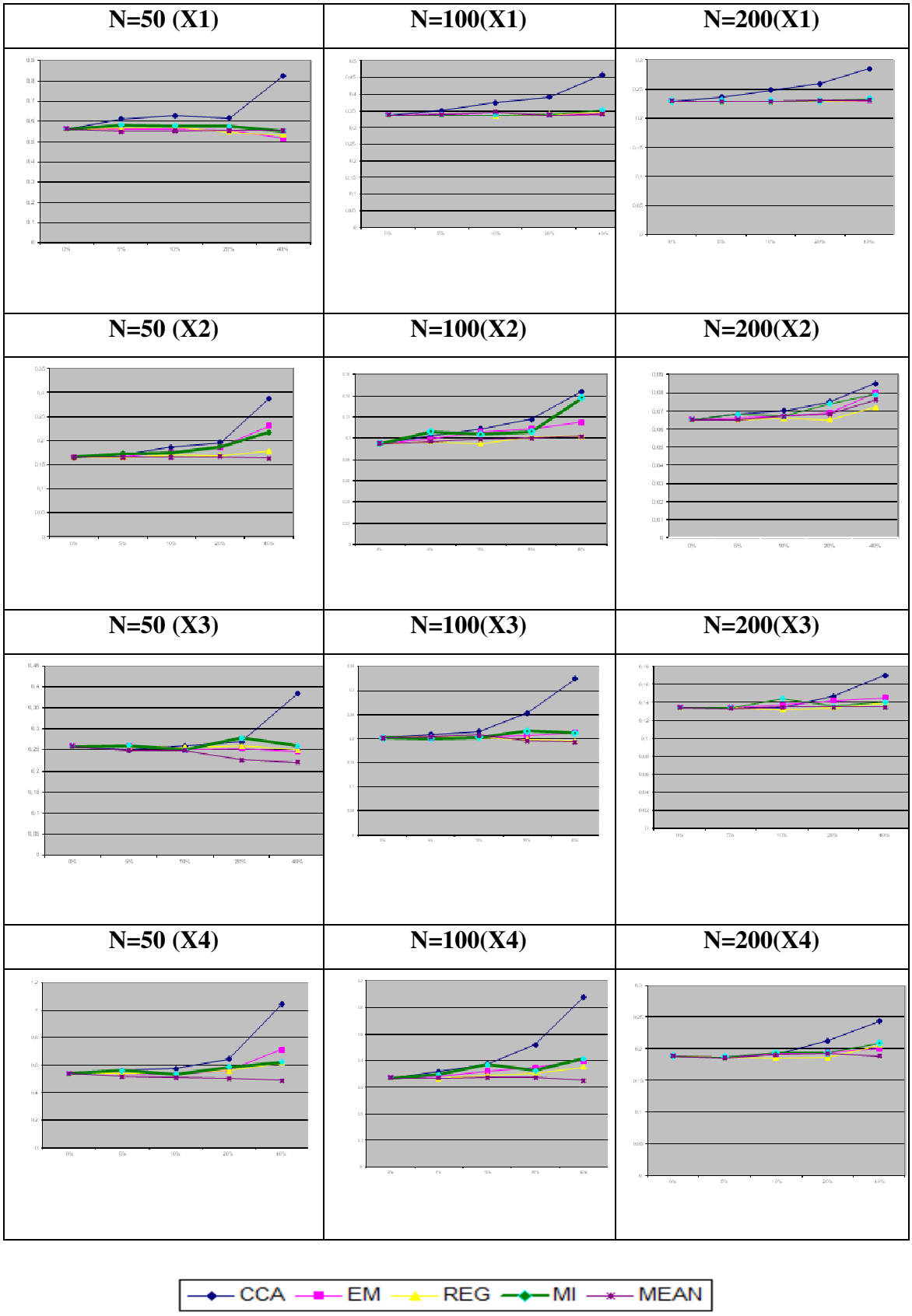
MEAN yönteminin farklı kayıp oranlarındaki performansı incelendiğinde %5 ve %10 kayıp oranlarında gerçek değere çok yakın değerler verdiği görülmüştür. %20 ve %40 kayıp değer oranlarında çok farklı sonuçlar vermemiş olmasına rağmen, gerçek değerden %5 ve %10'a göre biraz daha farklı sonuçlar bulunmuştur. MEAN yönteminin örnek genişliği arttıkça gerçek değere yakınlığı da artmıştır.

EM yöntemi genel olarak bütün kayıp oranlarında özelliklede %10'dan küçük kayıp oranlarında gerçek değere çok yakın sonuçlar vermiştir. EM yönteminde örnek genişliği arttıkça tahminlerdeki sapmada azalmıştır. Yani gerçek değere daha çok yaklaşılmıştır.

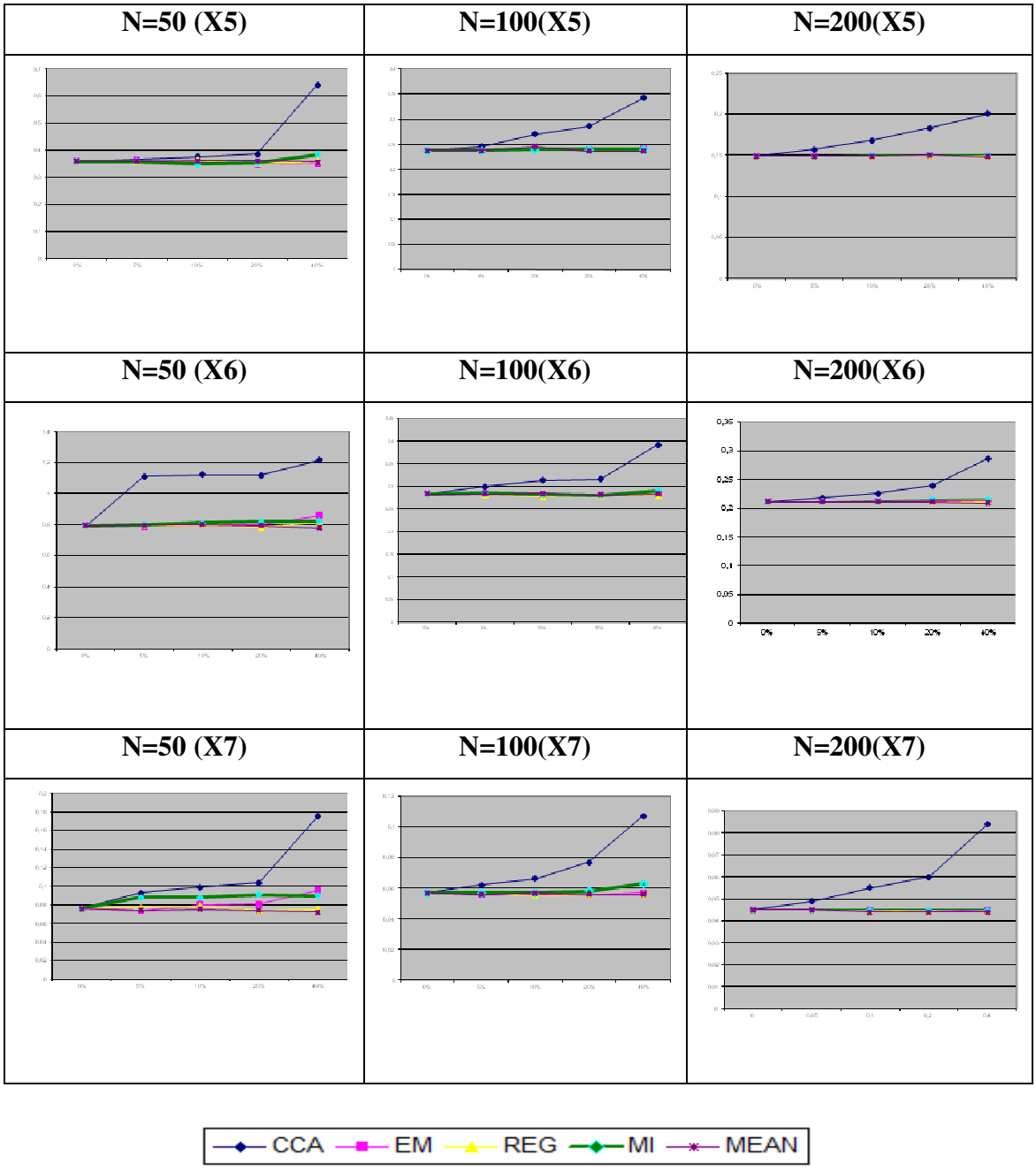
MI yönteminde hemen hemen bütün kayıp değer oranlarında gerçek parametre değerine yakın sonuçlar elde edilmiştir. Örnek genişliği arttıkça tahminler gerçek değere yaklaşmıştır.

REG yönteminde 200 birimlik örnek genişliğinde kayıp değer oranı %10 ve altı için tahminler gerçek değere yakın fakat %10'un üstü için gerçek değerden uzak sonuçlar elde etmiştir.

Farklı örnek genişlikleri ve farklı kayıp oranları için elde edilen standart hata grafikleri Şekil 4.4'de sunulmuştur.



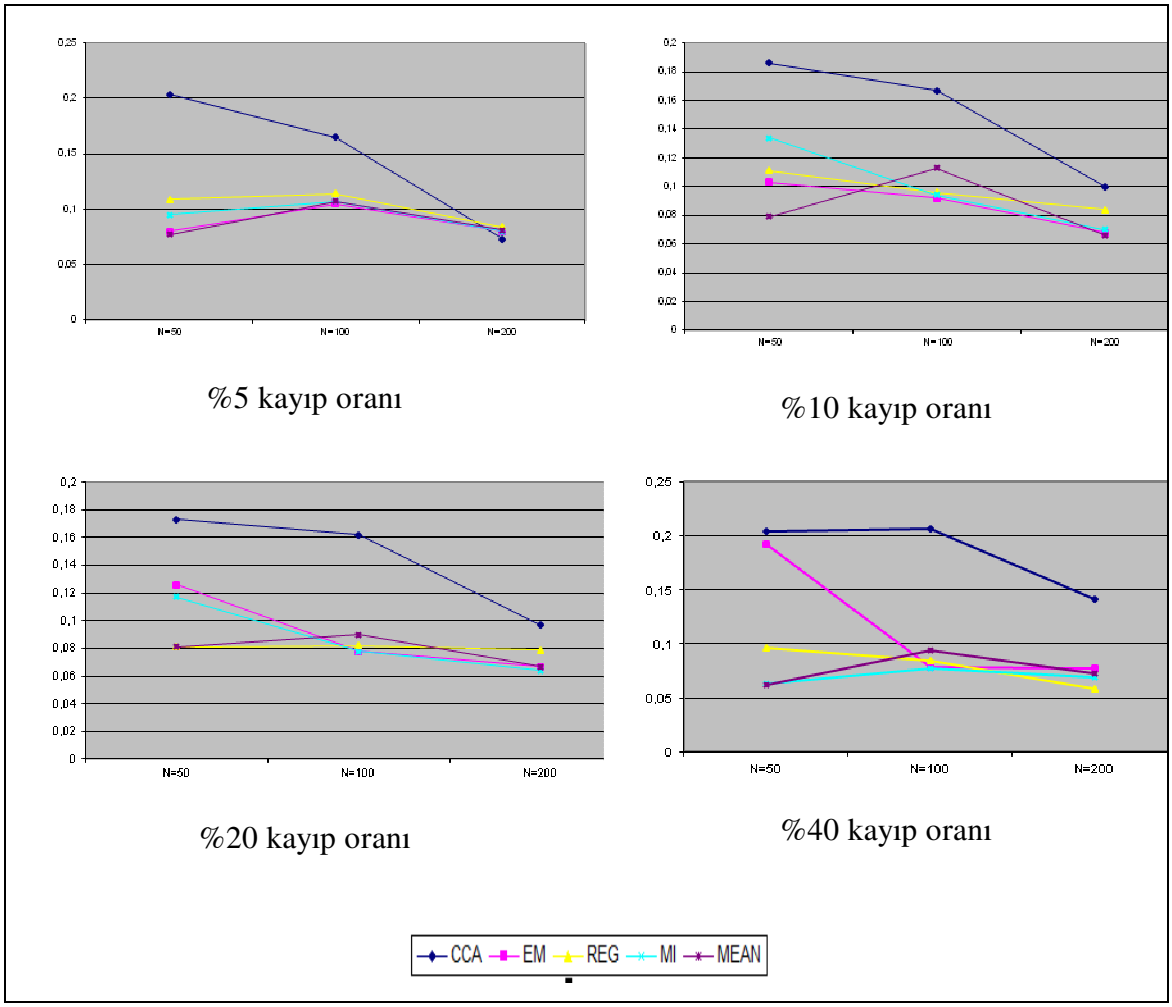
Şekil 4.4. Farklı örnek genişlikleri ve farklı kayıp oranları için standart hata grafikleri



Şekil 4.4. (Devamı) Farklı örnek genişlikleri ve farklı kayıp oranları için standart hata grafikleri

Şekil 4.4'de farklı kayıp oranlı ve örnek genişlikli verilerden elde edilen standart hataların grafikleri incelendiğinde CCA yönteminden elde edilen standart hatalar bütün örnek genişliklerinde diğer yöntemlere göre büyük bulunmuştur. Örnek genişliği arttıkça yöntemlerin uygulanması sonucu elde edilen standart hatalar birbirine ve gerçek değere daha çok yaklaşmaktadır.

Örnek genişliği arttıkça kayıp veri tamamlama yöntemlerinin benzerliklerinin incelenmesi amacıyla bütün değişkenler için farklı kayıp oranları elde edildi. Bütün değişkenler için ayrı ayrı yapılan grafikler incelendiğinde benzer sonuçlar elde edildiğinden, değişkenlerden biri için elde edilen grafikler Şekil 4.5'de verilmiştir.



Şekil 4.5. Farklı örnek genişlikli verilerde kayıp veri tamamlama yöntemlerinin grafiksel gösterimi

Şekil 4.5'de grafikler incelendiğinde %5 kayıp oranında bütün örnek genişlikleri için kayıp veri tamamlama yöntemleri arasındaki fark küçüktür. Bu kayıp oranında CCA yönteminin sadece 200 birimlik örnekte diğer yöntemlere benzer sonuç verdiği görülmüştür. Ayrıca kayıp oranı arttıkça yöntemler arasındaki farkların büyüdüğü buna karşın örnek genişliği arttıkça yöntemlerin birbirine benzerliğinin arttığı tespit edildi.

Yapılan deneysel çalışma sonucu örnek genişliği az (N=50) ise CCA yönteminin kullanılması veri sayısını daha da azaltmakta ve parametrelerin gerçek değerlerine yakın değerler bulunamamaktadır. Bu nedenle uygulaması çok kolay olan CCA yöntemi uygulanacaksa örnek genişliğinin çok olması (N>100) ve kayıp değerli gözlem oranının az (< 0,20) olması gerekir.

Örnek genişliği az olduğunda ortalama, regresyon, beklenen maksimizasyon ve çoklu değer atama yöntemleri kullanılabilir. Bu durumda kayıp değer oranının %20'den az olması ile doğru sonuç elde edilebilir. Örnek genişliği 50 ve kayıp değer oranı %20'den az olduğunda gerçek değere en yakın sonuçlar REG, MI ve MEAN yöntemleri ile elde edildiği için bu yöntemlerin kullanılması önerilmektedir.

Büyük örnek genişliklerinde (N>100) ve kayıp değer oranı %20 ve altında genel olarak tam veri ile elde edilen sonuçlara en yakın değerleri MI, EM ve REG yöntemleri vermektedir. Kayıp oranı %40 olması durumunda bütün örnek genişliklerinde çoklu değer atama yöntemi hariç diğer atama yöntemleri tam veri ile elde edilenlere her zaman yakın değerler bulamadığından tutarsızlık göstermişlerdir. Bu durumda bile çoklu değer atamanın sonuçları diğer yöntemlere göre nispeten daha iyidir.

Bu sonuçları göz önüne alarak her bir kayıp değere sadece bir değer atayan ortalama değer atama, regresyon değer ataması, beklenen maksimizasyon yöntemlerine göre birden fazla değer atayan çoklu değer atama yönteminin kullanılması önerilebilir.

Ayrıca örnek genişliği arttıkça kayıp değer tamamlama yöntemleriyle elde edilen sonuçlar ile tam veri kullanılarak elde edilen sonuçlar arasındaki fark oldukça azalmaktadır.

4.2. Uygulama II

Uygulamanın ikinci bölümünde gerçek bir veri seti olan ve Ondokuz Mayıs Üniversitesi Tıp Fakültesi'nden alınan akciğer kanserli hastaların sağkalım süresine ilişkin veri seti kullanıldı. Bu veri seti kullanılarak Bayesci Cox Regresyon (BCR) analizi yapıldı ve sonuçları Cox Regresyon (CR) analizi ile elde edilen sonuçlarla karşılaştırıldı. BCR bilgilendirici önselli ve bilgilendirici olmayan önselli olmak üzere iki şekilde uygulandı.

4.2.1. Sağkalım Verisinde Kayıp Değer Olmaması Durumunda (Tam) Cox ve Bayesci Cox Regresyon Yöntemlerinin Karşılaştırılması

BCR ve CR analizlerinin performanslarını değerlendirmek amacıyla Ondokuz Mayıs Üniversitesi Tıp Fakültesinden alınan akciğer kanserli sağkalım verisi kullanıldı.

Akciğer kanserli hastaların sağkalım verisinde 174 hastanın 136 tanesi ölmüş, 36 hasta ise çalışma sonunda hala yaşamaktadır. Yani sansürlü verilerdir. Akciğer kanserinde sağkalım süresini etkileyen risk faktörleri olarak, hastadaki kilo kaybı, hemoglobin değeri, platelet, protein, albümin, LDH değeri, ECOG performans skalasına göre durumu, histolojik tip ve tümör çapı değişkenleri ele alınmıştır. Bu değişkenlerden hemoglobin, platelet, protein, albümin ve LDH, tümor çapı değerleri sürekli değişkenler kilo kaybı, histolojik tip ve ECOG değişkenleri kesikli değişkenler olarak ele alındı.

