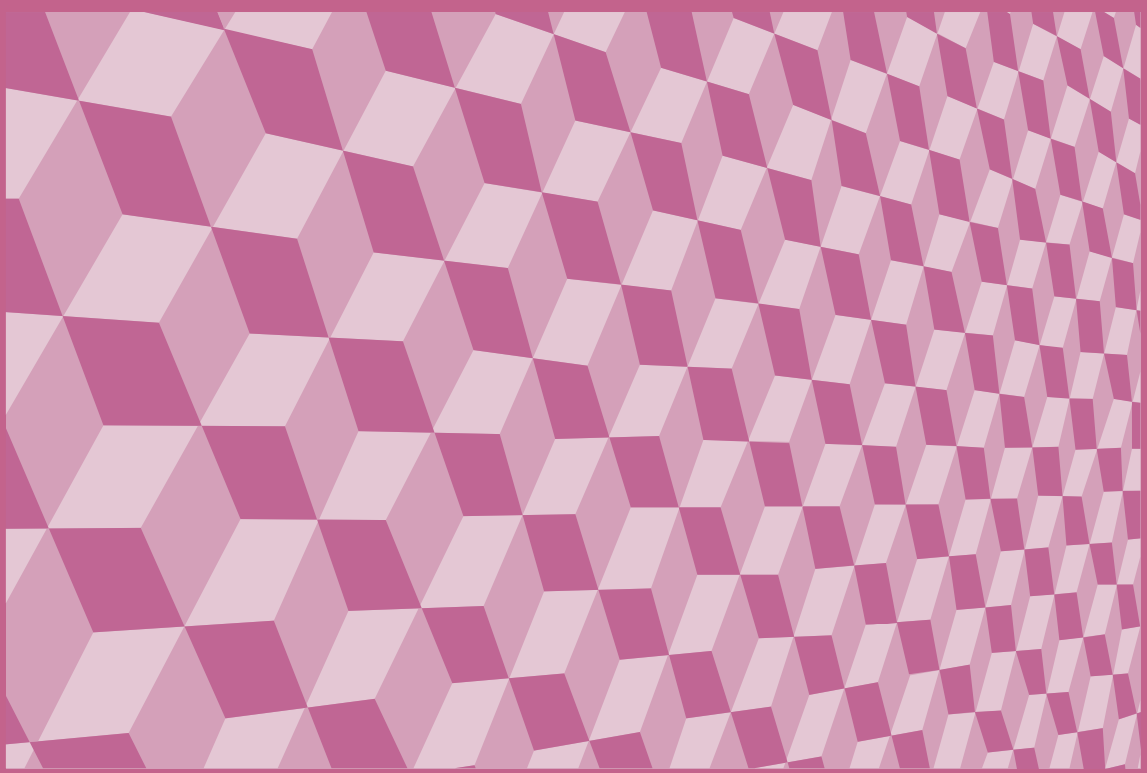
İSTATİSTİK ARAŞTIRMA DERGİSİ

Journal of Statistical Research

Cilt-Volume: 10 Sayı-Number: 02
Temmuz-July 2013

ISSN 1303-6319



TÜRKİYE İSTATİSTİK KURUMU
Turkish Statistical Institute

İSTATİSTİK ARAŞTIRMA DERGİSİ
Journal of Statistical Research

Cilt-Volume: 10 Sayı-Number: 02
Temmuz-July 2013

TÜRKİYE İSTATİSTİK KURUMU
Turkish Statistical Institute

Yayın istekleri için For publication order

Döner Sermaye İşletmesi Revolving Fund Management

Tel: + (312) 425 34 23 - 410 05 96 - 410 02 85
Faks-Fax**:** + (312) 417 58 86

Yayın içeriğine yönelik sorularınız için For questions about contents of the publication

Dergi Editörlüğü Journal Editorship

Tel: +90 (312) 410 03 67 - 233 13 63
Faks-Fax**:** +90 (312) 425 34 05

İnternet Internet
http://www.tuik.gov.tr http://www.turkstat.gov.tr

E-posta E-mail
dergi@tuik.gov.tr journal@tuik.gov.tr

Yayın No Publication Number
4165

ISSN
1303-6319

Türkiye İstatistik Kurumu Turkish Statistical Institute

Yücetepe Mah. Necatibey Cad. No: 114 06100 Çankaya-ANKARA / TÜRKİYE

Türkiye İstatistik Kurumu Matbaası, Ankara Turkish Statistical Institute, Printing Division, Ankara
Tel: +90 (312) 410 01 64 * Fax: +90 (312) 418 50 82
Haziran 2014 June 2014
MTB: 2014-349 - 500 **Adet**-Copies

Editör Notu

Değerli Okuyucular,

Türkiye İstatistik Kurumu tarafından 2001 yılından bu yana hakemli olarak yürütülmekte olan "İstatistik Araştırma Dergisi" ile istatistiki araştırmaların niteliğinin yükseltilmesi, kuramsal ve uygulama alanındaki araştırmacılar arasında iletişimin ortak çalışma ve yayınlarla güçlenmesi sağlanmaya çalışılmaktadır.

Evrensel bilimin paylaşılmasını sağlayan bilimsel dergilerin temel işlevi; bilimsel makale yazarının çalışmasını en etkin biçimde ifade etmesine yardımcı olmak ve bilimi anlaşılabilir bir biçimde yayınlamaktır.

Akademisyen, araştırmacı ve okuyucuların artan ilgisine paralel olarak bizlerin çabası, azmi ve kararlılığı da artacak olup, dergimiz daha üst seviyelere taşınacaktır. Dergimizin ulusal ve uluslararası endekslerde taranması çalışmaları da devam etmektedir. Bu kapsamda TÜBİTAK ULAKBİM'e on-line başvuru yapılmış olup, sonuç beklenmektedir. Bu konuya ilişkin olarak alınacak sonuçlar sizlerle paylaşılacaktır.

Bu sayımızda, kavramsal, kuramsal ve uygulamalı çalışmalar olmak üzere toplam beş adet çalışmayı siz değerli okuyucularımızla paylaşmanın gururunu taşıyoruz. Bu değerli çalışmaları, bizlerle ve siz değerli okuyucularımız ile paylaşan sayın yazarlara teşekkür ederiz. Ayrıca çalışmaların daha nitelikli hale gelmesinde çok değerli öneri, eleştiri ve katkılarını esirgemeyen sayın hakemlere de şükranlarımızı sunuyoruz.

Dergi'nin basım aşamasına gelmesinde emeğini ve desteklerini esirgemeyen TÜİK Başkanı Sayın Birol AYDEMİR'e, derginin her aşamasında emeği geçen Editör Yardımcısı Sayın Doç. Dr. Özlem İLK DAĞ'a, dergi çalışmalarını içtenlikle ve azimle yürüten Dergi Sekreteryası'na ve son olarak da emeği geçen diğer tüm TÜİK çalışanlarına teşekkürlerimi iletmek isterim.

Bu sayımızın da akademisyenler ile araştırmacılara faydalı olması temennisi ve gelecek sayılarda hedeflenenler ölçüsünde tekrar buluşmak dileği ile saygılar sunarım.

Prof. Dr. Fetih YILDIRIM
Dergi Editörü

TÜRKİYE İSTATİSTİK KURUMU
İSTATİSTİK ARAŞTIRMA DERGİSİ

TURKISH STATISTICAL INSTITUTE
JOURNAL OF STATISTICAL RESEARCH

Sahibi **Owner**
Türkiye İstatistik Kurumu Adına
Birol AYDEMİR
Türkiye İstatistik Kurumu Başkanı

On Behalf of Turkish Statistical Institute
Birol AYDEMİR
President, Turkish Statistical Institute

Editör **Editor**
Prof. Dr. Fetih YILDIRIM

Prof. Dr. Fetih YILDIRIM

Editör Yardımcısı **Assistant Editor**
Doç. Dr. Özlem İLK DAĞ

Assoc. Prof. Özlem İLK DAĞ

Sekreterya **Secretariat**
Buket AKGÜN
Z.Nur EMRE
Nurdan ELVER

AMAÇ VE KAPSAM

"*İstatistik Araştırma Dergisi (İAD)*", istatistik araştırmaların niteliğinin yükseltilmesi, istatistik yöntem ve uygulamalarının geliştirilmesi, literatürde yer alan çalışmaların tartışılması, istatistik uygulamalarıyla ilgili anket çalışmalarının ele alınması, kuramsal ve uygulama alanındaki araştırmacılar arasında iletişimin ortak çalışma ve yayınlarla güçlendirilmesi amacıyla, yayımlanan hakemli bir dergidir.

"*İstatistik Araştırma Dergisi*"nin kapsamında yer alan tematik konular aşağıda özet olarak verilmiştir:

• Bankacılık, Finans, Sigortacılık, Aktüerya ve Risk Yönetimi; Bayesci İstatistik; Benzetim Teknikleri; Bilgi Sistemleri; Biyoistatistik; Bulanık Teori; Demografi; Deney Tasarımı ve Varyans Analizi; Ekonometri; Genel Sayımlar ve Değerlendirmeleri; İstatistik Eğitimi; İstatistik Etiği; İstatistik Kuramı; İstatistiksel Kalite Kontrolü; Kamuoyu ve Piyasa Araştırmaları; Klinik Denemeler; Mühendislikte İstatistik Uygulamaları; Olasılık ve Stokastik Süreçler; Optimizasyon; Örnekleme ve Araştırma Tasarımları; Parametrik Olmayan İstatistiksel Yöntemler; Resmi İstatistikler; Toplum Bilimlerinde İstatistik; Veri Analizi ve Modelleme; Veri Madenciliği; Veri Yönetimi ve Karar Destek Sistemleri; Verimlilikte İstatistiksel Yaklaşımlar; Yönetsel Süreçlerde Performans Analizi; Yöneylem Araştırması; Zaman Serileri; Diğer İstatistiksel Yöntemler gibi istatistiğin her dalında yeni bilgi üretimine yönelik tüm araştırmalar.