Kilo kaybı : 0-Kilo kaybı yok .
1- Kilo kaybı %10'un üzerinde.

Histolojik tip: 1-Küçük hücreli akciğer kanseri(KHAK).
2-Küçük hücre dışı akciğer kanseri (KHDAK).

ECOG : 1- Hastalık öncesi işlerini yapabilir veya ayakta hafif işler yapabilir.
2- Hasta zamanının %50'den daha azını yatakta geçirir.
3-Hasta kendi bakımını yapmakta zorlanır ve gündüz zamanının %50'den fazlasını yatakta geçirir.
4-Hasta yatalaktır yani kendi bakımını yapamaz ve tam olarak sandalyeye ya da yatağa bağlıdır.

Akciğer kanserli hastalarda sağkalım süresini etkileyen önemli faktörleri belirlemek için Cox Regresyon (CR) modeli, bilgilendirici olmayan önselli Bayesci Cox Regresyon (BOÖBCR) modeli, bilgilendirici önselli Bayesci Cox Regresyon (BÖBCR) modelleri kullanıldı. Bilgilendirici önseller daha önce yapılmış benzer çalışmalardan elde edildi.

4.2.1.1. Cox Regresyon Analizi

Akciğer kanserli hastalarda sağkalım süresini etkileyen önemli faktörleri belirlemek amacıyla uygulanan CR sonuçları Çizelge 4.4'de verilmiştir.

Çizelge 4.4. Akciğer kanserli hastaların sağkalım sürelerine ilişkin Cox regresyon analizi sonuçları

Parametre	SD	Parametre Tahmini	Standart Hata	Khi-Kare	P>Khi-Kare	Hazard Oranı
Kilokaybi	1	-0.12879	0.22864	0.3173	0.5732	0.879
Hemoglbn	**1**	**0.18904**	**0.06539**	**8.3575**	**0.0038**	**1.208**
Platelet	1	8.31463E-7	6.98415E-7	1.4173	0.2338	1.000
Protein	**1**	**-0.36053**	**0.13389**	**7.2504**	**0.0071**	**0.697**
Albumin	1	-0.11950	0.18784	0.4047	0.5247	0.887
LDH	**1**	**0.0002177**	**0.0000734**	**8.7966**	**0.0030**	**1.000**
ECOG	**1**	**1.08495**	**0.14909**	**52.9598**	**<.0001**	**2.959**
Histolojiktip	1	-0.23309	0.21210	1.2077	0.2718	0.792
Tumorcapi	1	0.07685	0.04485	2.9357	0.0866	1.080

Çizelge 4.4'e göre sağkalım süresi üzerinde kilokaybı, platelet seviyesi, albümin, histolojik tip ve tümör çapı değişkenlerinin önemli etkilerinin olmadığı %95 güvenle bulunmuştur. Buna karşın hastanın hemoglobin seviyesi, protein değeri LDH, ECOG performans durumu sağkalım süresi üzerinde önemli etkiye sahiptir ($p<0,05$).

Elde edilen sonuçlara göre hemoglobin seviyesi normal seviyeden uzaklaştıkça hastanın ölüm riskinin 1,208 kat arttığı, protein seviyesi düşük seviyeden normal seviyeye yükseldikçe riskin 0,697 kat azaldığı, LDH seviyesi arttıkça riskin 1 kat arttığı, ECOG seviyesi için tedavisini ayakta devam ettirenlere göre diğerlerinin 2,959 kat risk taşıdığı söylenebilir.

4.2.1.2. Bilgilendirici Önselli Bayesci Cox Regresyon (BÖBCR)

Akciğer kanseri için daha önce yapılmış çalışmalar incelendi ve analiz için kullanılan değişkenlerden hemoglobin, protein, albümin, LDH ve tümör çapı için önsel bilgiler elde edildi. Önsel bilgiye ulaşılamayan değişkenler için normal önsel kullanıldı. Değişkenler için elde edilen önsel bilgiler Çizelge 4.5'de verilmiştir.

Çizelge 4.5. Akciğer kanserli hastalara ilişkin ulaşılabilen önsel bilgiler

Değişkenler	Hazard oranı	HO için Güven aralığı	Kaynak
Hemoglobin	2,2	1,1 - 4,5	Mohan ve ark. (2006)
Protein	1,3	0,6 – 2,9	Mohan ve ark. (2006)
LDH	0,8	0,4- 1,6	Mohan ve ark. (2006)
Albümin	0,95	0,92-0,98	Lam ve ark.(2007)
Tümör çapı	1,08	0,95-1,25	Abreu ve ark.(2003)

Çizelge 4.5'de verilen bilgilendirici önsel bilgilerin kullanılmasıyla elde edilen BCR model sonuçları Çizelge 4.6'da verilmiştir. HPD (highest posterior density) aralığının sıfırı kapsaması o değişkenin istatistiksel olarak önemli olmadığını gösterir.

Çizelge 4.6. Akciğer kanserli hastaların sağkalım sürelerine ilişkin BÖBCR analizi sonuçları

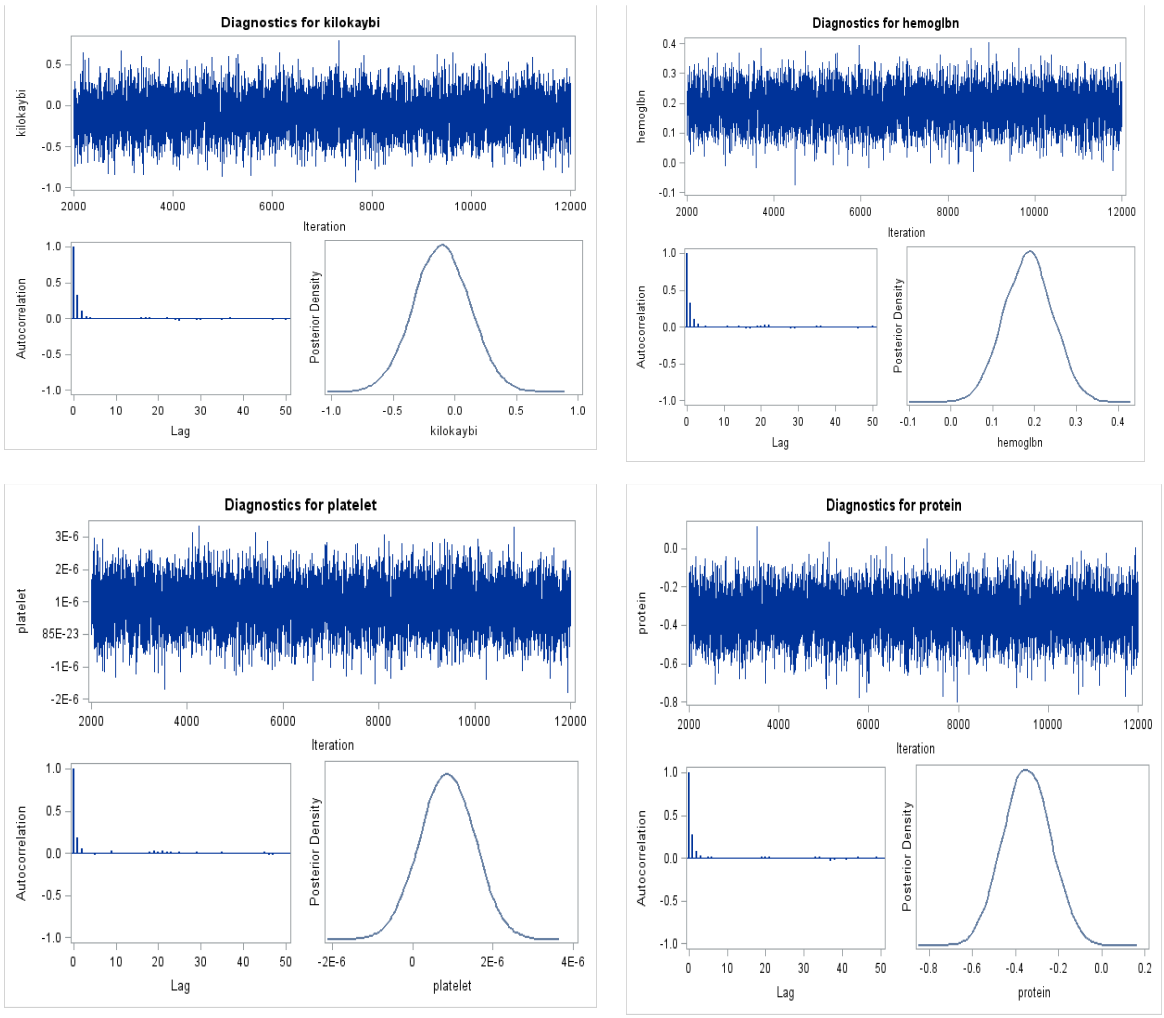
Değişkenler	Parametre Tahminleri	Standart Hatalar	HPD Aralığı		Hazard Oranı
Kilokaybi	-0.1047	0.2255	-0.5598	0.3244	0,901
Hemoglbn	**0.1856**	**0.0580**	**0.0718**	**0.2968**	**1,204**
Platelet	8.659E-7	6.776E-7	-4.68E-7	2.173E-6	1,000
Protein	**-0.3467**	**0.1104**	**-0.5625**	**-0.1343**	**0,707**
Albumin	**-0.0526**	**0.0157**	**-0.0846**	**-0.0233**	**0,949**
LDH	**0.000213**	**0.000081**	**0.000049**	**0.000367**	**1,000**
ECOG	**1.1008**	**0.1480**	**0.8097**	**1.3852**	**3,006**
Histolojiktip	-0.2425	0.2111	-0.6562	0.1757	0,785
Tümör çapi	**0.0761**	**0.0368**	**0.00586**	**0.1494**	**1,079**

Çizelge 4.6'da verilen sonuçlar incelendiğinde kilo kaybı, platelet, histolojik tip değişkenlerinin akciğer kanseri hastalarında sağkalım sürelerini etkilemediği, buna karşın hemoglobin seviyesi, protein, albümin, LDH, ECOG ve tümör çapı sağkalım süresi üzerinde önemli bir etkiye sahiptir ($p<0,05$). Hemoglobin seviyesi normal seviyeden uzaklaştıkça ölüm riskinin 1,204 kat arttığı görülmektedir.

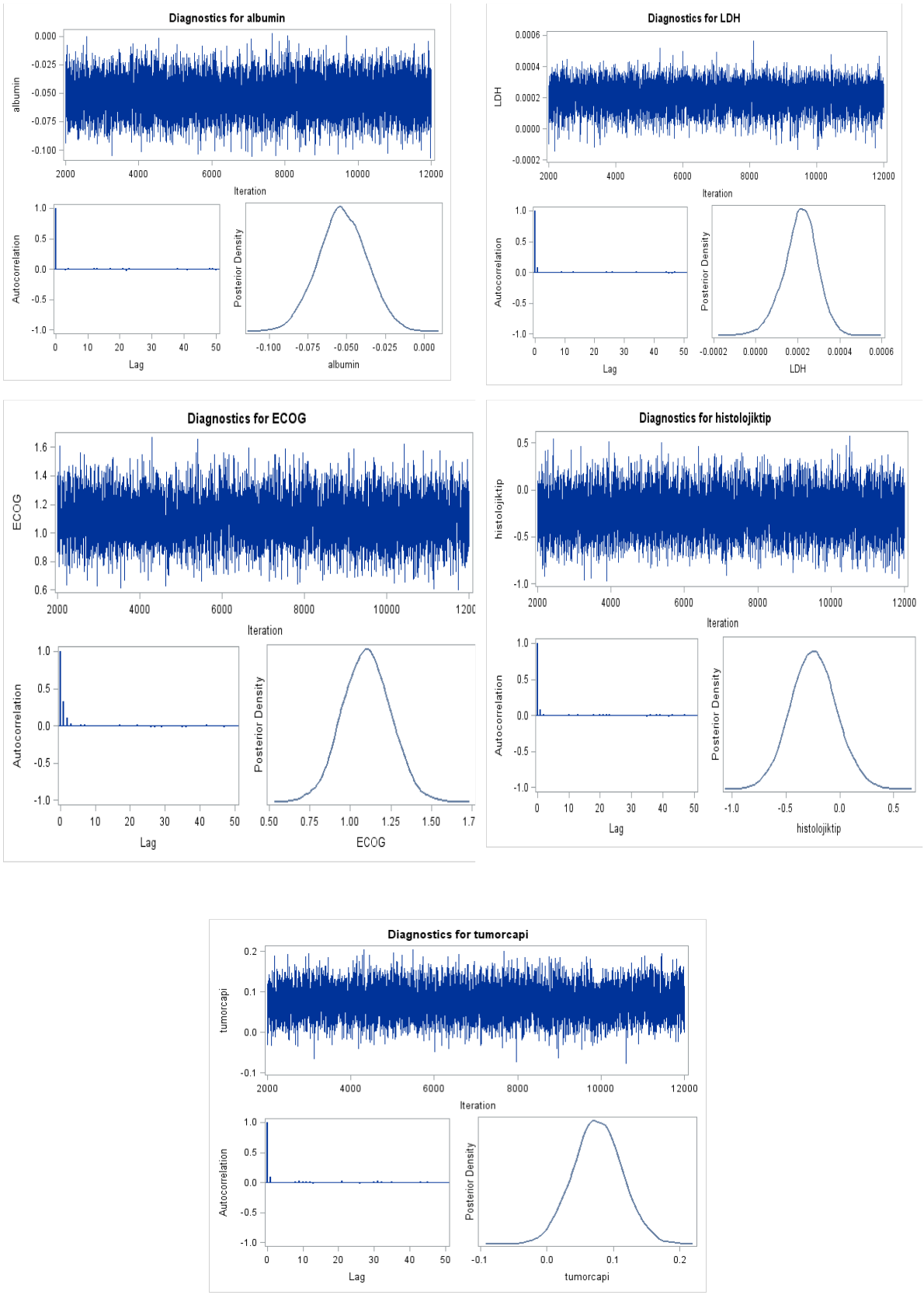
Protein ve albümin seviyesinin düşük seviyeden normal seviyeye yükselmesi hastanın taşıdığı ölüm riskini sırasıyla 0,707 ve 0,949 kat azaltmaktadır. LDH seviyesi arttıkça riskin 1 kat arttığı söylenebilir.

ECOG performans ölçüsü hastaların tedavisini ayakta devam ettirenlere göre 3,006 kat daha fazla ölüm riski taşır. Tümör çapı arttıkça ölüm riski 1,079 kat artmaktadır.

Ulaşılabilen önseller yardımıyla akciğer verisine uygulanan BCR analizinde Markov zincirinin sonsal dağılıma yakınsamasını değerlendirmek amacıyla iz grafikleri, Geweke testi sonuçları, sonsal otokorelasyonlar Şekil 4.6'da verildiği gibidir.



Şekil 4.6. Akciğer kanserli hastaların sağkalım süresine ilişkin BÖBCR için MCMC'nin yakınsamasının iz grafikleri



Şekil 4.6. (Devamı) Akciğer kanserli hastaların sağkalım süresine ilişkin BÖBCR için MCMC'nin yakınsamasının iz grafikleri

Grafikler incelendiğinde bütün parametreler için Markov zincirinin yakınsadığı Şekil 4.6'da görülmektedir.

Akciğer kanserli hastaların sağkalım süresine ilişkin BÖBCR için MCMC'nin yakınsamasının Geweke test sonuçları Çizelge 4.7'de verilmiştir.

Çizelge 4.7. Akciğer kanserli hastaların sağkalım süresine ilişkin BÖBCR için MCMC'nin yakınsamasının Geweke test sonuçları

Geweke Testi		
Parametre	**z**	**Pr > |z|**
kilokaybi	-1.3319	0.1829
hemoglbn	0.6598	0.5094
platelet	1.0220	0.1025
protein	0.4811	0.6305
albumin	-0.0459	0.9634
LDH	1.4853	0.1375
ECOG	0.5553	0.5787
histolojiktip	-0.7089	0.4784
tumorcapi	-0.9444	0.3450

Geweke test sonuçlarına göre her bir değişken için zincir sonsal dağılıma yakınsamaktadır (P>0,05).

Çizelge 4.8. Akciğer kanserli hastaların sağkalım süresine ilişkin BÖBCR için MCMC'nin yakınsamasının sonsal otokorelasyonlarla değerlendirilmesi

Sonsal Otokorelasyonlar (Posterior Autocorrelations)				
Parametre	**Lag 1**	**Lag 5**	**Lag 10**	**Lag 50**
kilokaybi	0.3203	-0.0006	-0.0067	-0.0205
hemoglbn	0.3297	0.0082	0.0069	0.0102
platelet	0.1848	-0.0190	0.0024	-0.0011
protein	0.2742	0.0135	-0.0048	-0.0026
albumin	-0.0066	-0.0014	-0.0027	-0.0219
LDH	0.0824	-0.0059	0.0030	-0.0073
ECOG	0.3268	0.0004	-0.0162	-0.0085
histolojiktip	0.0828	-0.0134	0.0098	0.0045
tumorcapi	0.0957	-0.0060	0.0174	0.0023

Çizelge 4.8'de verilen sonsal otokorelasyonlar incelendiğinde gecikmeler arasındaki korelasyonun küçük çıkması yakınsamanın sağlandığını gösterir.

4.2.1.3. Bilgilendirici Olmayan Önselli Bayesci Cox Regresyon (BOÖBCR)

Akciğer kanseri hastalarında sağkalım süresini etkileyen önemli değişkenleri belirlemek amacıyla uygulanan BOÖBCR modeli sonuçları Çizelge 4.9'da verilmiştir. Değişkenlerle ilgili önsel bilgi olmadığında bilgilendirici olmayan düzgün önsel kullanılmaktadır.

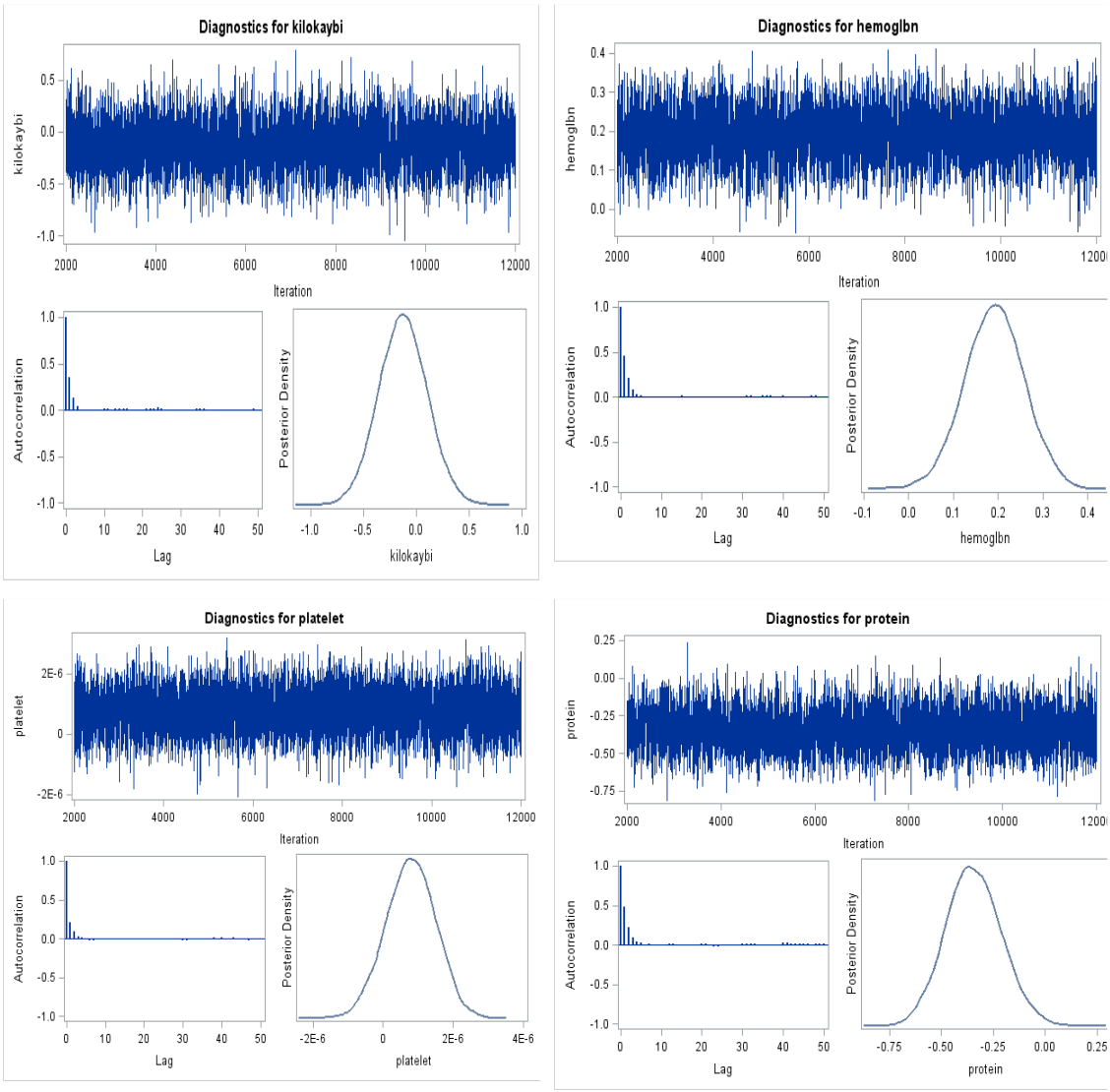
Çizelge 4.9. Akciğer kanserli hastaların sağkalım sürelerine ilişkin BOÖBCR analizi sonuçları

Değişkenler	**Parametre Tahminleri**	**Standart Hatalar**	**HPD Aralığı**		**Hazard Oranı**
Kilokaybi	-0.1251	0.2302	-0.5935	0.3110	0,882
Hemoglbn	**0.1914**	**0.0662**	**0.0630**	**0.3222**	**1,211**
Platelet	7.877E-7	7.042E-7	-5.85E-7	2.135E-6	1,000
Protein	**-0.3429**	**0.1331**	**-0.6095**	**-0.0863**	**0,710**
Albumin	-0.1481	0.1889	-0.5147	0.2154	0,862
LDH	**0.000205**	**0.000081**	**0.000040**	**0.000356**	**1,000**
ECOG	**1.0909**	**0.1498**	**0.7843**	**1.3745**	**2,976**
Histolojiktip	-0.2270	0.2140	-0.6264	0.2027	0,797
Tümör çapi	0.0761	0.0449	-0.00929	0.1670	1,079

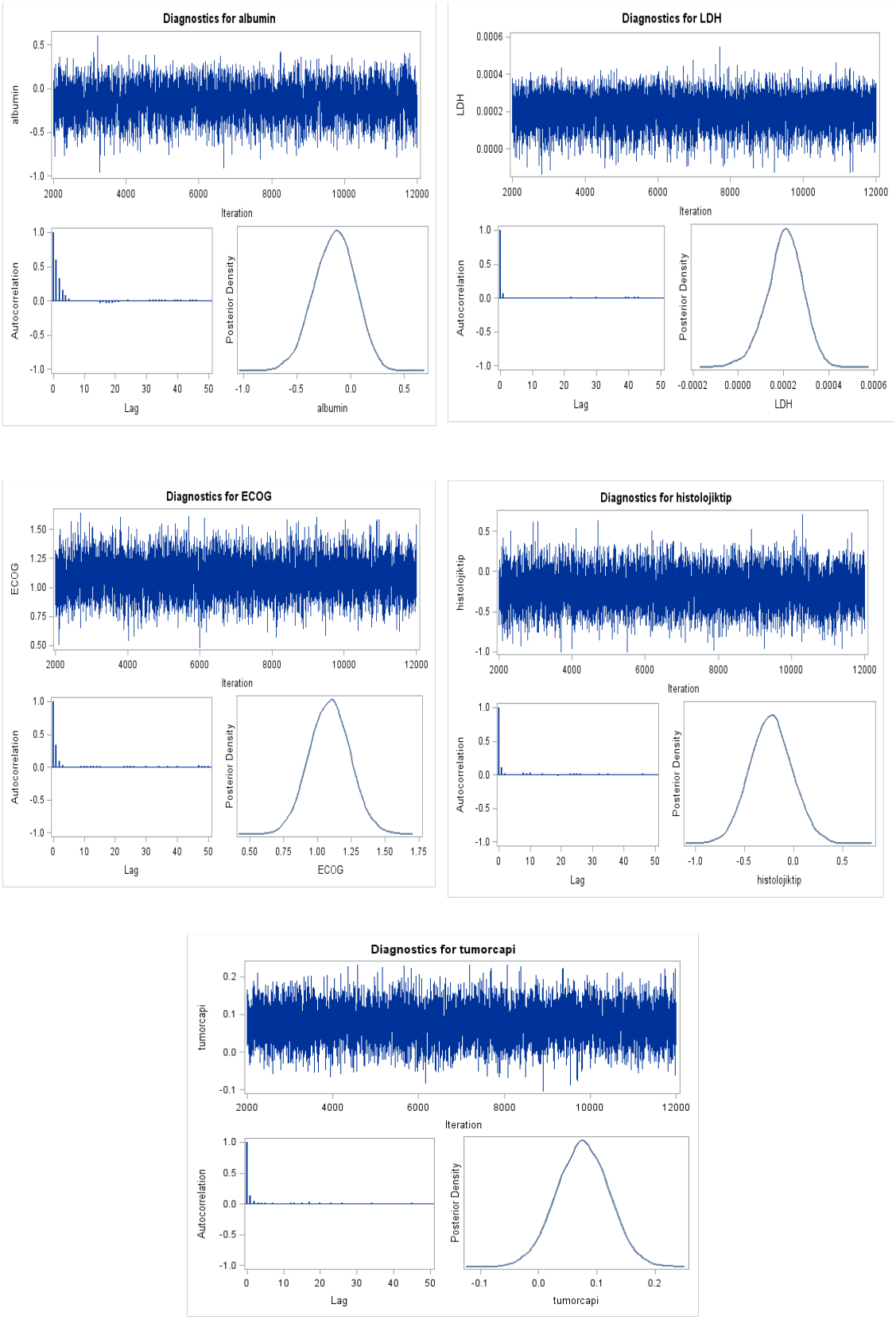
Çizelge 4.9'a göre sağkalım süresi üzerinde kilokaybı, platelet seviyesi, albümin, histolojik tip ve tümör çapı değişkenlerinin önemli bir etkilerinin olmadığı %95 güvenle bulunmuştur. Buna karşın hastanın hemoglobin seviyesi, protein değeri LDH, ECOG performans durumu sağkalım süresi üzerinde önemli etkiye sahip değişkenlerdir.

Elde edilen sonuçlara göre hemoglobin seviyesi normal seviyeden uzaklaştıkça hastanın ölüm riskinin 1,211 kat arttığı, protein seviyesi düşük seviyeden normal seviyeye yükseldikçe riskin 0,710 kat azaldığı, LDH seviyesi arttıkça riskin 1 kat arttığı, ECOG seviyesi için tedavisini ayakta devam ettirenlere göre diğerlerinin 2,976 kat risk taşıdığı tespit edilmiştir.

BOÖBCR analizinde Markov zincirinin yakınsamasını değerlendirmek amacıyla elde edilen iz grafikleri, Geweke test sonuçları ve otokorelasyon değerleri Şekil 4.7'de verilmiştir.



Şekil 4.7. Akciğer kanserli hastaların sağkalım süresine ilişkin BOÖBCR için MCMC'nin yakınsamasının iz grafikleri



Şekil 4.7. (Devamı) Akciğer kanserli hastaların sağkalım süresine ilişkin BOÖBCR için MCMC'nin yakınsamasının iz grafikleri

Şekil 4.7'deki grafikler incelendiğinde bütün değişkenler için Markov zincirinin sonsal dağılıma yakınsadığı görülmektedir.

Çizelge 4.10. Akciğer kanserli hastaların sağkalım süresine ilişkin BOÖBCR için MCMC'nin yakınsamasının Geweke test sonuçları

Geweke Testi		
Parametre	**z**	**Pr > |z|**
kilokaybi	-0.0189	0.9849
hemoglbn	-0.4833	0.6289
platelet	-0.5327	0.4968
protein	0.9662	0.3339
albumin	-0.5785	0.5629
LDH	-0.7398	0.4594
ECOG	-0.9143	0.3606
histolojiktip	1.3356	0.1817
tumorcapi	0.5594	0.5759

BOÖBCR analizinin yakınsamasını değerlendirmek amacıyla Geweke testi uygulandı ve bütün değişkenlerde Markov zincirinin yakınsadığı Çizelge 4.10'da görülmektedir ($p>0,05$).

Çizelge 4.11. Akciğer kanserli hastaların sağkalım süresine ilişkin BOÖBCR için MCMC'nin yakınsamasının sonsal otokorelasyonlarla değerlendirilmesi

Sonsal Otokorelasyonlar (Posterior Autocorrelations)				
Parametre	**Lag 1**	**Lag 5**	**Lag 10**	**Lag 50**
kilokaybi	0.3451	-0.0084	0.0130	-0.0023
hemoglbn	0.4593	0.0104	-0.0045	-0.0115
platelet	0.2091	-0.0104	-0.0141	-0.0146
protein	0.4757	0.0242	-0.0019	0.0174
albumin	0.5969	0.0279	-0.0146	0.0039
LDH	0.0653	-0.0043	0.0021	-0.0032
ECOG	0.3437	-0.0076	0.0148	0.0106
histolojiktip	0.0987	-0.0102	0.0292	0.0043
tumorcapi	0.1319	0.0136	0.0045	-0.0112

Çizelge 4.11'de verilen sonsal otokorelasyonlar incelendiğinde bütün gecikme değerleri için korelasyonlar küçük bulunduğundan, Markov zincirinin sonsal dağılıma yakınsaması sağlanmıştır.

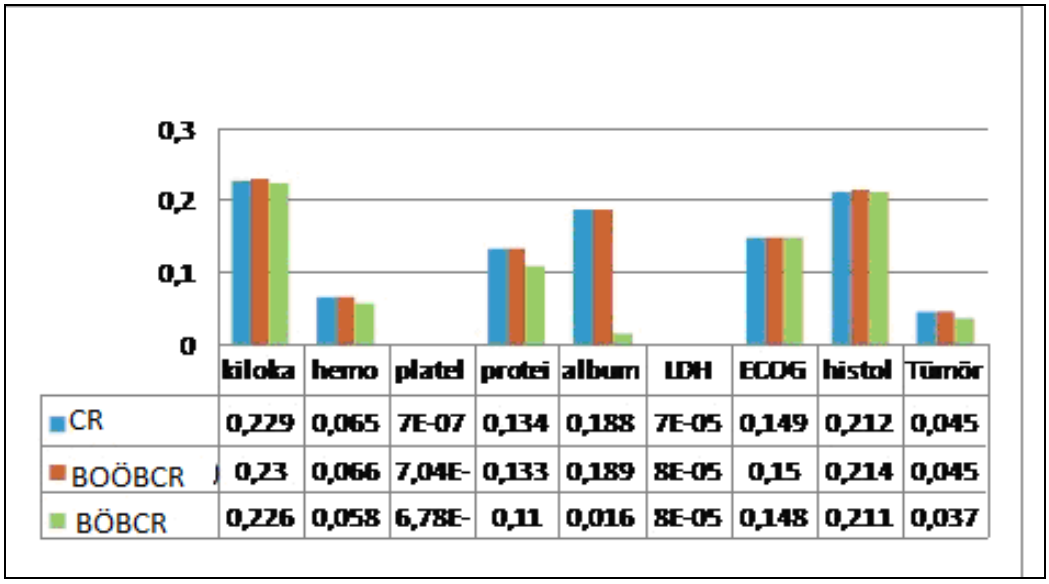
Akciğer kanseri verisine CR, BÖBCR ve BOÖBCR ayrı ayrı uygulanmış ve uygulanan yöntemlerin karşılaştırılması amacıyla standart hatalar ve modele uygunluk kriterleri dikkate alınmıştır. CR için AIC değeri Bayesci Cox Regresyon (BCR) için ise DIC değeri hesaplandı. Spiegelhalter ve ark. (1998) çalışmalarında verdiği (3.25) eşitliği yardımıyla BCR için AIC yaklaşık olarak hesaplandı. Her bir yöntem için modele uygunluk kriterleri Çizelge 4.12'de verilmiştir.

Çizelge 4.12. Akciğer kanserli hastaların sağkalım süresine ilişkin her bir yöntem için uygunluk kriterleri

	AIC	**DIC** (p_D)
CR	1064.251	
BOÖBCR	≈1064.360	1064.412 (9.052)
BÖBCR	≈1063.186	1061.785 (7.599)

Çizelge 4.12 incelendiğinde CR ile benzer sonuçlar veren BOÖBCR'nin DIC değeri (1064.412) BÖBCR'nun DIC değerinden (1061.785) daha büyüktür. Bu durumda BÖBCR modeli daha iyi bir model olduğu sonucuna varılır. Ayrıca BCR için yaklaşık olarak hesaplanan AIC değerleri dikkate alındığında en küçük AIC değerini yine BÖBCR vermiştir.

Çalışmada her bir analiz yönteminden elde edilen standart hatalar ayrı ayrı incelenmiş ve sonuçları Şekil 4.8'de sunulmuştur.



	kiloka	hemo	platel	protei	album	LDH	ECOG	histol	Tümör
CR	0,229	0,065	7E-07	0,134	0,188	7E-05	0,149	0,212	0,045
BOÖBCR	0,23	0,066	7,04E-	0,133	0,189	8E-05	0,15	0,214	0,045
BÖBCR	0,226	0,058	6,78E-	0,11	0,016	8E-05	0,148	0,211	0,037

Şekil 4.8. CR, BOÖBCR, BÖBCR yöntemleri tarafından bulunan her bir değişkenin standart hatalarının grafiksel gösterimi

Standart hatalar bakımından elde edilen Şekil 4.8 incelendiğinde en küçük standart hataları BÖBCR yöntemi vermiştir. Her iki kriter dikkate alındığında verilerde kayıp gözlem olmadığında ve değişkenlerle ilgili önsel bilgi olduğunda BÖBCR uygulanması önerilebilir. BOÖBCR ile CR benzer sonuçlar verdiğinden, BOÖBCR bir üstünlüğü olmamaktadır. Ayrıca CR sonucu bulunan standart hatalar düzgün önselli Bayesci Cox Regresyona (BOÖBCR) göre biraz daha küçük bulunduğundan, daha hassas olabileceği söylenebilir.

4.4.Uygulama III

Uygulamanın üçüncü bölümünde veride kayıp değer olması durumunda BCR performansını değerlendirmek amacıyla akciğer kanserli veri seti kullanılarak, MAR varsayımına göre %20 kayıp değerli gözlem içeren yeni bir kayıp değerli veri seti oluşturuldu. Araştırmacılar verilerinde kayıp değer olması durumunda en az bir kayıp değerli değişkene sahip olan gözlemi işlem dışı tutarak eksiksiz veri analizi yöntemini (Complete Case Analysis-CCA) sıklıkla kullanmaktadır. Bu sebeple kayıp değer içeren akciğer kanserli hastaların sağkalım verisi için çok sık kullanılan CCA yöntemi ile BCR analiz yönteminin performansı CR analizi ile karşılaştırıldı.

4.3.1. Sağkalım Verisinde Kayıp Değer Olması Durumunda Cox Regresyon (CR) ve Bayesci Cox Regresyon (BCR) Yöntemlerinin Karşılaştırılması

Veride kayıp değer olması durumunda BCR performansını değerlendirmek amacıyla akciğer kanserli gerçek veri seti kullanılarak, MAR varsayımına göre %20 kayıp değerli gözlem içeren yeni kayıp değerli veri seti oluşturuldu. Araştırmacılar verilerinde kayıp değer olması durumunda en az bir kayıp değerli değişkene sahip olan gözlemi işlem dışı tutarak eksiksiz veri analizi yöntemini (Complete Case Analysis-CCA) kullanmaktadır. Bu sebeple kayıp değer içeren akciğer kanserli hastaların sağkalım verisi için çok sık kullanılan CCA yöntemi ile BCR analiz yönteminin performansı CR analizi ile karşılaştırıldı. Ayrıca kayıp değerli veriye CCA ile birlikte uygulanan CR ve BCR analizlerinden elde edilen regresyon katsayıları tam veriden elde edilen katsayılarla karşılaştırılarak hangi yöntemin gerçek verinin sonucuna daha yakın sonuç verdiği araştırıldı.