Makale Dili ve Genel Kurallar

- Bu yayının 5846 Sayılı Fikir ve Sanat Eserleri Kanunu'na göre her hakkı Türkiye İstatistik Kurumu Başkanlığı'na aittir. Gerçek veya tüzel kişiler tarafından izinsiz çoğaltılamaz ve dağıtılamaz.
- Makale taslakları WORD yazım dilinde, Times New Roman yazı tipinde, 12 punto büyüklükte, satırlar arasında bir satır boşluk bırakılarak yazılmalı, şekil ve grafikler JPG dosyaları olarak hazırlanmalıdır.
- A4 sayfa boyutunda; soldan 3,5 cm, sağdan, yukarıdan ve aşağıdan 2,5 cm boşluk bırakılmalıdır.
- Ana bölüm başlıklarının tümü büyük harf, 12 punto büyüklükte, koyu, ortalı ve Arap rakamları ile numaralandırılarak; alt bölüm başlıklarında ise sadece kelimelerin baş harfleri büyük diğerleri küçük harfle, 12 punto büyüklükte, koyu, sola dayalı ve ana bölüm başlığına endeksli olarak Arap rakamları ile numaralandırılarak yazılmalıdır.
- Makale taslağı yazımında, okuyucunun, çalışmanın her aşamasını anlama ve değerlendirmesine olanak verecek bir anlatım ve plana uyulmalıdır.
- Anlatım olabildiğince sade, anlaşılabilir, öz ve kısa olmalıdır. Gereksiz tekrarlardan, desteklenmemiş ifadelerden ve konu ile doğrudan ilişkisi olmayan açıklamalardan kaçınılmalıdır.
- Yazımda çok genel ifadeler kullanılmamalıdır. Yargı veya kesinlik içeren ifadeler mutlaka verilere/ referanslara dayandırılmalıdır.
- Araştırmacı/araştırmacılar tarafından probleme, hangi kuramsal/kavramsal açıdan yaklaşıldığı, gerekçeleri ile birlikte belirtilmelidir.
- Kullanılan araştırma yönteminin seçilme gerekçesi açıklanmalıdır. Bütün veri toplama araçlarının geçerliliği ve güvenilirliği belirtilmelidir.
- Araştırma sonucunda elde edilen veriler bir bütünlük içinde sunulmalıdır.
- Sadece elde edilen verilere dayanan sonuçlar sunulmalıdır.
- Sonuçların yorumları, varsa, literatürdeki diğer kaynaklarla desteklenerek, değerlendirilmelidir.
- Yararlanılan kaynaklar, çalışmanın kapsamını yansıtacak zenginlik ve yeterlikte olmalıdır.
- Türkçe ve İngilizce özetler; çalışmanın amacı, yöntemi, kapsamı ve temel bulgularını içermelidir.

Ayrıntılı bilgi için, http://www.tuik.gov.tr adresinden *"İstatistik Araştırma Dergisi Kılavuzu"*na bakınız.

AIM AND SCOPE

"Journal of Statistical Research (JSR)" is a refereed journal published with the aim to raise the quality of statistical researches, improve the statistical methodology and applications, discuss the studies included in literature, consider survey studies regarding the statistical application, and strengthen the communication between researchers in the field of theory and application by joint studies and publications.

The contents of the *"Journal of Statistical Research"* are summarized below:

• Researches aimed at producing new knowledge in every field of statistics such as Banking, Finance, Insurance Trade, Actuarial and Risk Management; Bayesian Statistics; Biostatistics; Clinic Tests; Data Analysis and Modeling; Data Management and Decision Support Systems; Data Mining; Demography; Econometrics; Experimental Design and Variance Analysis; Fuzzy Theory; General Census and Evaluation; Information Systems; Non-Parametric Statistical Methods; Official Statistics; Operational Research; Optimization; Sampling and Research Designs; Performance Analysis in Managerial Process; Probability and Stochastic Processes; Public Opinion and Market Researches; Statistical Applications in Engineering; Statistical Approaches in Efficiency; Statistical Ethics; Statistical Quality Control; Statistical Training; Statistics in Social Science; Statistics Theory; Simulation Techniques; Time Series; Other Statistical Methods.

Article Language and General Rules

• Turkish Statistical Institute reserves all the rights of this publication. Unauthorized duplication and distribution of this publication is prohibited under Law No: 5846.
• Article drafts should be prepared in WORD, using Times New Roman font, in 12 point size, with a blank line in between lines. Figures and tables should be prepared as JPG files.
• On A4 paper size; margins should be set as: left 3,5 cm; right, top and bottom 2,5 cm.
• Titles of the main sections should be all capitalized, in 12 point size, bold, centered and numbered with Arabic numerals; only the first letter of the words in the titles of the subsections should be capitalized, with 12 point size, bold, left justified and numbered with Arabic numerals indexed to the titles of the main sections.
• In article draft writing, writer should follow such a plan that reader should be able to understand and evaluate all the steps of the study.
• Narration should be as plain as possible, as well as comprehensible, compact and short. Unnecessary repetitions, unsupported declarations and explanations that are not in direct relation to the topic should be avoided.
• General statements should be avoided in writing. Statements that include judgment or facts must be supported by data/references.
• It should be stated, with justifications, from which theoretical/conceptual aspect the researcher/researchers have approached the problem.
• The reason of choosing the research methodology that is used should be explained. The validity and reliability of all the data collection tools should be presented.
• Data obtained as the result of the research should be presented in unity.
• Results that only rely on the obtained data should be presented.
• The interpretation of the results should be supported and evaluated by the other resources, if any, in the literature.
• Used resources should be in good wealth and proficiency that reflect the scope of the study.
• Turkish and English abstracts should include the goal, methodology, scope and main findings of the study.

For detailed information, please see "A Guide for Journal of Statistical Research" at http://www.tuik.gov.tr.

DERGİNİN BU SAYISINA BİLİMSEL KATKI SAĞLAYAN HAKEMLER
REFEREES WHO PROVIDED SCIENTIFIC CONTRIBUTIONS FOR THIS VOLUME OF THE JOURNAL

Doç. Dr	Ayten YİĞİTER	Hacettepe Üniversitesi
Doç. Dr	Fikret Er	Anadolu Üniversitesi
Doç. Dr	Halil AYDOĞDU	Ankara Üniversitesi
Doç. Dr.	Mehmet Ali ERYURT	Hacettepe Üniversitesi
Doç. Dr.	Mehtap AKÇİL OK	Başkent Üniversitesi
Doç. Dr	Meral SUCU	Hacettepe Üniversitesi
Doç. Dr	Özlem İLK DAĞ	Orta Doğu Teknik Üniversitesi
Doç. Dr.	Vilda PURUTÇUOĞLU	Orta Doğu Teknik Üniversitesi
Yrd. Doç. Dr	B.Burçak BAŞBUĞ ERKAN	Orta Doğu Teknik Üniversitesi
Yrd. Doç. Dr.	Necla GÜNDÜZ TEKİN	Gazi Üniversitesi
Yrd. Doç. Dr.	Rukiye DAĞALP	Ankara Üniversitesi

Volume: 10 Number: 02 Category: 02 Page: 1-24 ISSN No: 1303-6319 Cilt: 10 Sayı: 02 Kategori: 02 Sayfa: 1-24

TÜRKİYE GENELİ ÖLÜM VERİLERİ KULLANILARAK YAŞAM TABLOSUNUN OLUŞTURULMASI

Hanife TAYLAN* **Güçkan YAPAR****

ÖZET

Bu çalışmada amaç değişen veri kayıt sistemi ile TÜİK tarafından 2009 yılı itibariyle yayımlanmaya başlanan Türkiye geneli ölüm verileri kullanılarak yaş grubu ve cinsiyet bazında Türkiye Dönem Yaşam Tablosunun oluşturulmasıdır. Yaşam tablosu elde edilirken Türkiye geneli ölüm verileri ve Adrese Dayalı Nüfus Kayıt Sisteminden elde edilen nüfus verileri kullanılmıştır. İlk olarak belirli bir zaman aralığı için yaşa özel ölüm hızları elde edilmiştir. Daha sonra her bir yaş grubu için yaşa özel ölüm ve yaşam olasılıkları ile beklenen yaşam süreleri Türkiye nüfusu için elde edilmiştir. Bunun sonucunda 2011 yılında kadınlarda beklenen yaşam süresinin 79.4, erkeklerde ise 74.1 olduğu görülmüştür. Yapılan çalışmalar sonucunda Türkiye nüfusunun giderek yaşlandığı ve kadınların erkeklerden ortalama 5.3 yıl daha uzun yaşadıkları ve yıllar itibariyle bu farkın hızla azalmaya başladığı görülmüştür. Son olarak elde edilen sonuçlar değerlendirilmiş ve tartışılmıştır.

Anahtar Kelimeler: Beklenen yaşam süresi, Özetlenmiş dönem yaşam tablosu, Yaşa özel ölüm olasılığı, Yaşam tablosu.

1. GİRİŞ

Yaşam tabloları demografik veriler kullanılarak oluşturulan tablolardır. Yaşa göre gerçek ölüm sayıları yaşam tablolarında belirli bir süre içinde hayatta kalan ve ölen bireylerin sayılarından faydalanılarak zaman ve yaş aralığı bazında ölüm ve yaşam olasılıklarının elde edildiği tablolardır. Hayat Sigortaları Matematiği için temel oluşturan yaşam tablosu ilk kez (John Graunt, 1662); (Edmont Halley, 1693) tarafından oluşturulmuştur. Edmont Halley tarafından yapılan çalışmada yaşa özgü ölüm sayıları kullanılmıştır. Yaşam tablosu, yaşa özgü ölen ve yaşayan kişi sayıları kullanılarak demografik yaklaşımlar yardımıyla oluşturulur (Shryock, Siegel, ve Associates, 1971). Aktüeryal/demografik yaşam tablolarının oluşturulmasında kullanılan yöntemler sırasıyla (Reed ve Merrel, 1939; Greville, 1943; Chiang, 1968, 1972; Fergany, 1971; and Keyfitz and Frauenthal, 1975) tarafından geliştirilmiştir. Bu çalışmalarda ölüm olasılığı belirli bir zaman aralığı için yaşanılan ortalama kişi yıl sayısı ve ölüm hızı kullanılarak hesaplanmıştır.

Yaşam tabloları dönem ve kuşak yaşam tabloları olmak üzere iki tipte oluşturulur (Preston vd., 2001). Kuşak yaşam tablosu aynı zaman aralığında doğan tüm bireylerin her bir üyesinin ölümüne kadar geçen sürenin modellenmesi ile oluşturulan tablolara denir. Dönem yaşam tabloları ise belirli bir zaman aralığında yaşayan mevcut nüfusun yaşam sürelerinin modellenmesiyle oluşturulur. Dönem yaşam tabloları gelişmiş ülkelerde bir yıllık ya da 3 yıllık tablolar halinde düzenli olarak ilgili birimler tarafından yayımlanır. Birleşik Amerika Devletleri'nde Hastalık Korunma ve Önleme Merkezi (NCHS) tarafından her yıl kuşak ve dönem yaşam tabloları yaş, cinsiyet ve ırka göre

*Arş. Gör., Dokuz Eylül Üniversitesi, Fen Fakültesi, İstatistik Bölümü, İzmir, e-posta: hanife.taylan@deu.edu.tr
**Doç. Dr., Dokuz Eylül Üniversitesi, Fen Fakültesi, İstatistik Bölümü, İzmir, e-posta: guckan.yapar@deu.edu.tr

oluşturulur ve yayımlanır (Arias, 2010). Benzer şekilde Avustralya'da ABS tarafından, Birleşik Krallıkta Ulusal İstatistik Ofisi (ONS) tarafından, Yeni Zelanda'da Yeni Zelanda İstatistik Merkezi (SNZ) tarafından belirli periyotlar halinde yayımlanmaktadır. Kuşak yaşam tabloları ise bir kuşağın ölüm davranışlarını incelediğinde yaklaşık yüz yıllık ölüm ve yaşam verisine ihtiyaç duyar ve bu çoğu ülke için henüz mevcut olmayan bir veri setidir. Ülkemizde ise ölüm ve yaşam verileri TÜİK tarafından yayımlanmaktadır. Bu verilerin doğru ve tüm nüfusu yansıtacak şekilde yayımlanması gerekir. 2009 yılı itibariyle ölüm verileri Türkiye geneli için yayımlanmaya başlanmıştır ve kayıt sistemindeki bu değişiklik ülkemiz için dönem yaşam tablolarının doğru bir şekilde oluşturulmasına olanak tanımıştır.

Yaşam tablolarının barındırdığı en önemli parametre olan beklenen yaşam süresi, demografik olarak bir gelişmişlik göstergesi olmasının yanında sigortacılık sektörü içinde önemli bir bilgidir. Gelecekte bireylerin ortalama ne kadar yaşayacağının bilinmesi sosyal güvenlik sistemi ve özel sigortacılık sektöründe prim, rezerv, teminat hesaplamaları, emeklilik sürelerinin belirlenmesi ya da ürün oluşturma gibi finansal yönetimin temelini oluşturan alanlarda kullanılır. Günümüzde yaşam koşullarının iyileşmesi, tıp ve teknolojik alanda yaşanan gelişmeler demografik yapının değişmesine yol açmıştır. Bu gelişmeler ise uzun ömürlülüğe ve yaşlı nüfusun artmasına neden olmuştur. Yaşlı nüfusun artması ve uzun ömürlülük ise çalışma süresinin uzunluğu, emeklilik yaşının yükselmesi ve emeklilik maaşının daha uzun periyotta ödenmesi, genç nüfusun üzerine yük binmesi ve sosyal problemlerin artması gibi sorunları da beraberinde getirmiştir. Değişen bu yapıya uyum sağlamak, sektörün yanlış hesaplamalardan korunması ve gelecekte karşılaşabileceği riski yönetebilmesi açısından toplumun demografik yapısını yansıtan yaşam tablolarının sektörde kullanılması büyük önem taşımaktadır. Gelişmiş ülkeler kendi yaşam tablolarını oluştururken ülkemizde hala başka ülkelerin yaşam tabloları kullanılmaktadır. Bu çalışma bu eksikliğin ortadan kaldırılması ve yapılacak çalışmalar için örnek teşkil etmesi amacıyla oluşturulmuştur.

Tuzgöl, 2005, tarafından yapılan çalışmada SSK ölüm istatistikleri incelenmiş ve çalışmakta olan sigortalılar, maluliyet aylığı ve yaşlılık aylığı alan gruplar için 2000-2003 yılı dönem yaşam tablosu ayrı ayrı elde edilmiştir. Yine elde edilen beklenen yaşam sürelerine göre diğer ülkelerin yaşam tablolarıyla karşılaştırılmış ve demografik yapının farklılığının yapılan hesaplamalardaki etkileri incelenmiştir.

2009 yılı öncesinde Türkiye için yaşam tablosu birçok kez farklı çalışmalar yapılarak oluşturulmuştur. Bu konuda yapılan önemli çalışmalar sırasıyla (Alpay, 1969; Özsoy, 1970; Öcal, 1974; Demirci, 1987; Duransoy, 1993; Hoşgör, 1992, 1997; Toros, 2000; Demirbüken, 2001; Coşkun, 2002; Kıkbeşoğlu, 2006; Eryurt ve Koç, 2010) olmuştur. Ülkemizde bu konu ile ilgili yapılan son çalışma ise Türkiye Hayat ve Hayat Anüite Tablolarının oluşturulması projesidir. Bu proje Hacettepe Üniversitesi Fen Fakültesi Aktüerya Bilimleri Bölümü'nün yöneticiliğinde, BNB Danışmanlık Şirketi, Marmara Üniversitesi ve Başkent Üniversitesi uzmanları ile birlikte gerçekleştirilmiştir. Türkiye 2010 yılı Kadın ve Erkek Hayat ve Hayat Anüite Tabloları oluşturulmuştur. Çalışmanın kaynağını TÜİK'ten 1927 den 2000 yılına kadar yapılan 15 nüfus sayımının yaş ve cinsiyet dağılımındaki verisi oluşturmuştur. Farklı nüfus sayımlarına Preston-Bennet yöntemi uygulanarak elde edilen model hayat tablo düzeylerindeki değişim regresyon modelleriyle açıklanmıştır. Sonuç bölümünde bu çalışmada oluşturulan Türkiye Kadın-Erken Hayat Tablosu (TRH-2010) ile projede oluşturulan hayat tabloları karşılaştırılacaktır. Bu çalışmada, Türkiye'de ve dünyada yapılan çalışmalar ve yenilenen veri kayıt sisteminin sağladığı olanaklar dikkate alınmıştır. Bu doğrultuda

ülkenin kendi demografik özelliklerini yansıtan yaşam tablosuna duyduğu ihtiyaç amacıyla Türkiye için Özetlenmiş Dönem Yaşam Tablosu cinsiyet bazında oluşturulmuştur.