4.3.1.1. Kayıp Değerli Veri Seti İçin Cox Regresyon Analizi

Akciğer kanserli 174 hastanın 35 tanesi en az bir kayıp değer içeren değişkene sahip olması nedeniyle analizden çıkarılarak, CR uygulandı. Kayıp değerli veriye sahip akciğer kanserli hastalarda sağkalım süresini etkileyen önemli faktörleri belirlemek amacıyla uygulanan CR sonuçları Çizelge 4.13'de verilmiştir.

Çizelge 4.13. Kayıp değerli akciğer kanserli hastaların sağkalım süresine ilişkin CR sonuçları

Parametre	SD	Parametre Tahmini	Standart Hata	Ki-Kare	Pr > Ki-Kar.	Hazard Oranı
kilokaybi	1	-0.09961	0.24526	0.1650	0.6846	0.905
hemoglbn	**1**	**0.17770**	**0.07674**	**5.3626**	**0.0206**	**1.194**
platelet	1	1.18528E-6	7.83118E-7	2.2908	0.1301	1.000
protein	**1**	**-0.36677**	**0.13847**	**7.0156**	**0.0081**	**0.693**
albumin	1	-0.05731	0.20173	0.0807	0.7763	0.944
LDH	**1**	**0.0006359**	**0.0003086**	**4.2454**	**0.0394**	**1.001**
ECOG	**1**	**1.01882**	**0.16796**	**36.7927**	**<.0001**	**2.770**
histolojiktip	1	-0.15452	0.24639	0.3933	0.5306	0.857
tumorcapi	1	0.05350	0.05407	0.9788	0.3225	1.055

Çizelge 4.13'e göre sağkalım süresi üzerinde kilokaybı, platelet seviyesi, albümin, histolojik tip ve tümör çapı değişkenlerinin önemli bir etkilerinin olmadığı %95 güvenle bulunmuştur. Buna karşın hastanın hemeoglobin seviyesi, protein değeri LDH, ECOG performans durumu sağkalım süresi üzerinde önemli etkiye sahiptir.

Hemoglobin seviyesi normalden uzaklaştıkça hastanın ölüm riskinin 1,194 kat arttığı, protein seviyesi düşük seviyeden normal seviyeye yükseldikçe riskin 0,693 kat azaldığı, LDH seviyesi arttıkça riskin 1 kat arttığı, ECOG seviyesi için tedavisini ayakta devam ettirenlere göre diğerlerinin 2,770 kat risk taşıdığı belirlenmiştir.

4.3.1.2. Kayıp değerli Veri Seti İçin BOÖBCR Analizi

Akciğer kanseri hastalarında sağkalım süresini etkileyen önemli değişkenleri belirlemek amacıyla uygulanan BOÖBCR modeli sonuçları Çizelge 4.14'de verilmiştir. Değişkenlerle ilgili önsel bilgi olmadığında bilgilendirici olmayan düzgün önsel kullanıldı.

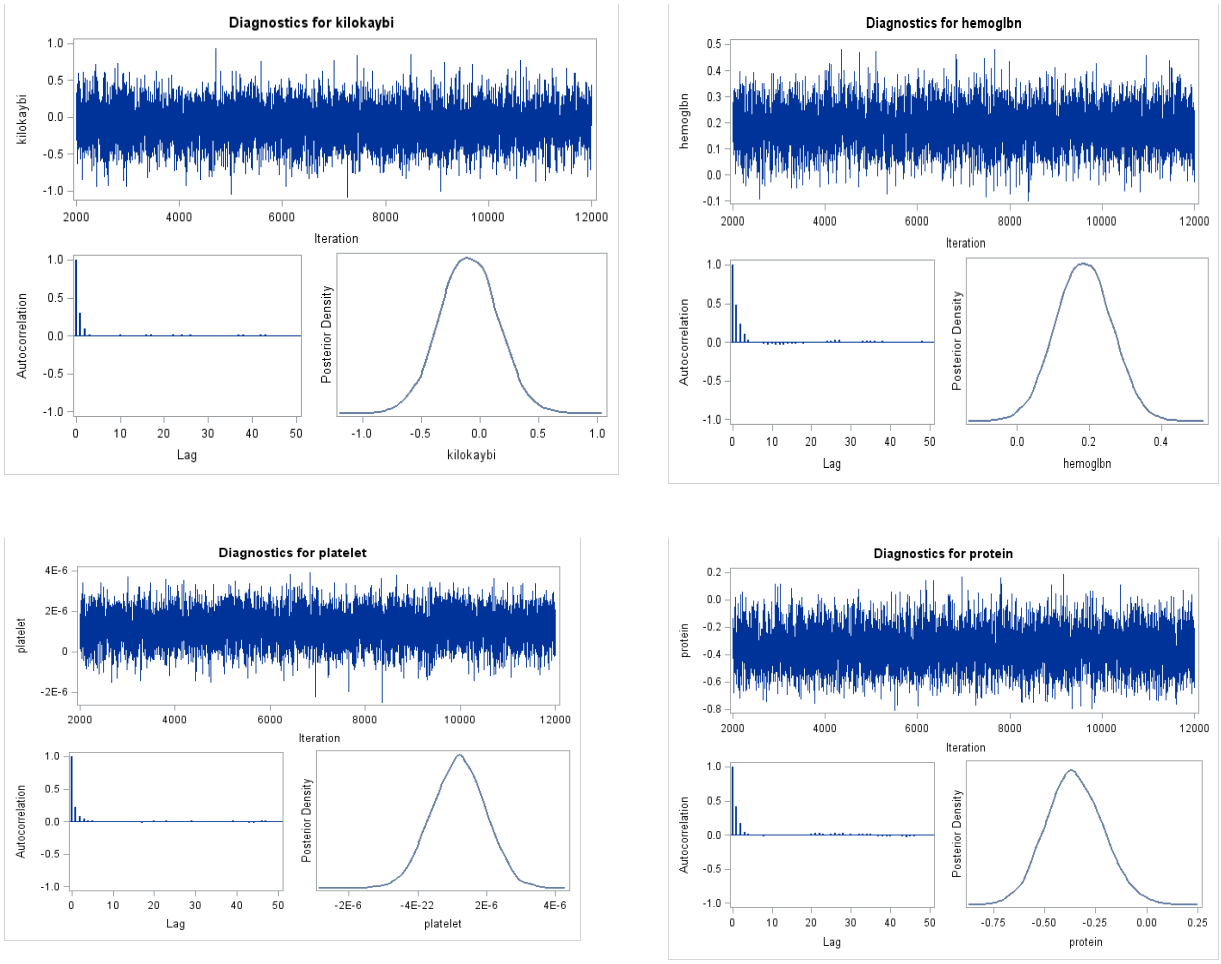
Çizelge 4.14. Kayıp değerli akciğer kanserli hastaların sağkalım süresine ilişkin BOÖBCR sonuçları

Değişkenler	**Parametre Tahminleri**	**Standart Hatalar**	**HPD aralığı**		**Hazard Oranı**
Kilokaybi	-0.0938	0.2463	-0.5796	0.3826	0,911
Hemoglbn	**0.1849**	**0.0770**	**0.0376**	**0.3357**	**1,203**
Platelet	1.19E-6	7.819E-7	-2.72E-7	2.741E-6	1,000
Protein	**-0.3524**	**0.1391**	**-0.6147**	**-0.0718**	**0,703**
Albumin	-0.0914	0.2068	-0.5001	0.3109	0,913
LDH	0.00058	0.00031	-0.00003	0.0012	1,001
ECOG	**1.0304**	**0.1686**	**0.6948**	**1.3592**	**2,802**
Histolojiktip	-0.1519	0.2490	-0.6380	0.3419	0,859
Tümör çapi	0.0500	0.0542	-0.0577	0.1547	1,051

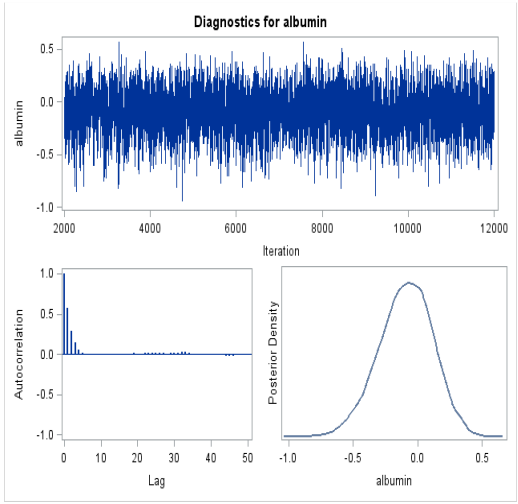
Çizelge 4.14'e göre sağkalım süresi üzerinde kilokaybı, platelet seviyesi, albümin, LDH, histolojik tip ve tümör çapı değişkenlerinin önemli etkilerinin olmadığı %95 güvenle bulunmuştur. Buna karşın hastanın hemoglobin seviyesi, protein değeri, ECOG performans durumu sağkalım süresi üzerinde önemli etkiye sahiptirler.

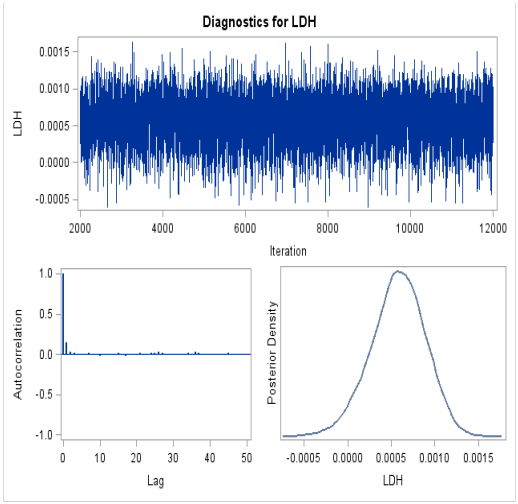
Hemoglobin seviyesi normal seviyeden uzaklaştıkça hastanın ölüm riskinin 1,203 kat arttığı, protein seviyesi düşük seviyeden normal seviyeye yükseldikçe riskin 0,703 kat azaldığı, ECOG seviyesi için tedavisini ayakta devam ettirenlere göre diğerlerinin 2,802 kat risk taşıdığı görülmektedir.

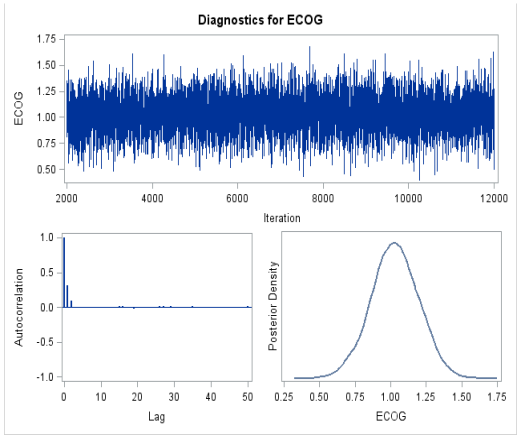
BOÖBCR analizinde Markov zincirinin yakınsamasını değerlendirmek amacıyla elde edilen iz grafikleri, Geweke test sonuçları ve otokorelasyon değerleri Şekil 4.9'da verilmiştir.

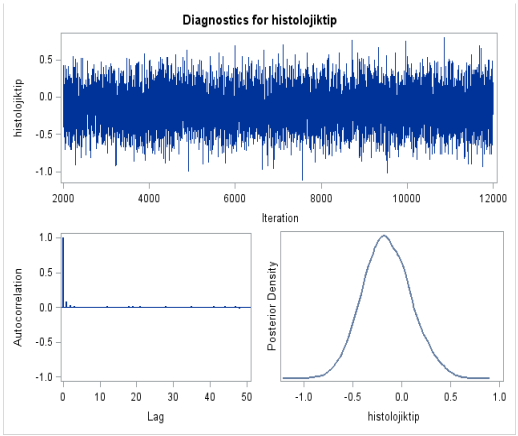


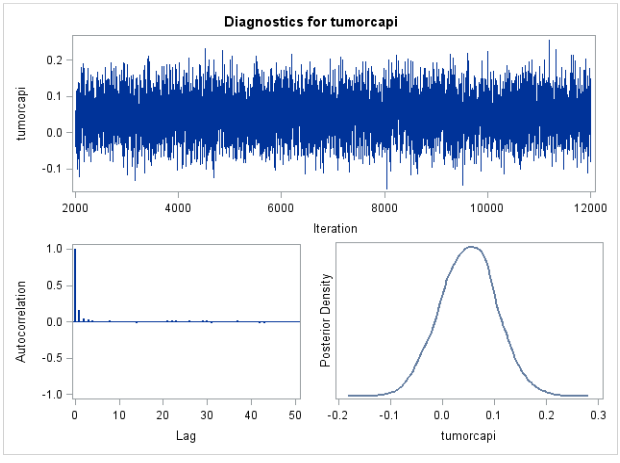
Şekil 4.9. Kayıp değerli akciğer kanserli hastaların sağkalım süresine ilişkin bilgilendirici olmayan önselli BCR için MCMC'nin yakınsamasının iz grafikleri











Şekil 4.9. (Devamı) Kayıp değerli akciğer kanserli hastaların sağkalım süresine ilişkin bilgilendirici olmayan önselli BCR için MCMC'nin yakınsamasının iz grafikleri

Şekil 4.9'da verilen grafikler incelendiğinde bütün değişkenler için Markov zincirinin sonsal dağılıma yakınsadığı görülmektedir.

Çizelge 4.15. Kayıp değerli akciğer kanserli hastaların sağkalım süresine ilişkin BOÖBCR için MCMC'nin yakınsamasının Geweke test sonuçları

Geweke Testi		
Parametre	**z**	**Pr > |z|**
kilokaybi	0.9974	0.3186
hemoglbn	-0.0261	0.9791
platelet	-0.5645	0.5123
protein	0.2878	0.7735
albumin	-0.3229	0.7468
LDH	1.7279	0.0840
ECOG	-1.9731	0.0510
histolojiktip	-0.1846	0.8535
tumorcapi	-0.6041	0.5458

Çizelge 4.15'te BOÖBCR analizinin yakınsamasını değerlendirmek amacıyla Geweke testi uygulandı ve bütün değişkenlerde Markov zincirinin yakınsadığı görülmektedir ($p>0,05$).

Çizelge 4.16. Kayıp değerli akciğer kanserli hastaların sağkalım süresine ilişkin BOÖBCR için MCMC'nin yakınsamasının sonsal otokorelasyonlarla değerlendirilmesi

Sonsal Otokorelasyonlar (Posterior Autocorrelations)				
Parametre	**Lag 1**	**Lag 5**	**Lag 10**	**Lag 50**
kilokaybi	0.3055	-0.0139	0.0127	-0.0041
hemoglbn	0.4848	0.0078	-0.0288	0.0070
platelet	0.2199	0.0102	-0.0051	0.0013
protein	0.4193	0.0015	-0.0097	-0.0025
albumin	0.5659	0.0194	-0.0062	-0.0010
LDH	0.1421	0.0044	-0.0232	-0.0011
ECOG	0.3099	-0.0095	-0.0031	0.0098
histolojiktip	0.0833	-0.0018	0.0059	-0.0105

Çizelge 4.16.'ya göre sonsal otokorelasyonlar incelendiğinde bütün gecikme değerleri için korelasyonlar küçük bulunduğundan, Markov zincirinin sonsal dağılıma yakınsaması sağlanmıştır.

4.3.3. Kayıp değerli Veri Seti İçin Bilgilendirici Önselli Bayesci Cox Regresyon (BÖBCR) Analizi

Akciğer kanseri için daha önce yapılmış çalışmalar incelendi ve analiz için kullanılan değişkenlerden hemoglobin, protein, albümin, LDH ve tümör çapı için önsel bilgiler elde edildi. Önsel bilgiye ulaşılamayan değişkenler için normal önsel kullanıldı. Değişkenler için elde edilen önsel bilgiler Çizelge 4.5'de verilmiştir.