2. DÖNEM YAŞAM TABLOSUNUN OLUŞTURULMASI

Yaşam Tabloları bir nüfusun ölen ve yaşayan kişi sayılarından faydalanılarak oluşturulan geleneksel bir yöntemdir ve birçok çalışmaya konu olmuştur (Chiang, 1972; Kintner, 2004; Keyfitz ve Caswell, 2005; Bell ve Miller, 2005 vb). Bir ülkenin belirli bir zaman aralığındaki mevcut demografik yapısını yansıtan ve bu aralıktaki yaşa özel ölüm hızları baz alınarak oluşturulan yaşam tablosuna dönem yaşam tablosu denir (Pfaff vd, 2012). Bu bölümde 2009, 2010 ve 2011 yılları için ölüm ve yaşam verileri kullanılarak Türkiye Özetlenmiş Dönem Yaşam Tablosu cinsiyet ve yaş grupları bazında elde edilmiştir. Bu çalışmanın yöntemi aşağıda anlatılmıştır. Öncelikle demografi biliminde hız herhangi bir olayın (ölüm, doğum, göç vb.) gerçekleşme sayısının bu riske maruz kalan kişi yıl sayısına bölümüyle elde edilir. Tablonun oluşturulması için tanımlanacak ilk değişken yaşa özel ölüm hızı olarak adlandırılan, ${}_nM_x$, belirli bir zaman aralığında x ve $x+n$ yaşları arasında ölen kişi sayısının yine aynı yaş aralığında yaşanılan kişi yıl sayısına bölünmesiyle elde edilir (Arias, 2010).

$$ {}_nM_x = \frac{{}_nD_x}{{}_nN_x} \approx {}_nm_x \qquad (1) $$

${}_na_x$ ile gösterilen ortalama yaşanan kişi yıl sayısı dört farklı yönteme göre hesaplanabilir (Preston vd., 2001). Bu çalışmada gerçek gözlemler kullanılmış yani ortalama yaşanan kişi yıl sayısı ölüm sayılarının her bir yaş grubu için ağırlıklı ortalamaları alınarak elde edilmiştir.

$$ {}_na_x = \frac{d_x.(0) + d_{x+1}.(1) + d_{x+2}.(2) + d_{x+3}.(3) + d_{x+4}.(4)}{{}_nd_x} \quad , \quad n = 5 \qquad (2) $$

Yaşa özel ölüm olasılığı, ${}_nq_x$, x yaşında bir kimsenin n yıl içindeki ölüm olasılığını ifade eder ve yaşa özel ölüm hızı ve yaşanan ortalama kişi yıl sayısı değişkenleri kullanılarak elde edilir. Bu dönüşüm gerçek kuşak davranışlarını baz alır ve aşağıdaki gibi elde edilir.

$$ {}_nL_x = n.l_{x+n} + {}_na_x \,.\, {}_nd_x \rightarrow {}_nq_x = \frac{{}_nd_x}{l_x} = \frac{n.{}_nm_x}{1+(1-{}_na_x)\,{}_nm_x} $$

${}_nL_x$: Bir kuşakta x ve x+n yılları arasında yaşanan kişi yıl sayısı (kuşak davranışlarını baz alır.)

$n.l_{x+n}$: Bir kuşakta yaşayan kişilerden bu aralıkta yaşayan kişi yıl sayısı

${}_na_x$: Yaşanan ortalama kişi yıl sayısı

${}_nd_x$: Bir kuşak nüfusundan bu aralıkta ölen kişilerin sayısı

Yaşayan kuşak kişi yıl sayısı aşağıdaki gibi yeniden düzenlenir:

$$
\begin{aligned}
{}_nL_x &= n\left(l_x - {}_nd_x\right) + {}_na_x \cdot {}_nd_x \\
n.l_x &= {}_nL_x + n.{}_nd_x - {}_na_x \cdot {}_nd_x \\
l_x &= \frac{1}{n}\left[{}_nL_x + \left(n - {}_na_x\right).{}_nd_x\right]
\end{aligned}
$$

Kuşak davranışları kullanılarak elde edilen l_x yani x yaşında yaşayan bireylerin sayısı ${}_nq_x$ formülünde yerine yazılır

$${}_nq_x = \frac{{}_nd_x}{l_x} = \frac{n.{}_nd_x}{{}_nL_x + \left(n - {}_na_x\right).{}_nd_x}$$

Pay ve payda ${}_nL_x$'e bölünür:

$${}_nq_x = \frac{n.\frac{{}_nd_x}{{}_nL_x}}{\frac{{}_nL_x}{{}_nL_x} + \left(n - {}_na_x\right).\frac{{}_nd_x}{{}_nL_x}} = \frac{n.{}_nm_x}{1 + \left(n - {}_na_x\right).{}_nd_x}$$

Daha sonra yaşa özel ölüm olasılığı aşağıdaki gibi elde edilir:

$${}_nq_x = \frac{n.{}_nm_x}{1 + \left(1 - {}_na_x\right){}_nm_x} \quad \text{and} \quad q_\infty = 1 \qquad (3)$$

Bu dönüşüm Greville (1943) and Chiang (1968) tarafından tanımlanmıştır ve sadece bir parametre gerektirir ve o parametre ise ${}_na_x$'tir. Ortalama yaşanan kişi yıl sayısı dönem yaşam tablolarının oluşturulmasında önemli bir yere sahiptir ve ${}_nm_x \rightarrow {}_nq_x$ dönüşümü ile yaşa özel ölüm olasılıklarının gözlenen verilerden elde edilmesine olarak sağlar (Preston vd., 2001). Daha sonra sırasıyla yaşam tablosu fonksiyonları olan yaşama olasılıkları, ${}_np_x$, yaşanan kişi yıl sayısı, ${}_nL_x$, toplam kişi yıl sayısı, T_x, elde edilmiştir (Selvin, 2008). Yaşam tablosu analizinden faydalanılarak x yaşında bir kimsenin beklenen yaşam ömrü aşağıdaki gibi hesaplanır.

$$\overset{0}{e}_x = \frac{T_x}{l_x}$$

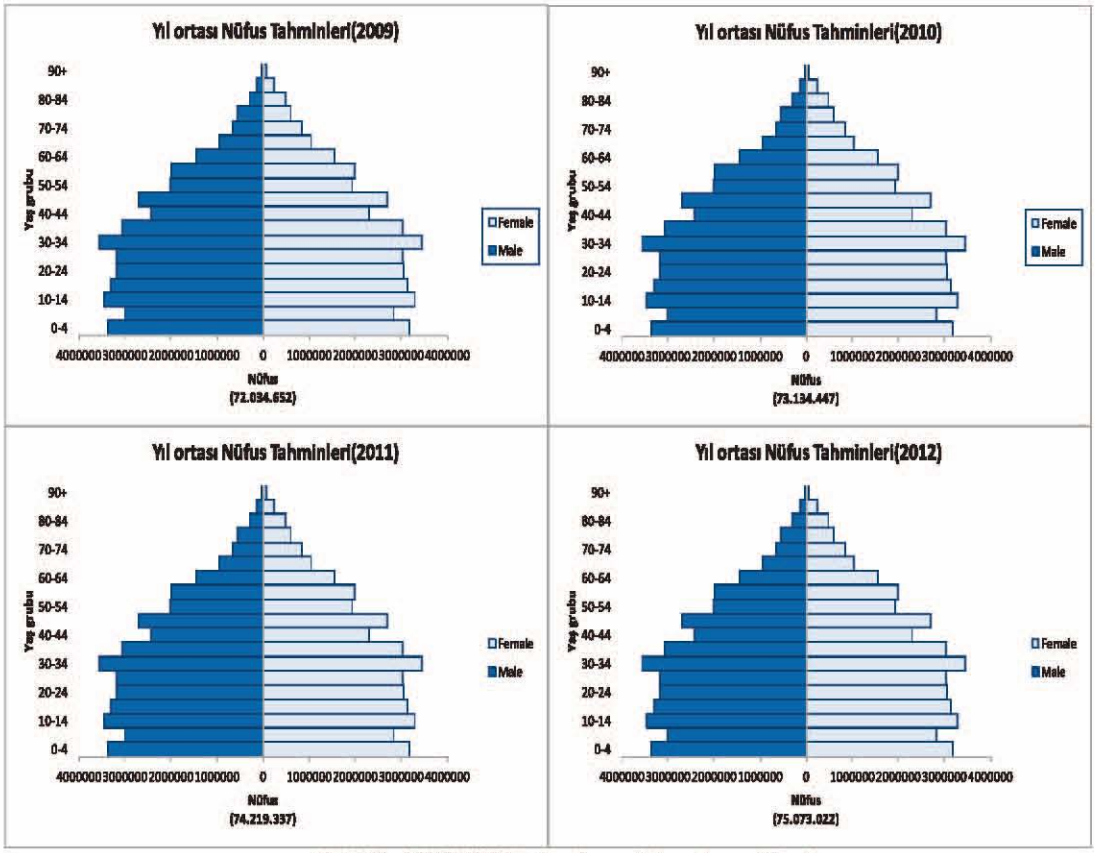
Doğuşta beklenen yaşam ömrü ise aşağıdaki gibi elde edilir.

$$\overset{0}{e}_0 = \frac{T_0}{l_0}$$

3. UYGULAMA

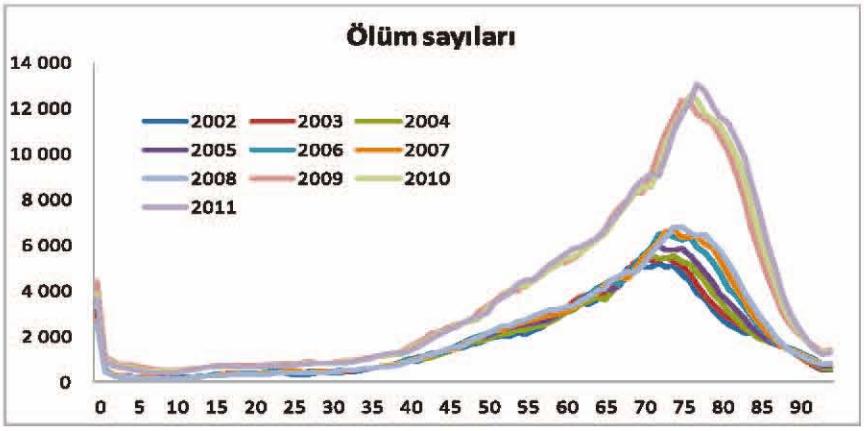
Bu çalışmada TÜİK tarafından yayımlanan 2009, 2010 ve 2011 yılı Türkiye geneli nüfus, ölüm ve doğum verileri kullanılarak 2009, 2010 ve 2011 yılları için cinsiyet bazında Özetlenmiş Dönem Yaşam Tabloları oluşturulmuştur. Adrese Dayalı Nüfus

Kayıt Sistemi'nden elde edilen Türkiye geneli nüfus verileri incelendiğinde yıllık nüfus sayısında artışın olduğu ancak bu artışa karşılık nüfus artış oranında bir azalma olduğu görülmüştür. Yaşam tablosunun oluşturulması için ilk olarak yaşa özel ölüm hızlarının hesaplanması gerekmektedir. Ölüm hızları yaş grupları için o aralıkta ölen kişi sayısının aynı aralıktaki yıl ortası nüfusa bölünmesiyle elde edilir (Bkz., Formül (1)). Nüfus artışının üstel olduğu varsayımı altında yıllık nüfus artış oranları tahmin edilmiş ve yıl ortası nüfuslar cinsiyet bazında aşağıdaki grafikte gösterildiği gibi hesaplanmıştır.



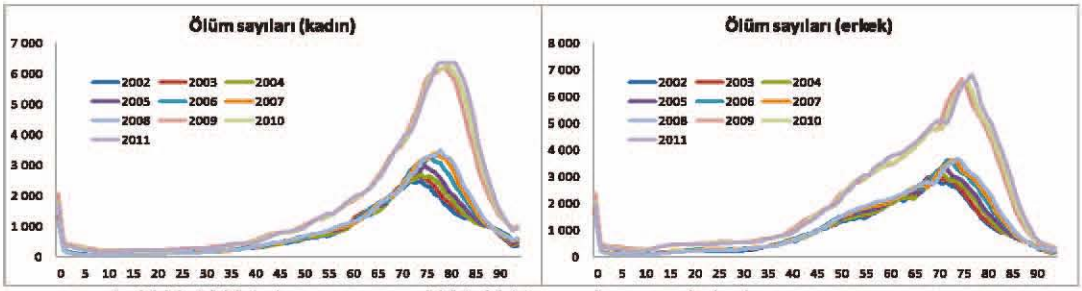
Şekil 1. 2009-2012 yıl ortası nüfus piramitleri

TÜİK tarafından ölüm verileri 2002 yılından 2009 yılına kadar il-ilçe bazında yayımlanmaktaydı. Ancak 2009 yılı itibariyle yenilenen veri kayıt sistemi ile birlikte Türkiye genelinde yayımlanmaya başlanmıştır. İl-ilçe bazında yayımlanan ölüm sayıları ile elde edilen ölüm hızları Türkiye geneli nüfusunu yansıtmadığı için gerçek değerinden daha düşük çıkmaktadır ve bu da beklenen ömrün olduğundan daha yüksek hesaplanmasına yol açmaktadır. Bu yüzden 2009 yılı ve öncesindeki ölüm verilerinin yaşam tablosu oluşturmaya elverişli olmadığı ancak 2009 yılından sonra yayımlanan verilerin Türkiye nüfusunun tümünü yansıttığı için yaşam tablosu oluşturmaya olanak sağladığı söylenebilir.



Şekil 2. 2002-2008 il-ilçe bazında, 2009-2011 Türkiye geneli düzleştirilmiş ölüm sayısı verileri (Beşli hareketli ortalamalar yöntemi)

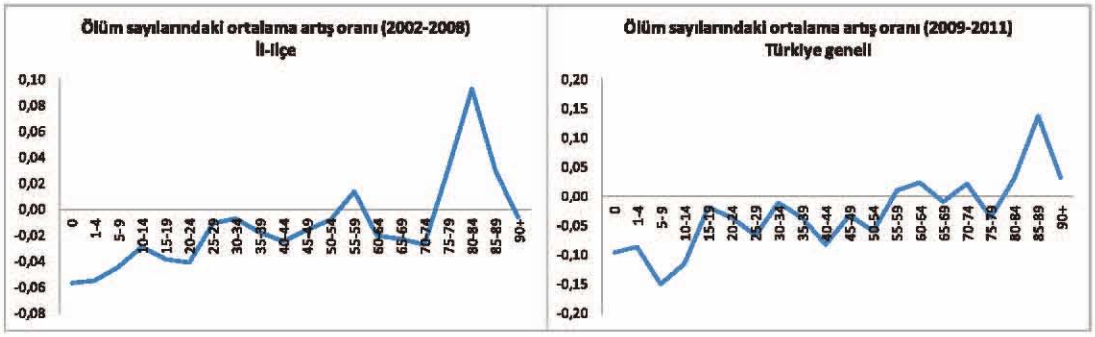
Şekil 2, 2002 ve 2011 yılları arasındaki düzleştirilmiş ölüm sayılarını göstermektedir. Buna göre 2002 yılında 175,434 olan ölüm sayısı 2008 yılında 206,296'ya ulaşmıştır. Veri kayıt sistemine köy ve bucaklarında eklenmesiyle 2009 yılında bu sayı 367,863 olarak 2010 yılında 365,532 olarak 2011 yılında ise 375,262 olarak gerçekleşmiştir. Grafikte de görüldüğü üzere ölüm sayısı yıllar itibariyle artış göstermektedir ve ölüm sayısını gösteren eğri ileri yaşlara doğru bir kayma eğilimi içerisindedir.



Şekil 3. 2002-2008 il-ilçe bazında , 2009-2011 Türkiye geneli cinsiyete göre düzleştirilmiş ölüm sayısı verileri

Şekil 3. te düzleştirilmiş ölüm sayıları yıl ve cinsiyet bazında gösterilmiştir ve cinsiyet ayrımı olmaksızın gözlenen ölüm davranışları kadın ve erkeklerde de gözlemlenmiştir. Veriler incelendiğinde yıllar itibariyle kadın ve erkek için ölüm sayıları eğrisinin ileri yaşlara doğru ilerlediği görülmektedir. Bunun anlamı ülkemizde yaşayan insanların zaman ilerledikçe daha geç yaşlarda öldükleridir. Ülkemizde yaşayan insanların ömürlerinin uzadığının bir kanıtı olarak da ifade edilebilmektedir. Ölüm sayısının en düşük olduğu yaş aralığının 10-14 olduğu gözlemlenmiştir. En fazla ölümün gerçekleştiği yaş aralığı ise 2002 yılında 70-74 yaş aralığı olurken 2005 yılından sonra 75-79 yaş aralığında gerçekleştiği gözlemlenmiştir. Şekil 3'te elde edilen gözlemler cinsiyet ayrımı olmaksızın gözlenen ölüm davranışlarının kadın ve erkekler için de geçerli olduğunu göstermiştir. Ancak erkeklerde 70 yaş ve öncesinde ölüm sayısının kadınlara oranla daha fazla olduğu görülmektedir. 70 yaş sonrasında da yine ölüm sayısının erkeklerde kadınlara göre daha fazla artış gösterdiği gözlemlenmektedir. Erkeklerde ölüm sayısının en düşük olduğu yaş aralığının 10-14 yaş aralığı olduğu gözlemlenmiştir. En fazla ölümün gerçekleştiği yaş aralığı ise 2002 yılında 70-74 yaş aralığı olurken 2007 yılından sonra 75-79 yaş aralığı olmuştur. Kadınlarda ölüm

sayısının en düşük olduğu yaş aralığının 10-14 yaş aralığı olduğu gözlemlenmiştir. En fazla ölümün gerçekleştiği yaş aralığı ise 2002 yılında 75-79 yaş aralığı olurken 2009 yılından sonra 80-84 yaş aralığı olmuştur. Görsel ifadeler değerlendirildiğinde kadınların beklenen ömrünün erkeklerden daha uzun olduğu söylenebilir ancak beklenen ömrün hesaplanması yaşam tablolarının oluşturulmasıyla mümkün kılınmaktadır.



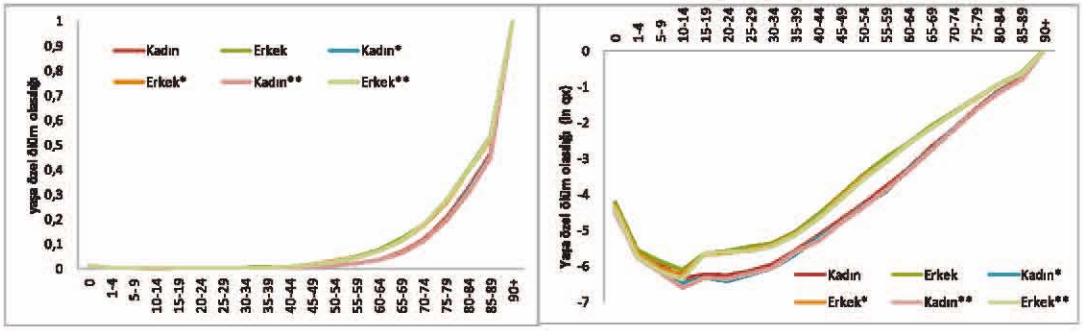
Şekil 4. İl-ilçe ve Türkiye geneli ölüm verileri için ortalama artış oranları

Şekil 4'de verilen grafikte il-ilçe ve Türkiye geneli olarak yayımlanan ölüm verilerinden elde edilen ortalama artış oranları incelenmiştir. Her iki grafiğe bakıldığında ölümlerdeki artış oranlarının 75-80 yaş aralığına kadar düşüş gösterdiği ve sonraki yaşlarda bu düşüşün artış olarak devam ettiği görülmektedir. Ölüm davranışları her iki grafik için belli yaş gruplarında farklılık gösterse de genel eğilimin ölüm sayılarındaki artışın ileriki yaş aralıklarına kaydığı ve Türkiye nüfusunun yıllar itibariyle yaşlandığıdır.