Çizelge 4.17. Kayıp değerli akciğer kanserli hastaların sağkalım süresine ilişkin BÖBCR sonuçları

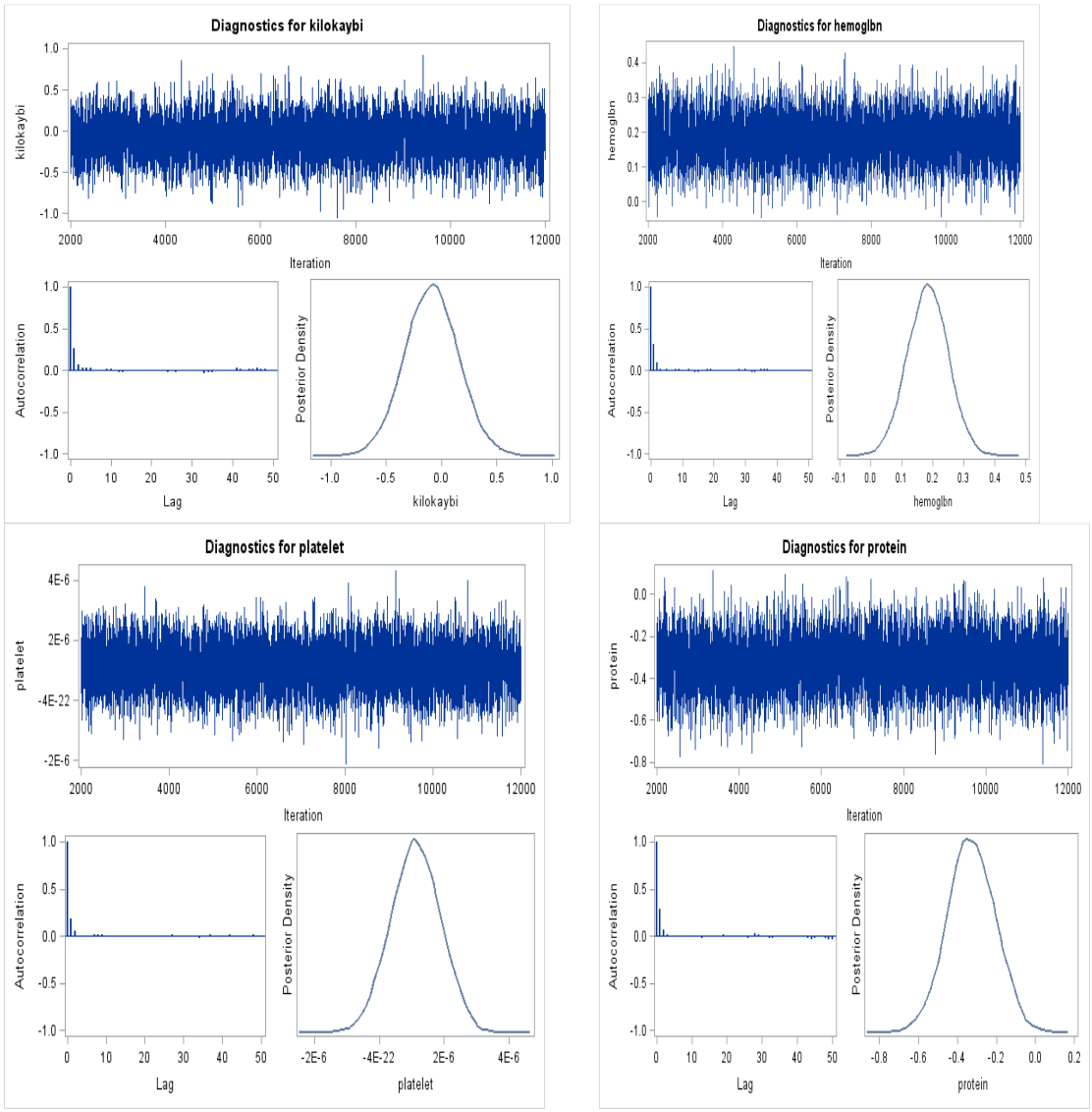
Değişkenler	**Parametre Tahminleri**	**Standart Hatalar**	**HPD Aralığı**		**Hazard Oranı**
kilokaybi	-0.0878	0.2424	-0.5728	0.3807	0,9160
hemoglbn	**0.1843**	**0.0653**	**0.0598**	**0.3156**	**1,2020**
platelet	1.131E-6	7.632E-7	-3.63E-7	2.593E-6	1,0000
protein	**-0.3268**	**0.1222**	**-0.5585**	**-0.0820**	**0,7210**
albumin	**-0.0520**	**0.0155**	**-0.0818**	**-0.0211**	**0,9490**
LDH	0.000593	0.000311	-0.00005	0.00116	1,0010
ECOG	**1.0433**	**0.1677**	**0.7096**	**1.3699**	**2,8380**
histolojiktip	-0.1660	0.2472	-0.6506	0.3248	0,8470
Tümör çapi	0.0602	0.0410	-0.0205	0.1412	1,0620

Çizelge 4.17'ye göre kilokaybı, platelet, LDH, histolojik tip ve tümor çapı değişkenlerinin akciğer kanseri hastalarında sağkalım sürelerine etkileri önemsizdir. Hemoglobin seviyesinin sağkalım süresini etkilediği ve hemoglobin seviyesi normal seviyeden uzaklaştıkça ölüm riskinin 1,202 kat arttığı diğer bir önemli bulgudur.

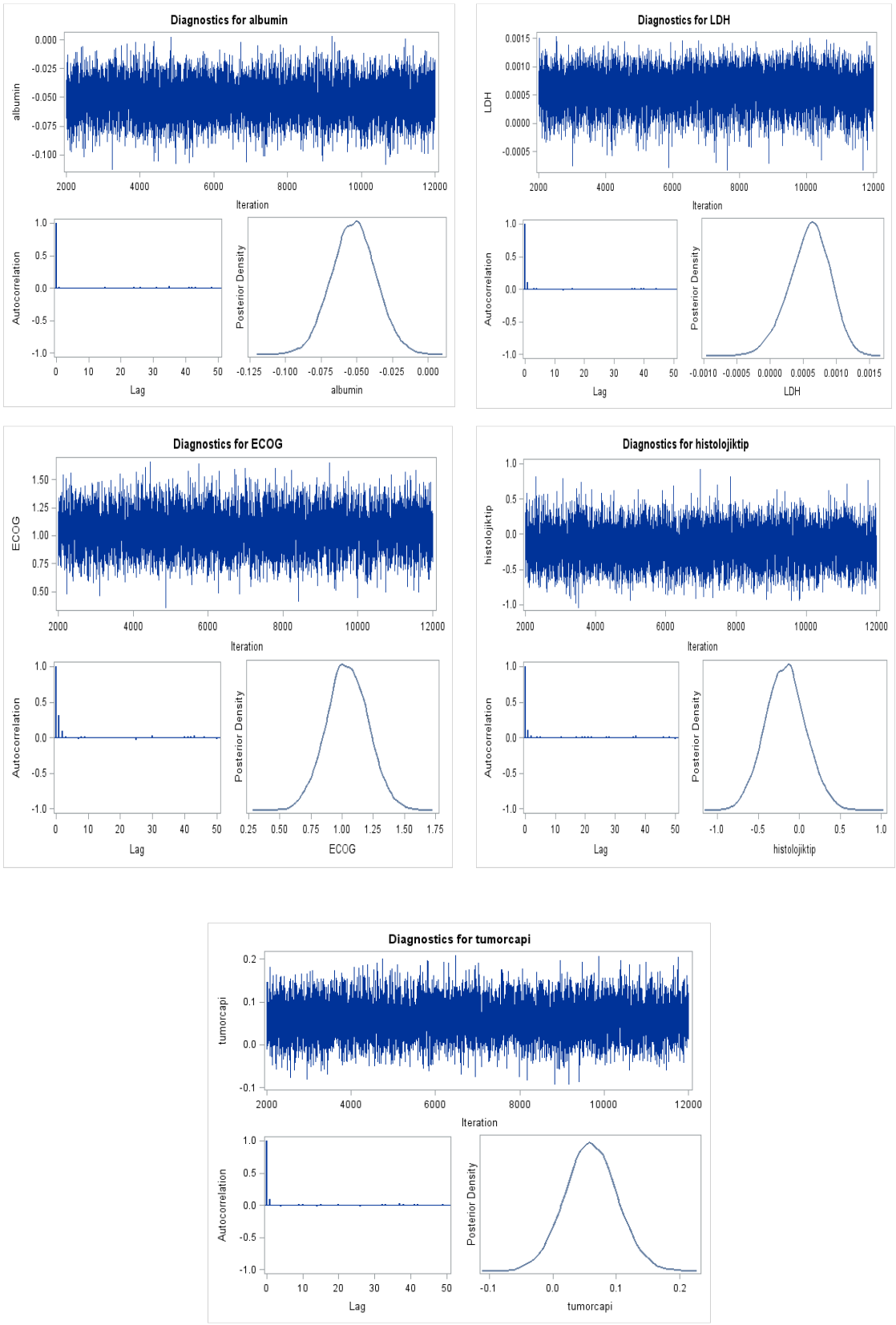
Protein ve albümin seviyesinin normal seviyeye yükselmesi hastanın taşıdığı ölüm riskini sırasıyla 0,721 ve 0,949 kat azalttığı görülmektedir.

ECOG performans ölçüsü sağkalım süresini etkileyen önemli bir değişken olarak bulunmuş ve hastaların tedavisini ayakta devam ettirenlere göre diğerlerinin 2,838 kat daha fazla ölüm riski taşıdığı belirlenmiştir.

Ulaşılabilen önseller yardımıyla akciğer verisine uygulanan BCR analizinde Markov zincirinin sonsal dağılıma yakınsamasını değerlendirmek için iz grafikleri, Geweke testi sonuçları, sonsal otokorelasyonlar Şekil 4.10'da sunulmuştur.



Şekil 4.10. Kayıp değerli akciğer kanserli hastaların sağkalım süresine ilişkin BÖBCR için MCMC'nin yakınsamasının iz grafikleri



Şekil 4.10. (Devamı) Kayıp değerli akciğer kanserli hastaların sağkalım süresine ilişkin BÖBCR için MCMC'nin yakınsamasının iz grafikleri

Şekil 4.10'daki her bir parametre için verilen grafikler incelendiğinde bütün parametreler için Markov zincirinin yakınsadığı görülmektedir.

Çizelge 4.18. Kayıp değerli akciğer kanserli hastaların sağkalım süresine ilişkin BÖBCR için MCMC'nin yakınsamasının Geweke test sonuçları

Geweke Testi		
Parametre	**z**	**Pr > |z|**
kilokaybi	0.0434	0.9654
hemoglbn	2.0233	0.0510
platelet	-1.2101	0.0980
protein	-1.3283	0.1841
albumin	0.5289	0.5969
LDH	-1.4237	0.1545
ECOG	0.4960	0.6199
histolojiktip	0.5733	0.5665
tumorcapi	-0.1936	0.8465

Çizelge 4.18'de Geweke test sonuçlarına göre her bir değişken zinciri sonsal dağılıma yakınsamaktadır (P>0,05).

Çizelge 4.19. Kayıp değerli akciğer kanserli hastaların sağkalım süresine ilişkin BÖBCR için MCMC'nin yakınsamasının sonsal otokorelasyonlarla değerlendirilmesi

Sonsal Otokorelasyonlar (Posterior Autocorrelations)				
Parametre	**Lag 1**	**Lag 5**	**Lag 10**	**Lag 50**
kilokaybi	0.2622	0.0286	0.0092	-0.0047
hemoglbn	0.3090	0.0080	-0.0011	-0.0087
platelet	0.1821	-0.0041	0.0028	-0.0014
protein	0.2827	0.0016	-0.0052	-0.0357
albumin	0.0179	0.0069	-0.0023	0.0077
LDH	0.1054	-0.0103	-0.0027	-0.0036
ECOG	0.3111	-0.0121	0.0064	-0.0222
histolojiktip	0.1047	0.0122	-0.0027	-0.0254
tumorcapi	0.0918	-0.0060	0.0120	-0.0059

Çizelge 4.19'a göre sonsal otokorelasyonlar incelendiğinde gecikmeler arasındaki korelasyonun küçük çıkması yakınsamanın sağlandığını gösterir.

Kayıp değerli akciğer kanseri verisine CR, BÖBCR ve BOÖBCR yöntemleri ayrı ayrı uygulanıp, uygulanan yöntemlerin karşılaştırılması amacıyla standart hatalar ve modele uygunluk kriterleri dikkate alınmıştır. Her bir yöntem için modele uygunluk

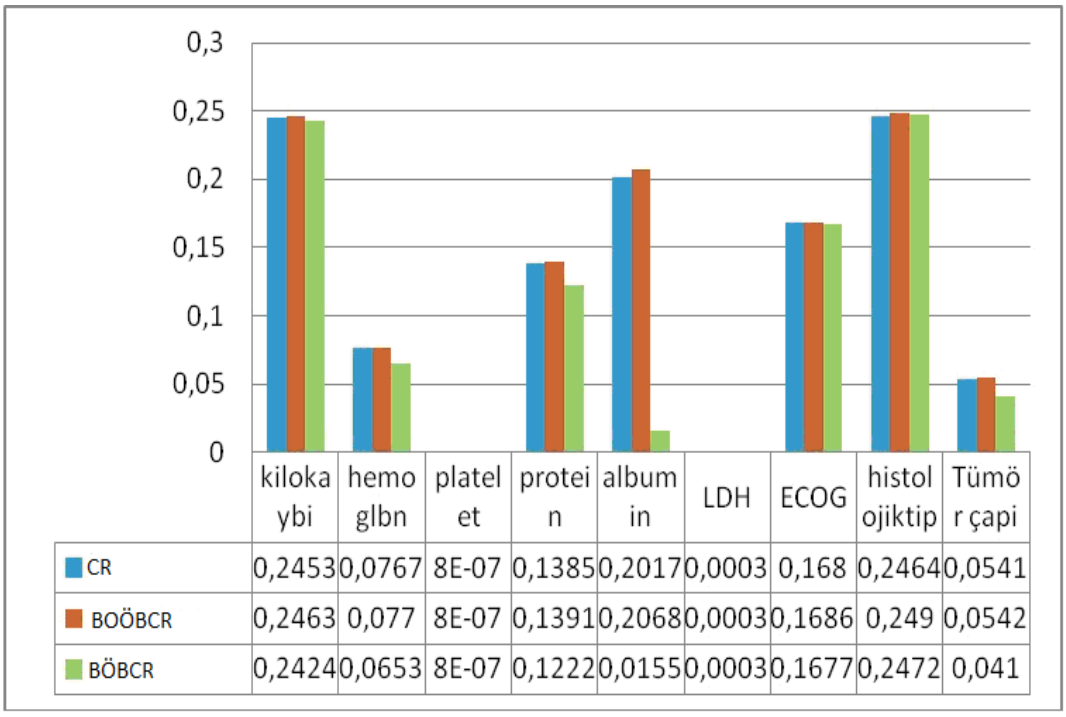
kriterleri Çizelge 4.20'de verilmiştir. BCR için AIC değeri hesaplanamadığından Spiegelhalter ve ark. (1998) önerdiği yaklaşık AIC değeri hesaplandı.

Çizelge 4.20. Kayıp değerli akciğer kanserli hastaların sağkalım süresine ilişkin her bir yöntem için hesaplanan uygunluk kriterleri

	AIC	**DIC** (p_D)
CR	831.584	-
BOÖBCR	≈831.621	831.592 (8.971)
BÖBCR	≈830.293	828.737 (7.444)

BÖBCR ve BOÖBCR için hesaplanan DIC değerleri incelendiğinde; CR ile benzer sonuçlar veren BOÖBCR'a ait DIC değeri (831.592) BÖBCR'a ait DIC değerinden (828.737) daha büyük bulundu. Bu da önsel bilginin varlığında BCR daha iyi bir performans sağladığını göstermektedir. Ayrıca CR analizinde bulunana AIC değeri ile BCR için hesaplanan yaklaşık AIC değerleri karşılaştırıldığında; BÖBCR en küçük AIC değerini verdiğinden diğer modellere göre daha iyi olduğu söylenebilir.

Her bir yöntem sonucu elde edilen standart hataların büyüklükleri hesaplanıp bulunan sonuçlar Şekil 4.11'de verilmiştir.

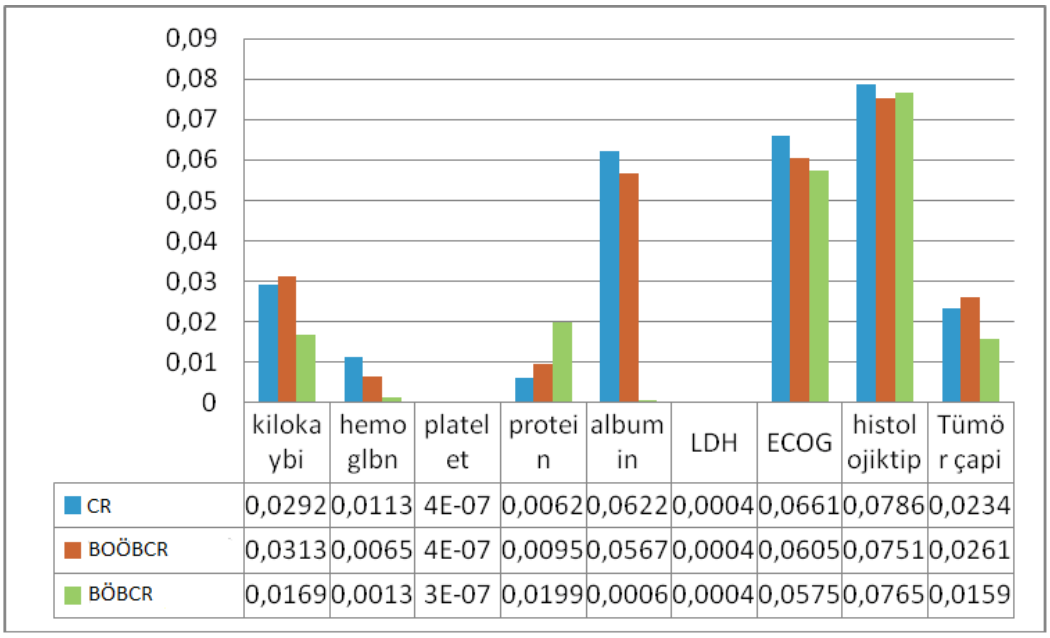


	kiloka ybi	hemo glbn	platel et	protei n	album in	LDH	ECOG	histol ojiktip	Tümö r çapi
CR	0,2453	0,0767	8E-07	0,1385	0,2017	0,0003	0,168	0,2464	0,0541
BOÖBCR	0,2463	0,077	8E-07	0,1391	0,2068	0,0003	0,1686	0,249	0,0542
BÖBCR	0,2424	0,0653	8E-07	0,1222	0,0155	0,0003	0,1677	0,2472	0,041

Şekil 4.11. Kayıp değerli veri seti için CR, BOÖBCR ve BÖBCR yöntemleri tarafından bulunan her bir değişkenin standart hatalarının grafiksel gösterimi

Şekil 4.11'de verilen standart hatalar incelendiğinde en küçük standart hataları BÖBCR yönteminin verdiği görülmüştür. Bu sonuçlara göre verilerde kayıp gözlem olması durumunda, değişkenlerle ilgili önsel bilgi varsa BÖBCR uygulanması önerilmektedir. Önsel bilgi olmadığında, BCR ile CR benzer sonuçlar verdiğinden, BOÖBCR'nun bir üstünlüğü de olmamaktadır.