Tablo 1. 2007-2011 yılları için kaba bebek ölüm hızları (Binde)

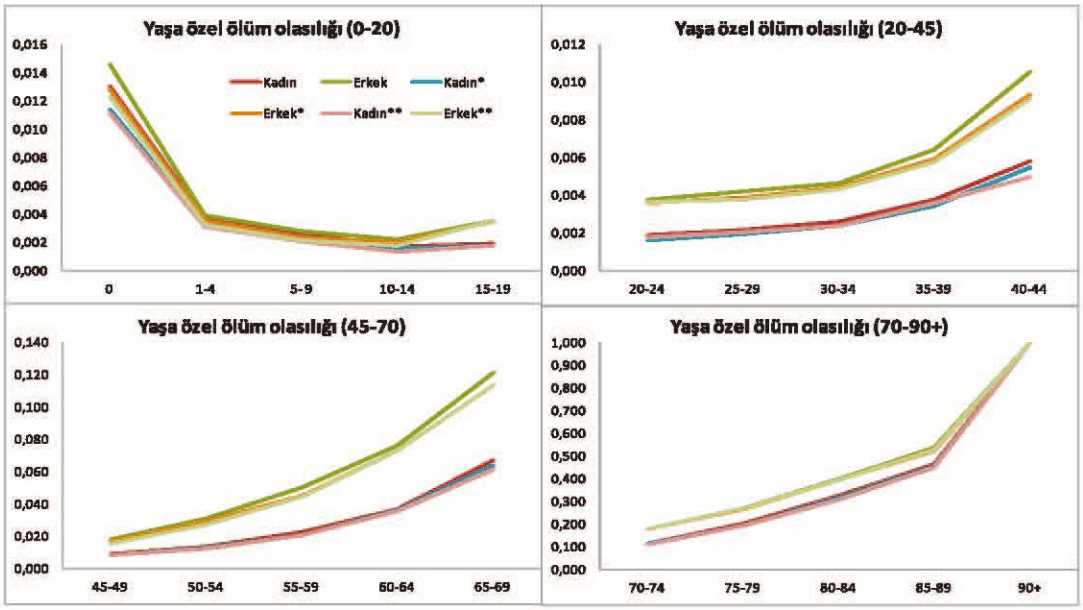
Kaba Bebek Ölüm Hızı			
Yıl	**Tüm**	**Erkek**	**Kız**
2007	**8,78**	**9,77**	**7,75**
2008	**8,56**	**9,47**	**7,60**
2009	**13,90**	**14,65**	**13,09**
2010	**12,18**	**12,88**	**11,43**
2011	**11,72**	**12,28**	**11,12**

Tablo 1'e bakıldığında yıllar itibariyle bebek ölümlerinin azaldığı görülmüştür. Ölüm olasılıkları hesaplanırken sıfır yaş için hiçbir dönüşüm yapılmadan ölüm olasılığı hesaplanmıştır yani bebek ölümleri için gerçek ölüm hızları kullanılmıştır. Ölüm olasılıkları bir önceki bölümde anlatılan dönüşümler kullanılarak gerçek gözlemlerden elde edilmiştir (Bkz., Formül (3)).



Şekil 5. 2009-2011 yılları kadın ve erkek nüfusu için yaşa özel ölüm olasılıkları
(" "=2009, "*"=2010, "**"=2011)

Şekil 5 incelendiğinde yaşa özel ölüm olasılıklarının kadınlarda erkeklere oranla daha düşük olduğu görülmektedir. Ölüm olasılıklarına ait genel grafiğe bakıldığında yıllar itibariyle düşüş görülmektedir.



Şekil 6. 2009-2011 yılları kadın ve erkek nüfusu için belirli yaş aralıklarına göre oluşturulmuş yaşa özel ölüm olasılıkları (" "=2009, "*"=2010, "**"=2011)

Şekil 6'da ölüm olasılıkları belli yaş gruplarına göre incelenmiştir. 2009 yılından 2010 yılına geçildiğinde bebek ölümlerinde genel nüfusta %14.12'lık bir azalış gözlemlenirken, kadınlarda %14.53 erkeklerde ise %13.78 olduğu görülmüştür. 2011 yılına geçildiğinde ise 2010 yılına göre genel nüfusta bebek ölümlerinde %3.88, erkeklerde %2.77 olurken kadınlarda %4.84 olarak gerçekleşmiştir. Bu azalış trendi 15 yaşına kadar devam etmektedir. 15 yaşından sonra ölüm olasılıklarında artış trendi gözlemlenirken 2010 yılında 2009 yılına göre ölüm olasılıkları düşüş göstermektedir. 40-44 yaş aralığında ölüm olasılığında meydana gelen artış erkeklerde %13.3, kadınlarda ise %6.2'dir. 2011 yılında 2010 yılına göre 40-44 yaş aralığında ölüm olasılığında meydana gelen artış erkeklerde %2.2, kadınlarda ise %10.1'dir. 65-69 yaş aralığında ölüm olasılıklarındaki azalış erkeklerde %6.6, kadınlarda ise %5.1 olduğu

ancak 2011 yılında bir önceki yıla göre bu aralıkta erkeklerde %0.2 'lik bir artış olduğu gözlemlenmiştir. 75-79 yaş aralığında ölüm olasılıklarındaki azalış erkeklerde %2.8, kadınlarda ise %4.4 olduğu ve 2011 yılında erkeklerde ölüm olasılığının %2.3 arttığı gözlemlenmiştir. 80-84 yaş aralığında ölüm olasılıklarında ki azalış erkeklerde %1.6, kadınlarda ise %2.9 olduğu ve 2011 yılında aynı trendin erkeklerde azalarak devam ettiği gözlemlenmiştir. Görüldüğü gibi bebek ve çocuk ölümlerinde ve 70 yaş sonrasında erkeklerde yıllara göre azalan ölüm olasılığı kadınlara göre daha düşüktür. Bunun tersine orta yaş aralığında daha yüksektir. Ancak 2011 yılından itibaren 70 yaşından sonra erkeklerin yaşa özel ölüm olasılıklarında bir artış trendinin başladığı görülmüştür.

Yaşam tablosu analizi kullanılarak elde edilen beklenen yaşam süreleri kadınların erkeklerden daha uzun yaşadığını göstermektedir. Tablo 2'de belirli yaşlar için elde edilen beklenen ömürler gösterilmiştir.

Tablo 2. 2009-2011 yılları cinsiyet bazında belirli yaşlar için ortalama yaşam süresi (beklenen yaşam süresi) (" "=2009, "*"=2010, "**"=2011)

Yaş	Tüm	Tüm*	Tüm**	Kadın	Kadın*	Kadın**	Erkek	Erkek*	Erkek**
0	75.8	76.5	76.8	78.5	79.1	79,4	73.1	73.9	74,1
20	57.6	58.1	58.3	60.2	60.7	60.9	55.0	55.6	55.7
40	38.3	38.8	39.0	40.7	41.1	41.4	35.9	36.4	36.5
60	20.5	20.9	21.0	22.2	22.6	22.8	18.6	19.0	19.0
80	7.4	7.6	7.7	7.9	8.1	8.3	6.6	6.8	6.8

Tablo 2'de elde edilen sonuçlara göre Türkiye'nin doğuşta beklenen yaşam süresi 2009 yılında 75.8 iken 2010 yılında 76.5, 2011 yılında ise 76.8'e yükselmiştir. Türkiye'nin doğuşta beklenen yaşam süresi %1.32'lik bir artış göstermiştir. Beklenen yaşam süresi kadınlarda 2009 yılında 78.5, 2010 yılında 79.1 iken 2011 yılında 79.4'e, erkeklerde ise 2009 yılında 73.1, 2010 yılında 73.9 iken 2011 yılında 74.1'e yükselmiştir. 2011 yılında 2009 yılına göre Türkiye nüfusunun beklenen yaşam süresi 1 yıl uzamıştır. Kadınların doğuşta beklenen yaşam süresi 10.8 ay uzarken erkeklerin ömrü 1 yıl uzamıştır.

Tablo 3. 2011 yılında 2009 yılına göre beklenen yaşam süresinde meydana gelen artış oranları

Yaş	Tüm	Kadın	Erkek
0	1.32%	1.15%	1.37%
20	1.22%	1.16%	1.27%
40	1.83%	1.72%	1.67%
60	2.44%	1.80%	2.15%
80	4.05%	2.53%	3.03%

Tablo 3'te elde edilen sonuçlara göre beklenen yaşam süresindeki artış oranı erkeklerde %1.37 olurken kadınlarda %1.15 olmuştur. Her bir yaş grubu için oranlar incelendiğinde erkeklerin beklenen yaşam süresinin 40 yaş grubu hariç kadınlardan daha hızlı uzadığı görülmüştür. Bu da erkeklerin beklenen yaşam süresinin zamanla kadınlarınkine yaklaşacağının bir göstergesidir. Tablo 3'te görülen bir başka sonuç ise beklenen yaşam süresindeki artış oranının yaşlı nüfusa doğru yükselmesidir. Görüldüğü

üzere 80 yaşında ortalama ömürdeki artış oranı %4.05'tir. Bu durum yaşlı nüfusun ömrünün yıllar itibariyle artış gösterdiğini açıklamaktadır.

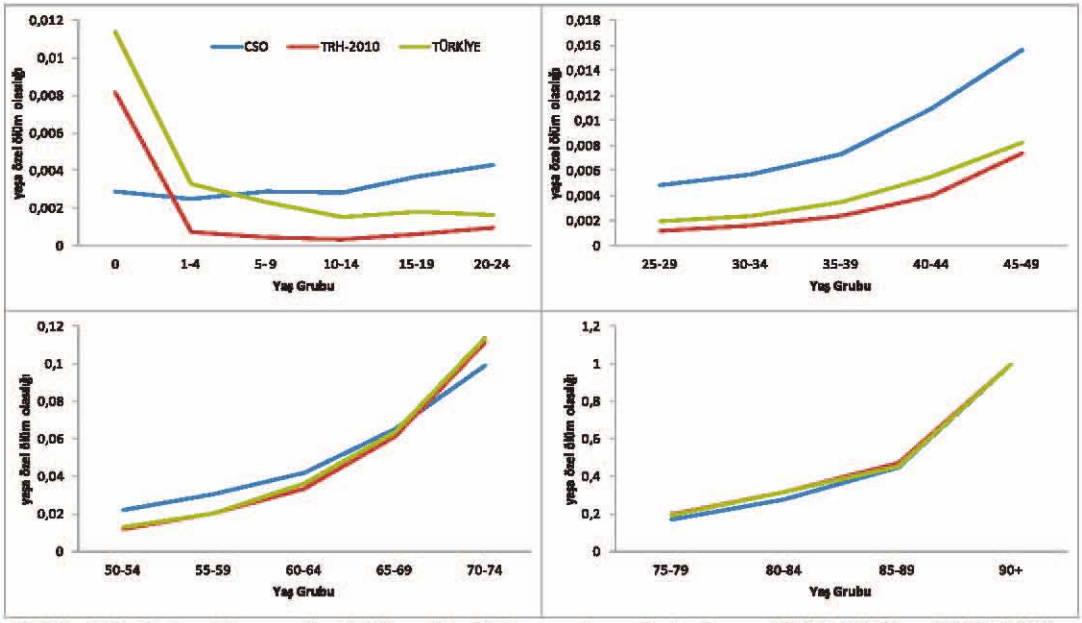
4. BULGULAR VE TARTIŞMA

Bu çalışmada gerçek nüfus ve demografik veriler kullanılarak Türkiye için Özetlenmiş Dönem Yaşam Tablosu oluşturulmuştur. Yaşam tablosu analizi ve çeşitli demografik yaklaşımlar kullanılarak ölüm davranışları incelenmiş ve yaş grubu ve cinsiyet bazında beklenen yaşam süreleri hesaplanmıştır. Buna göre yaşa özel ölüm olasılığının bebeklerde ve genç nüfusta düşüş gösterdiği ve bu olasılığın yaşlı nüfusa doğru artış gösterdiği görülmüştür. Bunun bir sonucu olarak Türkiye'nin yıllar itibariyle daha uzun yaşadığı görülmüştür. Kadınların erkeklerden daha uzun yaşadığı ancak beklenen yaşam süresindeki artışın erkeklerde daha fazla olduğu önemli bir bulgu olarak karşımıza çıkmıştır. Bunun anlamı ise erkeklerin ilerleyen yıllarda kadınların ortalama yaşam sürelerine yaklaşacaklarıdır.

Ölüm davranışları il-ilçe ve Türkiye geneli bazında yayımlanan veri ile karşılaştırılmış ve genel davranışın aynı olduğu ancak 2009 yılı itibariyle veri kayıt sistemine köy ve bucaklarının dahil edilmesiyle birlikte belli yaş grupları için farklılık gösterdiği görülmüştür. Buda Türkiye'nin gerçek demografik yapısını yansıtmadığı için 2009 yılı öncesinde yayımlanan ölüm verilerinin kullanılmasının yanlış sonuçlar doğurabileceğini göstermiştir. Ayrıca 2009 yılı itibariyle ölüm verilerinin güncellenmesiyle birlikte mevcut nüfusun demografik yapısını yansıtan Dönem Yaşam Tablolarının oluşturulması mümkün kılınmıştır.

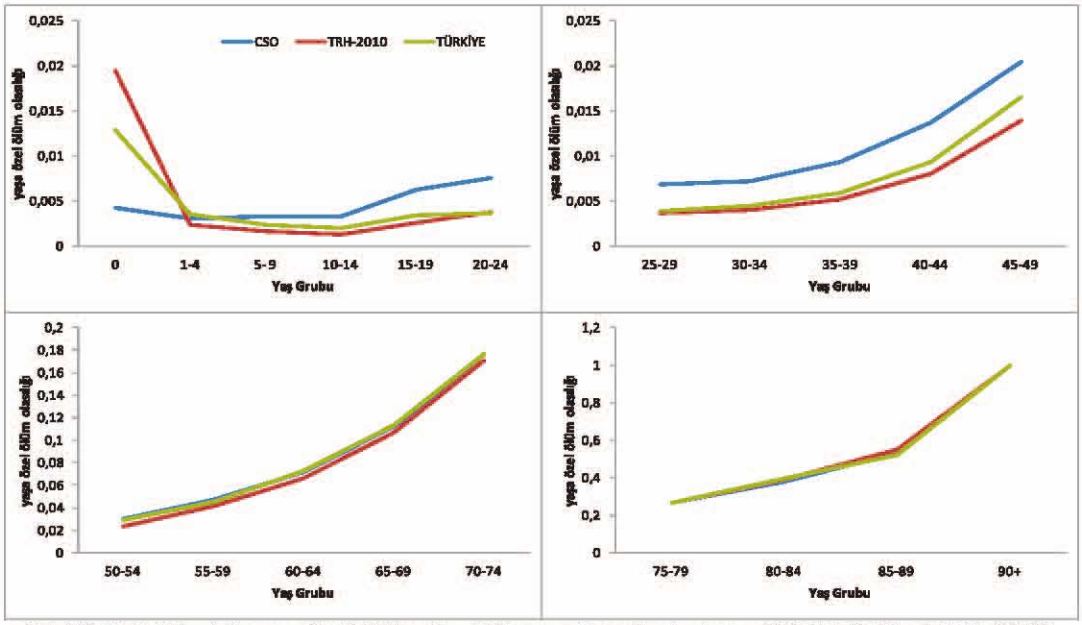
Şekil 7 ve Şekil 8'de bu çalışmada elde edilen yaşa özel ölüm olasılıkları üç ayrı yaşam tablosu ile cinsiyet bazında karşılaştırılmıştır. Bu tablolar sırasıyla ülkemizde halen kullanılan CSO1980 Yaşam Tablosu (The Commissioners 1980 Standard Ordinary Mortality Table), bu çalışmada oluşturulan yaşam tablosu (Turkiye) ve Hayat Sigortaları Bilgi Merkezi (HAYMER) için yapılan projede Türkiye için oluşturulan yaşam tablosudur (TRH-2010).

Şekil 7 incelendiğinde elde edilen yaşa özel ölüm olasılıklarının farklı yaş aralıkları için davranışları yorumlanmıştır. Buna göre 0-25 yaş aralığı için ülkemiz için oluşturulan iki tabloda CSO'ya göre aynı davranışları göstermiştir. Ancak bu çalışmada elde edilen tabloya göre bebek ölümleri CSO ve TRH-2010 tablosuna göre daha yüksek, 1-25 yaş aralığında ise ölüm olasılıkları TRH-2010 tablosuna göre CSO'ya daha yakın çıkmıştır. Benzer şekilde aynı durum 65-69 yaş grubuna kadar devam etmiştir. 70 yaş sonrasında ise CSO'ya göre ölüm olasılıkları yükselmiştir bu da ülke nüfusunun yaşlandığını göstermektedir. Ancak bu aralıkta ölüm olasılıkları TRH-2010 tablosunda bu çalışmaya göre CSO'ya daha yakın çıkmıştır. Yani bu çalışmada beklenen ömürdeki artışın kaynağı yaşlanan nüfus olurken TRH-2010 tablosunda artışın nedeni daha çok genç nüfustaki ölüm olasılıklarının düşüklüğü olmuştur.



Şekil 7. Kadınlar için yaşa özgü ölüm olasılıklarının karşılaştırılması (CSO-Türkiye-TRH-2010)

Şekil 8 incelendiğinde elde edilen yaşa özel ölüm olasılıklarının bebek ölümleri hariç tüm yaş gruplarında CSO'ya göre daha düşük olduğu görülmüştür. Ancak Türkiye ve TRH-2010 karşılaştırıldığında özellikle yaşlı nüfustaki ölüm olasılıklarındaki artışın bu çalışmada elde edilen tabloda TRH-2010'a göre daha yüksek olduğu görülmüştür



Şekil 8. Erkekler için yaşa özgü ölüm olasılıklarının karşılaştırılması (CSO-Türkiye-TRH-2010)

Yukarıda grafiklerde görüldüğü üzere Türkiye'nin beklenen yaşam süresi CSO1980 tablosuna göre hesaplanan süreden daha yüksektir. HAYMER için oluşturulan tabloya ait yaşa özel ölüm olasılıkları incelendiğinde beklenen ömürdeki artışın nedeninin bebek ve çocuk ölümlerindeki düşüş olduğu ancak gerçek veriler bu artışın nedenin

bebek ve çocuk ölümlerindeki düşüşten daha çok yaşlanan bir Türkiye nüfusunun olduğunu göstermiştir.