Kayıp değerli akciğer verisine uygulanan yöntemler ile tam (orijinal) veriye uygulanan yöntemlerin regresyon katsayıları sonuçları karşılaştırıldığında, BÖBCR yönteminin sonuçları diğer yöntemlere göre tam verinin regresyon katsayılarına en fazla benzediği tespit edilmiştir. Bununla ilgili olarak tam veriye her bir yöntemin uygulanması sonucu elde edilen regresyon katsayıları ile kayıp değerli veriye uygulanan yöntemlerden elde edilen regresyon katsayıları arasındaki fark alınarak her bir yöntemin tam verinin regresyon katsayısına yakınlığı incelendi ve sonuçlar grafiksel olarak Şekil 4.12'de sunulmuştur.



	kiloka ybi	hemo glbn	platel et	protei n	album in	LDH	ECOG	histol ojiktip	Tümö r çapi
CR	0,0292	0,0113	4E-07	0,0062	0,0622	0,0004	0,0661	0,0786	0,0234
BOÖBCR	0,0313	0,0065	4E-07	0,0095	0,0567	0,0004	0,0605	0,0751	0,0261
BÖBCR	0,0169	0,0013	3E-07	0,0199	0,0006	0,0004	0,0575	0,0765	0,0159

Şekil 4.12. Tam veriden elde edilen parametrelerle kayıp değerli veriden elde edilen parametreler arası farkın grafiği

Şekil 4.12'de verilen sonuçlar incelendiğinde kilo kaybı, hemoglobin, platelet, albümin, LDH, ECOG ve tümör çapı değişkenleri için BÖBCR analizi yöntemi tam (orijinal) veriden elde edilen regresyon katsayısına en yakın değerleri vermektedir. Protein ve histolojik tip değişkenleri ise BOÖBCR yöntemine daha yakın değer elde etmiştir.

4.4. Uygulama IV

Uygulamanın son bölümünde ise kayıp değer içeren akciğer kanseri verisi kayıp veri analizlerinden biri olan çoklu değer atama (Multiple İmputation-MI) yöntemi ile tamamlanıp daha sonra Bayesci ve klasik Cox regresyon uygulanarak elde edilen sonuçlar tam (orijinal) veriden ve CCA sonucu uygulanan yöntemlerden elde edilen bulgularla karşılaştırılmıştır.

4.4.1. Kayıp Değerli CR ve BCR İçin Çoklu Değer Atama (MI) Yöntemi

Bölüm 4.1.1.'de büyük örnek genişliklerinde (N>100) ve kayıp değer oranı %20 ve altında olduğunda tam veri ile elde edilen sonuçlara en yakın değerleri MI, EM ve REG yöntemleri ile ulaşılmıştı. Ayrıca her bir kayıp değere sadece bir değer atayan

MEAN, REG, EM yöntemlerine göre birden fazla değer atayan MI yönteminin kullanılması önerilmişti**.** Bu sonuçtan yola çıkarak akciğer kanserli 174 hastaya ait veriden MAR varsayımına göre değerler silinerek %20 kayıp değer içeren veri seti oluşturuldu. Bu kayıp değerli veri seti MI ile tamamlanarak, BCR analizinin performansı değerlendirildi ve CR analizi sonuçları ile karşılaştırıldı.

BCR bilgilendirici önselli ve bilgilendirici olmayan önselli olmak üzere iki şekilde uygulandı. Geçmiş benzer çalışmalar incelendi ve ulaşılan parametre değerleri ve standart hataları BCR için bilgilendirici önsel olarak kullanıldı.

MI yönteminde her bir kayıp değer için 5 farklı değer atanarak, toplam 5 tamamlanmış veri seti elde edildi. Bu tamamlanmış veri setleri ayrı ayrı ilgilenilen analiz yöntemi ile analiz edildi ve her birinden elde edilen çıkarımların Rubin'in (1987) kuralına göre birleştirilerek sonuç tablosu oluşturuldu.

Kayıp değerli akciğer kanseri verisine MI yöntemi ile CR, BÖBCR ve BOÖBCR yöntemleri ayrı ayrı uygulandı ve bulunan standart hatalar karşılaştırıldı. Ayrıca tam veriye (orijinal) uygulanan her bir yöntem sonucunda elde edilen parametre değerine en yakın değeri bulan yöntemi belirlemek için parametre değerleri karşılaştırıldı.

4.4.1.1. Çoklu Değer Atama (MI) ve Cox Regresyon (CR) Analizi

Kayıp değer içeren verideki her bir kayıp değer MI yöntemi ile tamamlanıp, 5 tamamlanmış veri seti oluşturuldu ve her birine CR yöntemi uygulanarak ayrı ayrı elde edilen sonuçlar birleştirildi. Elde edilen sonuçlar Çizelge 4.21'de verilmiştir.

Çizelge 4.21. MI ile tamamlanmış veri setinin CR analiz sonuçları

Değişkenler	**Parametre Tahminleri**	**Standart Hatalar**	**Güven aralığı**		**Hazard Oranı**
kilokaybi	-0.135461	0.228669	-0.58365	0.31272	0,873
hemoglbn	**0.176057**	**0.064013**	**0.05059**	**0.30153**	**1,193**
platelet	0.000000672	0.000000699	-0.00001	0.00000	1,000
protein	**-0.317694**	**0.140723**	**-0.59390**	**-0.04149**	**0,728**
albumin	-0.138365	0.194019	-0.51881	0.24208	0,871
LDH	0.000129	0.000129	-0.00015	0.00041	1,000
ECOG	**1.118593**	**0.150189**	**0.82422**	**1.41296**	**3,061**
histolojiktip	-0.243413	0.211949	-0.65883	0.17200	0,784
Tümör çapi	0.085078	0.045349	-0.00381	0.17397	1,089

Çizelge 4.21'e göre akciğer kanserli hastaların sağkalım süreleri üzerinde hastanın kilo kaybı, platelet seviyesi, albümin değeri, LDH değeri, histolojiktipi ve tümor çapı değişkenlerinin önemli bir etkilerinin olmadığı %95 güvenle söylenebilir. Buna karşın hastanın hemoglobin değeri, protein değeri ve ECOG performans durumunun sağkalım üzerinde önemli etkileri vardır.

Akciğer kanserli hastanın hemoglobin değeri normal seviyeden uzaklaştıkça ölüm riskinin 1,193 kat arttığı, protein seviyesi normal seviyeye yükseldikçe ölüm riskinin 0,728 kat azaldığı ve ECOG seviyesi için tedavisini ayakta devam ettirenlere göre diğerlerinin 3,061 kat risk taşıdığı söylenebilir.

4.4.1.2. Çoklu Değer Atama (MI) ve Bilgilendirici Olmayan Önselli Bayesci Cox Regreyon (BOÖBCR) Analizi

Kayıp değerli akciğer verisi çoklu değer atama yöntemi ile tamamlanmış ve 5 ayrı tamamlanmış veri seti elde edilmiştir. Her bir tamamlanmış veri setine BOÖBCR analizi uygulanmış ve birleştirilmiş sonuçlar Çizelge 4.22'de sunulmuştur.

Çizelge 4.22. MI ile tamamlanmış veri setinin BOÖBCR sonuçları

Değişkenler	Parametre Tahminleri	Standart Hatalar	Güven aralığı		Hazard Oranı
kilokaybi	-0.130686	0.228844	-0.57921	0.31784	0,878
hemoglbn	0.176498	0.065017	0.04907	0.30393	1,193
platelet	0.000000731	0.000000708	-0.000000	0.00001	1,000
protein	-0.307337	0.139507	-0.58086	-0.03381	0,735
albumin	-0.153072	0.192280	-0.52994	0.22380	0,858
LDH	0.000060556	0.000107	-0.00015	0.00027	1,000
ECOG	1.137787	0.149294	0.84517	1.43040	3,120
histolojiktip	-0.238301	0.212106	-0.65402	0.17742	0,788
Tümör çapi	0.071766	0.044787	-0.01602	0.15955	1,074

Çizelge 4.22'ye göre CR ile elde edilen sonuçlara benzer şekilde sağkalım süresini etkileyen önemli değişkenlerin hemoglobin, protein ve ECOG performans durumu olduğu görülmüştür. Hastanın sağkalım süresi üzerinde kilokaybı, platelet seviyesi, albümin değeri LDH seviyesi, histolojik tipi ve tümor çapı değişkenler açısından önemli etkileri bulunmadı.

Elde edilen bu sonuçlara göre hastanın hemoglobin değeri normal seviyeden uzaklaştıkça taşıdığı riskin 1,193 kat arttığı protein seviyesi normal seviyeye yükseldikçe riskin 0,735 kat azaldığı ve ECOG performans durumu için tedavisini ayakta devam ettirenlere göre riskin 3,120 kat arttığı söylenebilir.

4.4.1.3. Çoklu Değer Atama (MI) ve Bilgilendirici Önselli Bayesci Cox Regresyon (BÖBCR) Analizi

Kayıp değerli akciğer verisi MI yöntemi ile tamamlandı ve 5 ayrı tamamlanmış veri seti elde edildi. Her bir tamamlanmış veri setine BÖBCR analizi uygulandı ve birleştirilmiş sonuçlar Çizelge 4.23'de verilmiştir.

Akciğer kanseri için daha önce yapılmış çalışmalar incelenerek hemoglobin, protein, LDH, albümin ve tümör çapı değişkenleri için önsel bilgiler elde edildi. Önsel bilgisine ulaşılamayan diğer değişkenler için normal önsel kullanıldı. Değişkenler için elde edilen önsel bilgiler Çizelge 4.5'de verilmiştir.

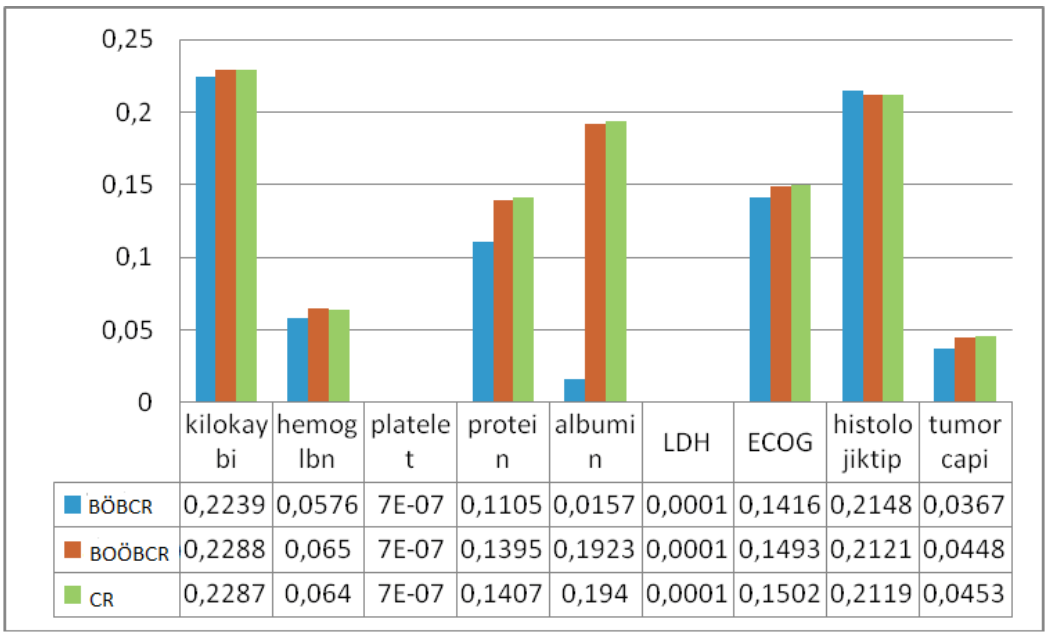
Çizelge 4.23. MI ile tamamlanmış veri setinin BÖBCR sonuçları

Değişkenler	**Parametre Tahminleri**	**Standart Hatalar**	**Güven aralığı**		**Hazard Oranı**
kilokaybi	-0.07364	0.223874	-0.5124336	0.365154	0,929
hemoglbn	0.17846	0.057603	0.06555792	0.291362	1,195
platelet	8.23E-07	6.91E-07	-5.311E-07	2.18E-06	1,000
protein	-0.35386	0.110475	-0.5703901	-0.13733	0,702
albumin	**-0.05288**	**0.015741**	**-0.0837321**	**-0.02203**	**0,949**
LDH	0.000188	0.000101	-1.683E-05	0.000393	1,000
ECOG	0.99868	0.141618	0.72110784	1.276252	2,715
histolojiktip	-0.23264	0.214798	-0.6536436	0.188364	0,792
Tümör çapi	0.06948	0.036657	-0.0023675	0.141327	1,072

Çizelge 4.23'e göre sağkalım süresi üzerinde hastanın kilokaybı, platelet seviyesi, albümin değeri, LDH değeri, histolojik tipi ve tümör çapı değişkenlerinin önemli etkileri tespit edilmemiştir. Hastanın hemoglobin seviyesi, protein değeri, albümin değeri ve ECOG performans değeri değişkenlerinin ise sağkalım süresini etkileyen önemli faktörler olduğu görülmüştür.

Çizelge 4.23'de verilen sonuçlara göre hastanın hemoglobin değeri normal seviyeden uzaklaştıkça ölüm riski 1,195 kat artmakta, protein değeri normal seviyeye yükseldikçe risk 0,702 kat azalmakta, albümin değeri normal seviyeye yükseldikçe risk 0,949 kat azalmakta ve ECOG performans durumu ayakta tedavi olanlara göre diğerlerinden 2,715 kat risk taşımaktadır.

Akciğer kanseri verisindeki kayıp değerlerin MI yöntemi ile tamamlanarak veri setine CR, BÖBCR ve BOÖBCR ayrı ayrı uygulandı. Uygulanan bu yöntemlerin karşılaştırılması amacıyla standart hatalar incelenmiş ve yöntemlerden elde edilen standart hatalar Şekil 4.13'de gösterilmiştir.

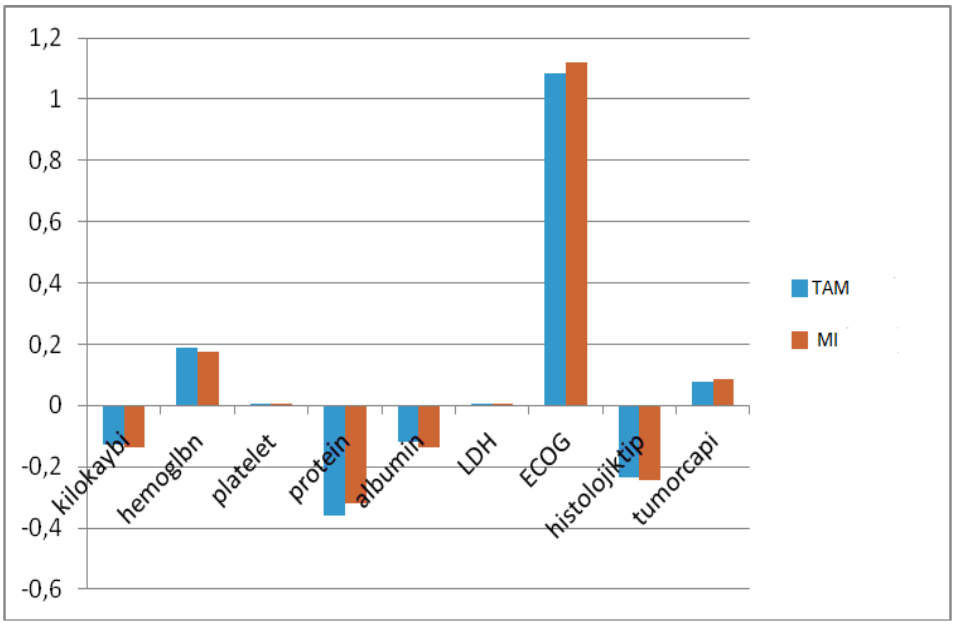


	kilokaybi	hemoglbn	platelet	protein	albumin	LDH	ECOG	histolojiktip	tumor capi
BÖBCR	0,2239	0,0576	7E-07	0,1105	0,0157	0,0001	0,1416	0,2148	0,0367
BOÖBCR	0,2288	0,065	7E-07	0,1395	0,1923	0,0001	0,1493	0,2121	0,0448
CR	0,2287	0,064	7E-07	0,1407	0,194	0,0001	0,1502	0,2119	0,0453

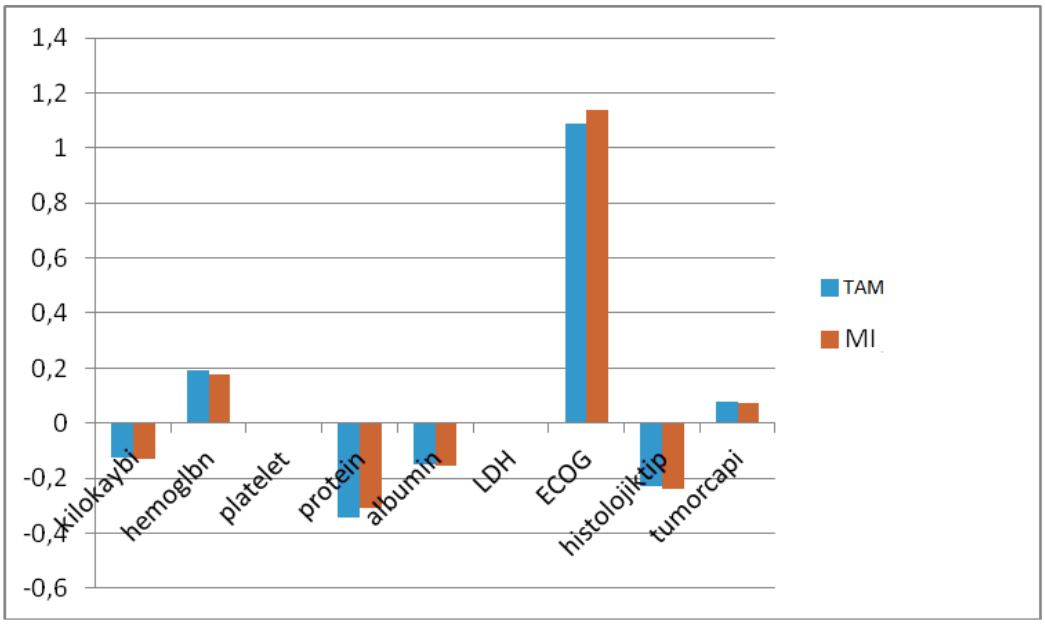
Şekil 4.13. MI ile tamamlanmış veri seti için CR, BÖBCR ve BOÖBCR yöntemleri tarafından bulunan her bir değişkenin standart hatalarının grafiksel gösterimi

Şekil 4.13'e göre hemen hemen bütün değişkenler için BÖBCR sonucu bulunan standart hatalar BOÖBCR ve CR analizlerinin bulduğu standart hatalardan daha küçüktür. Genel olarak BOÖBCR ve CR analizlerinin sonuçları benzer bulunmuştur.

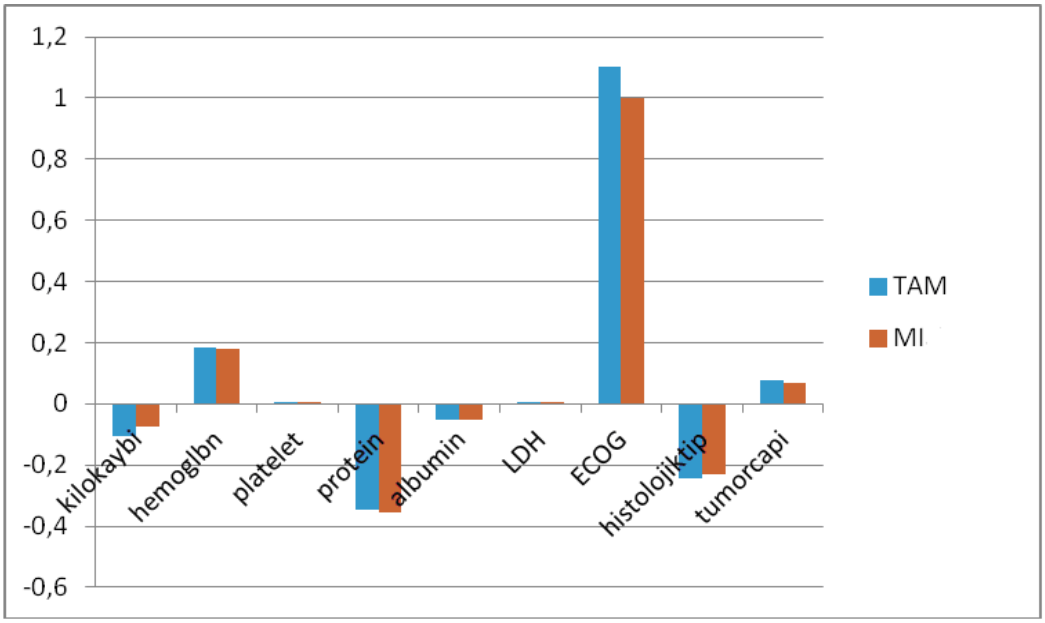
Kayıp veri tamamlama yöntemlerinden biri olan MI yönteminin performansını değerlendirmek amacıyla MI ile tamamlanmış verinin CR, BÖBCR ve BOÖBCR katsayıları ile orijinal veriden elde edilen regresyon katsayıları ayrı ayrı her bir yöntem için karşılaştırıldı. Sonuçlar Şekil 4.14, Şekil 4.15 ve Şekil 4.16'da verilmiştir.



Şekil 4.14. MI ile tamamlanmış verinin CR katsayıları ve orijinal (tam) veriden elde edilen regresyon katsayılarının karşılaştırması



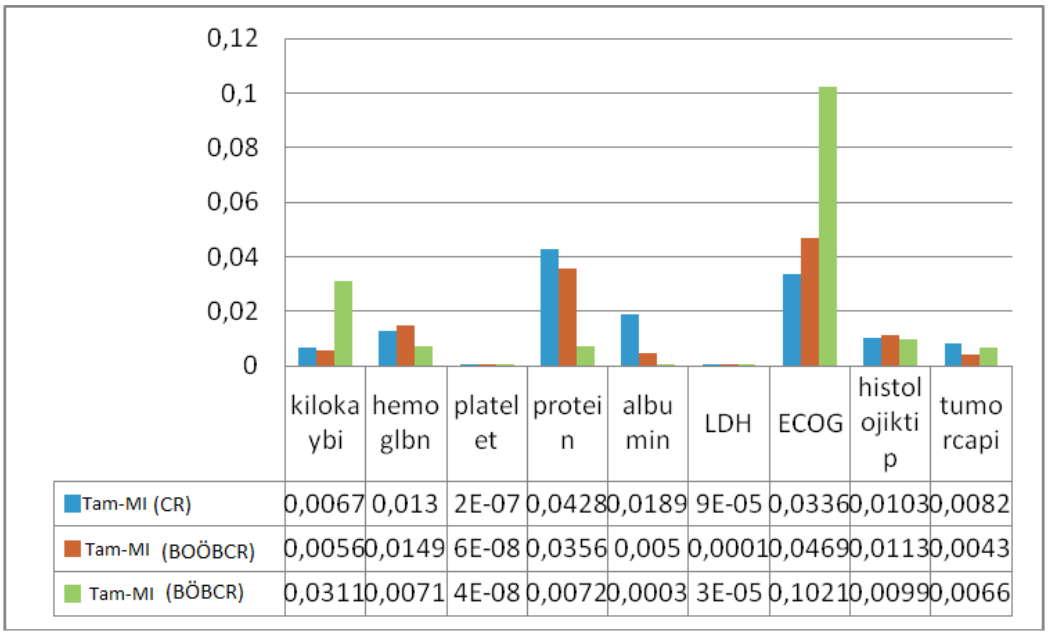
Şekil 4.15. MI ile tamamlanmış verinin BOÖBCR katsayıları ve orijinal (tam) veriden elde edilen regresyon katsayılarının karşılaştırması



Şekil 4.16. MI ile tamamlanmış verinin BÖBCR katsayıları ve orijinal (tam) veriden elde edilen regresyon katsayılarının karşılaştırması

Şekil 4.14, 4.15 ve 4.16'da verilen grafikler incelendiğinde bütün yöntemlerde MI sonucu uygulanan yöntemler ile orijinal veriden elde edilen regresyon katsayıları benzer bulunmuştur. Bu sonuç, kayıp değerli verilerde MI yönteminin kullanılmasının doğru sonuçlar vereceğini göstermiştir.

Yöntemler arasındaki farkı daha iyi görebilmek ve orijinal verideki regresyon katsayılarına her bir yöntemin ne kadar yakın sonuçlar verdiğini görmek için her bir parametre için farklar mutlak değerce alınmış ve sonuçlar Şekil 4.17'de sunulmuştur.



Şekil 4.17. CR, BOÖBCR ve BÖBCR yöntemlerinin her biri için tam ve MI yöntemi ile tamamlanmış veriden elde edilen regresyon katsayıların karşılaştırılması

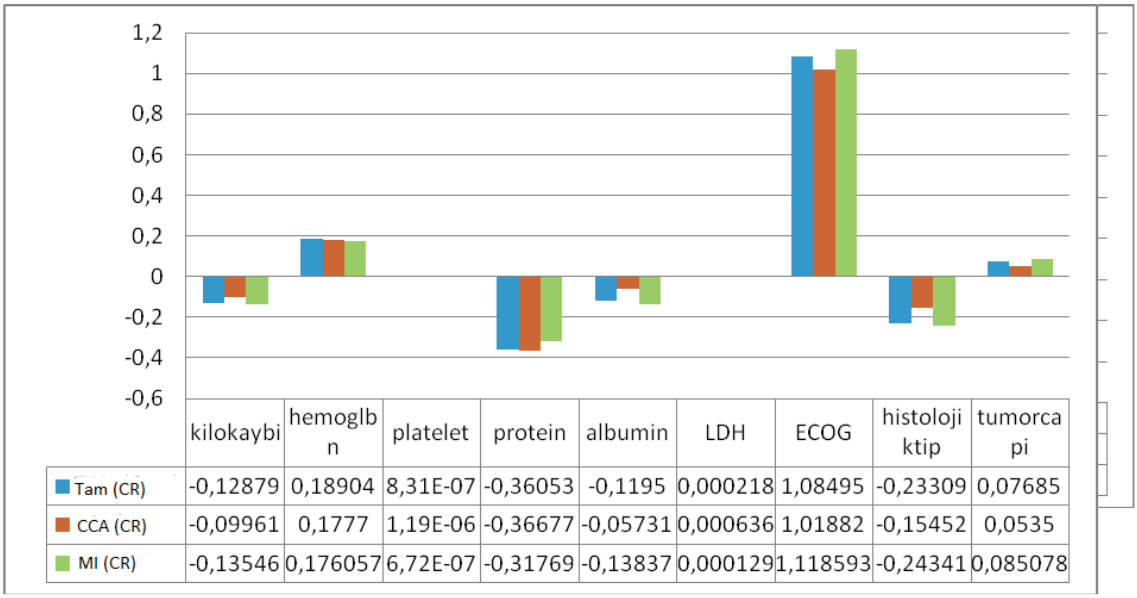
Şekil 4.17'de verilen grafik incelendiğinde kilokaybı ve tümor çapı için en yakın regresyon katsayısını BOÖBCR'nun verdiği görülmüştür. Hemoglobin, platelet, protein, albümin LDH, histolojik tip için ise en yakın regresyon katsayısını BÖBCR elde etmiştir. ECOG değişkeni için ise en yakın regresyon katsayısı, CR analizi sonucu elde edilmiştir.

Sonuç olarak MI ile tamamlanan veriden BÖBCR analizi ile en küçük standart hatalar elde edildi ve parametrelerin tamamında orijinal verinin analiz edilmesiyle bulunan parametrelere çok yakın değerler bulundu. Parametrelerin çoğunda BÖBCR diğer iki yönteme göre (CR, BOÖBCR) en yakın değerleri vermiştir.

Sağlık alanında yapılan çalışmalarda kayıp değerlerle sıklıkla karşılaşılmakta ve bu durumda araştırmacılar uygulama kolaylığı bakımından sıklıkla eksiksiz veri analizi (CCA) yöntemine başvurmaktadır. Bu çalışmada çok sık kullanılan CCA yöntemi ile kayıp veri problemlerinde iyi bir performans gösteren çoklu değer atama (MI) yöntemi sonuçlarının karşılaştırması yapılmıştır.

Bu amaçla kayıp veri giderme yöntemleri ile tamamlanan veriler CR, BÖBCR ve BOÖBCR yöntemleri ile veri seti analiz edildi ve elde edilen regresyon katsayıları orijinal verinin regresyon katsayıları ile karşılaştırıldı.

MI ve CCA sonucu elde edilen verilerin CR analizi sonucu elde edilen regresyon katsayıları ve orijinal veriden elde edilen regresyon katsayıları sonuçları karşılaştırılarak bulgular Şekil 4.18'de verilmiştir.

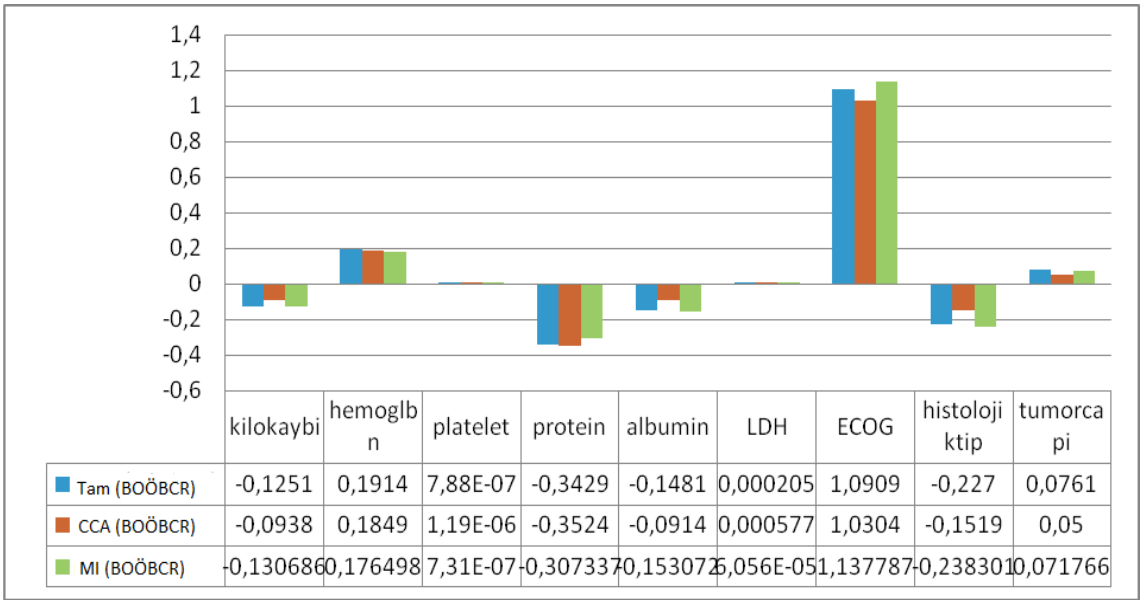


	kilokaybi	hemoglbn	platelet	protein	albumin	LDH	ECOG	histolojiktip	tumorcapi
Tam (CR)	-0,12879	0,18904	8,31E-07	-0,36053	-0,1195	0,000218	1,08495	-0,23309	0,07685
CCA (CR)	-0,09961	0,1777	1,19E-06	-0,36677	-0,05731	0,000636	1,01882	-0,15452	0,0535
MI (CR)	-0,13546	0,176057	6,72E-07	-0,31769	-0,13837	0,000129	1,118593	-0,24341	0,085078

Şekil 4.18. Tam, CCA ve MI yöntemlerine göre CR analizi sonucu elde edilen regresyon katsayıları

Şekil 4.18'de verilen grafikler incelendiğinde, kilo kaybı, platelet, albümin, LDH, ECOG, histolojik tip ve tümör çapı değişkenleri için tam veriden elde edilen regresyon katsayılarına yakın değerlere MI yöntemi ile uygulanan CR ile ulaşılmıştır.

MI ve CCA sonucu elde edilen verilerin BOÖBCR analizi sonucu elde edilen regresyon katsayıları ve orijinal verinin sonuçları karşılaştırılmış ve sonuçlar Şekil 4.19'de sunulmuştur.



	kilokaybi	hemoglbn	platelet	protein	albumin	LDH	ECOG	histolojiktip	tumorcapi
Tam (BOÖBCR)	-0,1251	0,1914	7,88E-07	-0,3429	-0,1481	0,000205	1,0909	-0,227	0,0761
CCA (BOÖBCR)	-0,0938	0,1849	1,19E-06	-0,3524	-0,0914	0,000577	1,0304	-0,1519	0,05
MI (BOÖBCR)	-0,130686	0,176498	7,31E-07	-0,307337	0,153072	5,056E-05	1,137787	-0,238301	0,071766

Şekil 4.19. Tam, CCA ve MI yöntemleri ile BOÖBCR analizi sonucu elde edilen regresyon katsayılarının karşılaştırılması

Şekil 4.19'de verilen grafiklere göre, kilo kaybı, platelet, albümin, LDH, ECOG, histolojik tip, tümör çapı değişkenleri için tam veriden elde edilen regresyon katsayısına en yakın değerin çoklu değer atama ile uygulanan BOÖBCR yöntemi ile bulunduğu gözlenmektedir.

MI ve CCA sonucu elde edilen verilerin BÖBCR analizi sonucu elde edilen regresyon katsayıları ve orijinal verinin sonuçları karşılaştırıldı ve sonuçlar Şekil 4.20'de gösterildi.

	kilokaybi	hemoglbn	platelet	protein	albumin	LDH	ECOG	histolojiktip	tumorcapi
Tam(BÖBCR)	-0,1047	0,1856	8,66E-07	-0,3467	-0,0526	0,000213	1,1008	-0,2425	0,0761
CCA (BÖBCR)	-0,0878	0,1843	1,13E-06	-0,3268	-0,052	0,000593	1,0433	-0,166	0,0602
MI (BÖBCR)	-0,07364	0,17846	8,23E-07	-0,35386	-0,05288	0,000188	0,99868	-0,23264	0,06948

Şekil 4.20. Tam, CCA ve MI yöntemlerine göre BÖBCR analizi sonucu elde edilen regresyon katsayıları

Şekil 4.20'de verilen grafiklere göre, platelet, protein, albümin, LDH, histolojik tip ve tümör çapı değişkenleri için tam veriden elde edilen regresyon katsayılarına en yakın değeri MI ile birlikte BÖBCR analizi ile bulunduğu görülmektedir.

Sonuç olarak kayıp değerlerin MI yöntemi ile tamamlanarak her bir yöntemin uygulanması sonucu elde edilen regresyon katsayıları neredeyse bütün değişkenler için tam (orijinal) veriden elde edilen regresyon katsayılarına yakın bulundu. CCA yöntemi ile uygulanan her bir yöntemde bulunan regresyon katsayıları gerçek verinin değerlerine MI ile uygulanan yöntemlerde bulunan regresyon katsayıları kadar yaklaşamamıştır.

5. SONUÇLAR VE ÖNERİLER

Araştırmaların çoğunda kayıp değerler ile karşılaşılmaktadır ve bu durum çok ciddi problem olarak görülmektedir. Çünkü istatistiksel analizler ve bilgisayar programları genelde verinin tümünün var olduğu durumlar için geliştirilmiştir. Bu durumla karşılaşan pek çok araştırmacı kayıp değer içeren gözlemi silerek çalışmalarını sürdürmektedir. Verilerdeki bu silme işlemi örnek genişliğini azaltmakta ve istatistiksel gücün küçülmesine neden olmaktadır. Bu açıdan kayıp veri probleminin giderilmesi önem kazanmaktadır.

Bu çalışmada, kayıp veri problemini gideren yöntemlerden en çok kullanılan ortalama değer atama (MEAN), regresyon değer atama (REG), beklenen maksimizasyon (EM), çoklu değer atama (MI) ve eksiksiz veri analizi (CCA) yöntemleri ayrıntılı olarak incelenmiş ve farklı örnek genişlikli (50, 100, 200 birimli örnekler) verilerde ve farklı kayıp oranlarında (%5, 10, 20 ve 40) yöntemlerin performansları Cox Regresyon (CR) analizi uygulanarak incelenmiştir.

Kayıp değerli gözlem oranı %5 olduğunda, CCA dışındaki bütün kayıp veri giderme yöntemleri örnek genişliği ne olursa olsun tam veriden elde edilen sonuçlara yakın sonuçlar verdiği gözlenmiştir.

Verideki kayıp değerli gözlem oranı arttıkça yöntemler arasındaki farklılığın arttığı görülmüştür. Örnek genişliği arttıkça kayıp veri giderme yöntemlerinin birbirine benzerliklerinde arttığı tespit edilmiştir.

Örnek genişliği az olduğunda (N<100) CCA yönteminin kullanılmasının veri sayısını daha da azaltacağından önerilmemektedir. Uygulaması çok kolay olan bu yöntem kullanılacaksa örnek genişliği (N>100) olmalı ve kayıp değerli gözlem oranı %20 ve daha az olmalıdır. Ancak bu koşulda bulunan regresyon katsayıları tam veriden elde edilen regresyon katsayılarına benzer olmaktadır.

Örnek genişliği 50 ve kayıp değer oranı %20'den az olduğunda gerçek değere en yakın sonuçları REG, MI ve MEAN yöntemleri vermektedir.

Büyük örnek genişliklerinde (N>100) ve kayıp değer oranı %20 ve altında ise tam veri ile elde edilen sonuçlara en yakın değerleri MI, EM ve REG yöntemleri vermektedir.

Kayıp oranı %40 olması durumunda bütün örnek genişliklerinde kayıp veri tamamlama yöntemleri tam veri ile elde edilen değerlere her zaman yakın değerler

bulamadığı için tutarsızlık göstermişlerdir. Bu durumda bile MI yönteminin sonuçları diğer yöntemlere göre daha iyi performans sunmaktadır.

Her bir kayıp değere sadece bir değer atayan MEAN, REG ve EM yöntemlerine göre birden fazla değer atayan MI yönteminin kullanılması önerilmektedir**.** Ayrıca örnek genişliği arttıkça kayıp değer tamamlama yöntemleri ile elde edilen sonuçlar ile tam veri kullanılarak elde edilen sonuçlar arasındaki farkların azaldığı tespit edilmiştir.

Gerçek bir veri setinde akciğer kanserli hastaların sağkalım süresini etkileyen prognostik faktörleri belirlemek için BCR ve CR analizi kullanıldı. BCR analizi bilgilendirici ve bilgilendirici olmayan önselli olmak üzere iki şekilde uygulandı. Bilgilendirici önseller daha önce yapılmış benzer çalışmalardan elde edildi.

CR ve BOÖBCR analizleri sonucunda hemoglobin, protein, LDH değerleri ile ECOG performans durumunun sağkalım süresi üzerinde önemli etkiye sahip faktörler olarak bulundu. Yani CR ve BOÖBCR yöntemleri aynı sonucu vermişlerdir. BÖBCR analizi sonucunda ise hemoglobin, protein, albümin, LDH değerlerinin, ECOG performans durumunun, tümör çapı sağkalım süresi üzerinde önemli etkiye sahip faktörler olarak bulundu. Bütün yöntemlerde akciğer kanserli hastanın sağkalım süresini etkileyen ortak değişkenler hemoglobin, protein, LDH değerlerinin ve ECOG performans durumudur. BÖBCR da diğer yöntemden farklı olarak albümin ve tümör çapı değişkenleri önemli bulunmuştur.

Her bir yöntem sonucu bulunan standart hatalar ve model uyum kriterleri incelendiğinde BÖBCR en küçük standart hataları ve model uyum kriterini verdiğinden, bu yöntemin kullanılması tercih edilmelidir. BOÖBCR ile CR analizi yöntemleri benzer sonuçlar vermiştir. İyi bir kaynak taraması yapılarak elde edilen veriye uygun ve gerçeği yansıtan önseller, BCR analizinin performansını artırmaktadır. Bu nedenle BCR uygulanacaksa önsel bilginin doğru seçilmesi önemlidir.

Sağkalım verisinde kayıp değer olması durumunda BCR performansını değerlendirmek amacıyla akciğer kanserli hastaların verisi MAR varsayımına göre % 20 kayıp değerli gözlem olacak şekilde azaltıldı. Kayıp veri analiz yöntemlerinin N>100 ve %20 ve daha az kayıp oranında iyi performans göstermesinden dolayı kayıp değerli gözlem oranı bu şekilde belirlendi. Kayıp değerli akciğer kanserli hastaların verisi uygulama kolaylığı yönünden çok sık kullanılan CCA yöntemi ile kayıp değer problemi giderildi ve CR ile BCR analizleri ayrı ayrı uygulandı. Her bir analiz sonucu elde edilen

regresyon katsayıları kayıp değer içermeyen orijinal verinin regresyon katsayıları ile karşılaştırıldı.

Kayıp değerli akciğer kanserli veriye CR analizi uygulandığında sağkalım süresini etkileyen önemli faktörler hemoglobin, protein, LDH ve ECOG performansı bulundu. BOÖBCR analizi sonucu ise önemli faktörler hemoglobin, protein ve ECOG değişkenleri anlamlı bulundu. BÖBCR analizi sonucunda da hemoglobin, protein, albümin ve ECOG performans anlamlı çıktı.

Uygulanan yöntemleri karşılaştırmak amacıyla standart hatalar ve model uygunluk kriterleri incelendiğinde en küçük kriteri ve standart hatayı BÖBCR analizi vermektedir. Ayrıca kayıp değerli veriye uygulanan her bir yöntem sonucu bulunan regresyon katsayıları kayıp değer içermeyen orijinal veriden elde edilen regresyon katsayıları ile karşılaştırıldığında gerçek değere en yakın değeri yine BÖBCR analizinin verdiği görülmüştür.

Kayıp değerli gözlem oranı %20 ve daha az olan aynı zamanda örnek genişliği N>100 durumunda iyi performans gösteren kayıp veri analiz yöntemlerinden biri olan MI kullanılarak tamamlanan akciğer kanseri verisi için CR ve BCR analizleri uygulandı.

Kayıp değerlerin MI yöntemi ile tamamlanarak CR uygulandığında akciğer kanserli hastaların sağkalım süresini etkileyen faktörlerden hemoglobin, protein ve ECOG performans değerlerinin anlamlı olduğu görüldü. Kayıp değerler MI yöntemi ile tamamlanıp BOÖBCR analizi uygulandığında, önemli faktörler yine hemoglobin, protein ve ECOG performans değeri bulundu. Kayıp değerler MI yöntemi ile tamamlanıp BÖBCR analizi uygulandığında önemli faktörler hemoglobin, protein, albümin ve ECOG performans değerleridir. BÖBCR yönteminde diğer iki yöntemden farklı olarak albümin değişkeni de anlamlı çıkmıştır.

Kayıp değerlerin MI ile tamamlanıp her bir yöntemin uygulanması sonucu elde edilen standart hatalar incelendiğinde BÖBCR analizi en küçük standart hataları verdi. Ayrıca BÖBCR sonucu bulunan regresyon katsayıları orijinal veriden elde edilen regresyon katsayılarına en yakın değerleri verdi.

Sonuç olarak veriye uygun gerçeği yansıtan önsel bilginin kullanılması durumunda veride kayıp değer olmadığında, kayıp değer olduğunda ve kayıp değerlerin MI yöntemi ile tamamlandığında BÖBCR yöntemi, CR ve BOÖBCR yöntemlerine göre daha uygun sonuçlar vermektedir. Çalışılan konu ile ilgili gerçeği yansıtan önsel bilgi

olmadığında BCR yönteminin CR yöntemine göre avantaj sağlamadığı ve sonuçların benzer olduğu tespit edildi.

Daha sonraki çalışmalarda, kullanılan veriye uygun farklı önseller kullanılarak BCR ile CR yöntemleri karşılaştırılabilir.

Ayrıca çalışmada kayıp veri analiz yöntemlerinden MI yöntemi diğer kayıp veri analizi yöntemlerine göre daha iyi sonuçlar vermiştir. Bu durum farklı kayıp veri oluşum mekanizmaları için de araştırılabilir.

6. KAYNAKLAR

Abreu, C.M., Chatkin, J.M., Fritscher, C.C., Wagner M.B., Pinto, J.A.L.F., 2003. Long-term survival in lung cancer after surgical treatment: is gender a prognostic factor? www.scielo.br/pdf/.../en_v30n1a03.pdf (15.03.2012).

Akaike, H., 1973. Information theory and an extension of the maximum likelihood principle. In Proc. 2 nd Int. Symp. Information Theory (eds B.N.Petrov and F.Csaki), s: 267-281, Budapest: Akademiai Kiado.

Allison, P.D., 2000. Multiple Imputation for Missing Data: A Cautionary Tale. Sociological Methods and Research, 28, 301–309.

Anderson, M.N., 2007. Bayesian and non-Bayesian techniques applied to censored survival data with missing values, Technical University of Denmark, IMM-PHD, ISSN 0909-3192.

Bayes, T., 1763. An essay towards solving a problem in the doctrine of Chances. Philosophical Transactions of the Royal Society of London, 53, 370-418.

Breslow. N., 1974. Covariance analysis of censored survival data. Biometrics, 30(1), 89–99.

Calle, M.L., Hough, G., Curia, A., Gómez, G., 2006, Bayesian survival analysis modeling applied to sensory shelf life of foods, Food Quality and Preference, 17(3- 4), 307-312.

Chen, M. H., Ibrahim, J. G. & Lipsitz, S. R., 2002, Bayesian methods for missing covariates in cure rate models, Lifetime Data Analysis 8, 117–146.

Congdon, P., 2006. *Bayesian Statistical Modeling*. John Wiley&Sons, England.

Congdon, P., 2003. *Applied Bayesian Modelling*. John Wiley&Sons,England.

Cox, D.R., 1972. Regression models and life tables. Journal of the Royal Statistical Society, 34, 187-220

Dalpatadu, R., Gewali, L., Singh, A.K., 2002. Computing the Bayesian Highest Posterior Density Credible Sets For the Lognormal Mean, Environmetrics, 13, 465-472.

Demirhan, H., 2004. Logaritmik Doğrusal Modellerde Parametrelerin ve Beklenen Göze Sıklıklarının Bayesci Kestirimi, Yayınlanmamıs Bilim UzmanlığıTezi, Hacettepe Üniversitesi, Fen Bilimleri Enstitüsü, Ankara.

Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society Ser. B, 39, 1-38.

Efron, B., 1975. The efficiency of Cox's likelihood function for censored data. Jornal of the American Statistical Association, 72, 557-565.

Ellison, A., E., 2004. Bayesian inference in ecology. Ecol. Lett, 7, 509-520.

Enders, C.K., 2010. *Applied Missing Data Analysis*. Guilford Pres, s:165-286, New York.

Gamerman, D., 1997. *Markov Chain Monte Carlo Stochastic Simulation for Bayesian Inference*. 235 s, Chapman and Hall, London.

Garcia, R.I., Ibrahim, J.G., Zhu, H., 2010, Variable Selection in the Cox Regression Model with Covariates Missing at Random, Biometrics 66, 97-104.

Gelfand, A.E., Hills, S.E., Racine-Poon, A., Smith, A.F.M., 1990. Illustration of Bayesian Inference in Normal Data Models Using Gibbs Sampling. Journal of the American Statistical Association, 85, 972–985.

Gelfand, A., Smith, A.F.M., 1990. Sampling-based approaches to calculating marginal densities. Journal of the American Statistical Association, 85, 398-409.

Geman, S., Geman, D., 1984. Stochastic Relaxation, Gibbs Distribution, and the Bayesian Restoration of Images. IEEE Transaction on Pattern Analysis and Machine Intelligence, 6, 721–741.

Geweke, J., 1992. Evaluating the accuracy of sampling based approaches to the calculation of posterior moments. Bayesian Statistics 4 (164-193), Oxford, UK: Oxford University Pres.

Gilks, W.R., Richardson, S., Spiegelhalter, D.J., 1996. *Markov Chain Monte Carlo in Practice*. Chapman and Hall, London.

Haitovsky, Y., 1968. Missing Data in Regression Analysis. Journal of the Royal Statistical Society Ser. B, 30, 67-82.

Hemming, K., Hutton, J.L., 2010, Bayesian sensitivity models for missing covariates in the analysis of survival data, Journal of Evaluation in Clinical Practice, ISSN 1356-1294.

Hosmer, D.W., Lemeshow, S., 1999. *Applied Survival Analysis: Regression Modeling of Time to Event Data*. Wiley, John and Sons, Incorporated, 408s.

Ibrahim, J.G., Chen, M.H., Sinha, D., 2001. *Bayesian Survival Analysis*. Springer-Verlag, s: 290-317, New York.

Ibrahim,J.G., Chen, M.H., Kim, S.,2008, Bayesian variable selection for the Cox regression model with missing covariates, Lifetime Data Anal. 14(4), 496-520.

Jeffreys, H., 1961. *Theory of Probability*. Third Edition, Oxford: Oxford University Press. 105s.

Kalbfleisch, J.D., Prentice, R.L., 1980. *The Statistical Analysis of Failure Time Data*. Wiley, 219 s, New York.

Kleinbaum, D.G., Klein, M., 1996. *Survival Analysis*, *A Self Learning Text*. Springer, 124 s, USA.

Kurt, İ., 2008. Bayesgil Yaşam Analizi ve Cox Regresyon Yaşam Analizinin Türetilmiş ve Gerçek Veri Setlerinde Uygulanması. Doktora Tezi, Eskişehir Osmangazi Üniversitesi Sağlık Bilimleri Enstitüsü, Eskişehir, 101s.

Lam, P.T., Leung, M.W., Tse, C.Y., 2007. Identifying prognostic factors for survival in advanced cancer patients: a prospective study. Hong Kong Med J, 13, 453-459.

Lee, E.T., 1992. *Statistical Methods for Survival Data Analysis*. Second Edition, John Wiley&Sons. Inc., 482s, New York.

Lee, E.T., Wang, J.W., 2003. *Statistical Methods for Survival Data Analysis*, John Wiley & Sons, s: 82-175, Canada.

Little, R.J.A., 1992. Regression with missing X's: a review. Journal of the American Statistical Association, 87, 1227-1237.

Little, R.J., Rubin, D.B., 2002. *Statistical Analysis With Missing Data*. Wiley&Son, s: 99-145, New Jersey.

Mohan, A., Goyal, A., Singh, P., Singh, S., Pathak, A.K., Bhutani, M., Pandey, R.M., Guleria, R., 2006. Survival in small cell lung cancer in India: Prognostic utility of clinical features, laboratory parameters and response to treatment. Indian Journal of Cancer, 43(2), 67-74.

Rubin, D.B., 1976. Inference and Missing Data. Biometrika, 63, 581–592.

Rubin, D.B., 1987. *Multiple Imputation for Nonresponse in Surveys*. Wiley&Sons, 303s, New York.

Rubin, D.B., 1996. Multiple Imputation After 18+ Years (with discussion). Journal of the American Statistical Association, 91, 473-489.

Sahlin, K., 2011. Estimating convergence of Markov Chain Monte Carlo Simulations. Stockholm University, Mathematical Statistics, Master Thesis.

SAS Institute, 2006. Preliminary capabilities for bayesian analysis in SAS/STATR Software, SAS Institute Inc., Cary, NC, USA.

SAS Institute, 2011. SAS/STAT 9.3 User's Guide, SAS Instute INC., Cary, NC, USA.

Schafer, J.L., 1997. *Analysis of Incomplete Multivariate Data*. Chapman&Hall, s: 67-117, London.

Schafer, J.L., Olsen, M.K., 1998. Multiple imputation for multivariate missing data problems: A data analyst's perspective. Multivariate Behavioral Research, 33, 545-571.

Spiegelhalter, D.J., Best, N.G., Carlin, B.P., 1998. Bayesian deviance, the effective number of parameters and the comparison of arbitrarily complex model. (http://yaroslavvb.com/papers/spiegelhalter-bayesian.pdf) (01.03.2012).

Spiegelhalter, D.J., Best, N.G., Carlin, B.P., Van der Linde, A., 2002. Bayesian measures of complexity and fit. Journal of Royal Stat. Soc. B, 64, 583-540.

Walsh, B., 2002. Markov Chain Monte Carlo and Gibbs Sampling, lecture notes for EB 596z, http://nitro.biosci.arizona.edu/courses/EEB596/handouts/ gibbs.pdf, erisim tarihi: 2005.

Wong, M.C.M., Lam, K.F., Lo, E.C.M., 2005, Bayesian analysis of clustered interval-censored data, J Dent Res 84(9), 817-821.

Yin, G., Ibrahim, J.G., 2006, Bayesian transformation hazard model, IMS Lecture Notes-Monograph Series 2nd Lehmann symposium- Optimality Vol. 49, 170-182.

ÖZGEÇMİŞ

Adı Soyadı :Nesrin ALKAN

Doğum Yeri :Mersin

Doğum Tarihi :02.12.1979

Medeni Hali **:**Evli

Bildiği Yabancı Diller **:**İngilizce

Eğitim Durumu (Kurum ve Yıl)

Lise **:**Özel İçel Lisesi (1994-1997)

Lisans **:**Ondokuz Mayıs Üniversitesi, Fen Edebiyat Fakültesi, İstatistik Bölümü, (1997-2001)

Yüksek Lisans **:**Ondokuz Mayıs Üniversitesi, Fen Bilimleri Enstitüsü, İstatistik Anabilim Dalı (2003-2006)

Çalıştığı Kurum/Kurumlar ve Yıl:

- Ondokuz Mayıs Üniversitesi, Sinop Fen Edebiyat Fakültesi, İstatistik Bölümü (2002-2007)
- Sinop Üniversitesi, Fen Edebiyat Fakültesi, İstatistik Bölümü (2007-…)

İletişim Bilgileri: Sinop Üniversitesi, Fen Edebiyat Fakültesi, İstatistik Bölümü, 57000, Sinop, Türkiye