Tablo 4. Belirli yaşlar için cinsiyet bazında beklenen yaşam süresi

	Kadın			Erkek		
Yaş	CSO	Türkiye	Haymer	CSO	Türkiye	Haymer
0	75.8	79.1	78.0	70.8	73.9	71.9
20	57.0	60.7	58.9	52.4	55.6	54.0
40	38.4	41.1	39.2	34.1	36.4	34.9
60	21.2	22.6	20.8	17.5	19.0	17.6
80	7.5	8.1	7.0	6.2	6.8	6.0

Tablo 4'te kadın ve erkekler için elde edilen beklenen yaşam süreleri karşılaştırılmıştır. Buna göre kadınlarda beklenen yaşam süresi TRH-2010'a göre 1.1 yıl, erkeklerde ise 2 yıl daha uzundur. Yaşlı nüfus için 60 yaş incelendiğinde bu çalışmada hesaplanan beklenen yaşam süresi TRH-201'a göre kadınlarda 1.8 yıl, erkeklerde 1.4 yıl daha uzundur.

Bu çalışmadan elde edilen bulgular göstermiştir ki ülkemizde halen aktif olarak kullanılan başka ülkelerin tabloları ülkemizin gerçek demografik yapısını yansıtmamaktadır. Bu durum başta sosyal güvenlik sistemi ve sigorta sektörü için tehlike arz etmektedir. Bu tehlikelerin önlenmesi için Türkiye kendi demografik yapısını yansıtan yaşam tablosunu oluşturmalı ve aktif olarak kullanmalıdır.

TÜİK tarafından yayımlanan ölüm verileri bilinmeyen yaşa ait verileri de içermektedir ancak bu çalışmada bilinmeyen yaşa ait olan ölüm sayıları veriye dahil edilmemiştir. Bu çalışma için bir eksikliktir ve çeşitli istatistiksel yöntemler kullanılarak bu eksiklik giderilebilir.

Ayrıca bu çalışmada Özetlenmiş Yaşam Tablosu 2009, 2010 ve 2011 yılları için ayrı ayrı oluşturulmuştur. Ancak dönem yaşam tablosu her yıl ayrı ayrı oluşturulabildiği gibi 3 yıllık ya da benzeri dönemler içinde oluşturulabilir. Bu çalışmada örnek olması için tek yıl ve yaş grupları için oluşturulmuştur. İlerleyen çalışmalarda tek yaşlar için oluşturulması hedeflenmektedir.

Son olarak gelecek yıllar itibariyle eklenecek yeni ölüm verileri her yıl ülkemiz için dönem yaşam tablosunun yenilenmesine ve uygun yıl sayısına ulaşıldığında ölüm olasılıklarının ve beklenen yaşam sürelerinin gelecek için kestrimine olanak sağlayacaktır.

5. KAYNAKLAR

Alpay A., 1969. Abridged Life Tables for Selected Regions and Cities of Turkey, Turkish Demography: Proceedings of a Conference. H.Ü. Nüfus Etütleri Enstitüsü, 83-108.

Arias E., 2010. United States Life Tables by Hispanic Origin, National Center for Health Statistics. Vital Health Stat 2(152).

Australian Bureau of Statistics (ABS). 2008. Life Tables. Retrieved 2012, from http://www.abs.gov.au/AUSSTATS/abs@.nsf/webpages/statistics?opendocument.

Bell, F., C., Miller, M., L., 2005. Life Tables for the United States Social Security Area 1900-2100, Social Security Administration. *Actuarial Study*, 120.

Chiang, C. L., 1968. Introduction to Stochastic Processes in Biostatistics, New York: John-wiley and Sons.

Chiang, C. L., 1972. On Constructing Current Life Tables, Journal of the American Statistical Association 67, 538-541.

Coşkun, Y., 2002. Estimation of Adult Mortality by Using the Orphanhood Method from the 1993 and 1998 Turkish Demographic and Health Surveys, H. Ü. Nüfus Etütleri, Ankara. (İngilizce).

Demirbüken, D., 2001. An Evaluation of Burial Records of Ankara City Cemeteries, H. Ü. Nüfus Etütleri Enstitüsü, Ankara. (İngilizce).

Demirci, M., 1987. Türkiye'nin Ölümlülük Yaş Yapısına Model Yaşam Tablolarından En Uygun Kalıbın Seçimi, H. Ü. Fen Bilimleri Enstütüsü, Ankara. (Türkçe).

Duransoy, M. L., 1993. Türk Mortalite Tablosu (1980-2000), Mimar Sinan Üniversitesi, İstanbul. (Türkçe).

Eryurt, M. A., Koç, İ., 2010. Türkiye için Hayat Tablolarının Sentetik Yetimlik Tekniği ile Oluşturulması, Nüfusbilim Dergisi, 28-29, 47-60.

Fergany, N., 1971. On the Human Survivorship Function and Life Table Construction, Demography 8, 331-334.

Government Actuary's Department (SNZ)., 2010. Life Tables, Retrieved 2012, from http://www.gad.gov.uk/Demography%20Data/Life%20Tables/.

Greville, T. N. E., 1943. Short Methods of Constructing Abridged Life Tables, Record of the American Institute of Actuaries 32, 28-34.

Hoşgör, Ş., 1992. Estimation of Post-Childhood Life Tables Using Age and Sex Distributions and Intercensal Growth Rates in Turkey, (1930-1990), H. Ü. Nüfus Etütleri Enstitüsü, Ankara, (İngilizce).

Hoşgör, Ş., 1997. Estimation of Post-Childhood Life Tables of Provinces and Regions in Turkey, by Using Age and Sex Distributions and Intercensal Growth Rates (1985-1990), H. Ü. Nüfus Etütleri Enstitüsü, Ankara, (İngilizce).

Keyfitz N., 1982. Choice of Function for Mortality Analysis: Effective Forecasting Depends on a Minimum Parameter Representation Theoretical Population Biology, 21, 329–352.

Keyfitz, N., Caswell, H., 2005. Applied Mathematical Demography, 3rd Edition. New York: Springer.

Keyfitz, N., Frauenthal, J., 1975. An Improved Life Table Method, Biometrics 31, 889-900.

Kırkbeşoğlu, E., 2006. Construction of Mortality Tables for Life Insurance Sector from the 2003 Turkey Demographic and Health Survey, H. Ü. Nüfus Etütleri Enstitüsü, Ankara. (İngilizce).

Kintner, H. J., 2004. The Life Table. In: Siegel, J.S. and Swanson, D. A. (eds.). The Methods and Materials of Demography, 2nd edition. San Diego: Elsevier, Academic Press: 301-340.

National Center For Health Statistics (NCHS). 2011. Life Table Analysis System. Retrieved 2011, from http://www.cdc.gov/niosh/LTAS/.

Öcal, M., 1974. Türkiye Ölüm Oranları Tablosu (1960/1961), İstanbul.

Özsoy, A., 1970. Türkiye için Ölüm Tabloları, Ankara, Ordu Yardımlaşma Kurumu Yayınları.

Preston, S. H., Heuveline, P., Guillot, M., 2001. Demography Measuring and Modeling Population Process, Blackwell, INC., Publications, USA.

Pfaff, Thomas, J., Seltzer, Stanley, 2012. "Period Life Tables: A Resource for Quantitative Literacy, 5(1), Article 5.

Reed, L. J., Merrell, M., 1939. A Short Method for Constructing an Abridged Life Table, American Journal of Hygiene 30.

Selvin, S., 2008. Survival Analysis for Epidemiologic and Medical Research, Cambridge University, INC., Publications, USA.

Shryock, H. S., Siegel, J. S., Associates, 1971. The Methods and Materials of Demography, U.S. Government Printing Office, Washington, D.C.

Sigorta Bilgi ve Gözetim Merkezi, 2010. Türkiye Hayat ve Hayat Annüite Tablolarının Oluşturulması Projesi, 2012. http://www.sbm.org.tr/?p=mortaliteIstatistik, 2012.

Statistics New Zealand (SNZ). (2009. *Period Life Tables*.Retrieved, 2012, from http://www.stats.govt.nz/browse_for_stats/health/life_expectancy/period-lifetables.aspx.

Toros, A., 2000. Life Tables for the Last Decade of XX. Century in Turkey, Nüfusbilim Dergisi, 22, 57-110.

Tucek, D., G., A., 2011. Comparison of Period and Cohort Life Tables, Journal of Legal Economics, 17(2), 99-112.

Tuzgöl H., 2005. SSK Ölüm İstatistiklerinin İncelenmesi ve Farklı Gruplar için Yaşam Tablosunun Oluşturulması, Sosyal Sigortalar Kurumu Başkanlığı, Ankara. (Türkçe)

TÜİK, Population and Demographic Statistics, 2012. Deaths.Retrieved 2012, from http://www.tuik.gov.tr/AltKategori.do?ust_id=11.

TÜİK, Population and Demographic Statistics, 2012. Births.Retrieved 2012, from http://www.tuik.gov.tr/AltKategori.do?ust_id=11.

TÜİK, Population and Demographic Statistics, 2012. Population.Retrieved 2012, from http://www.tuik.gov.tr/AltKategori.do?ust_id=11.

CONSTRUCTION OF A LIFE TABLE BY USING THE TURKISH GENERAL DEATH DATA

ABSTRACT

The purpose of this study is the construction of a period life table by age group and gender by using the Turkish general death data which has been started to be published by TurkStat since 2009 with the renewed death data registration system. Turkish general death data and the population data which is obtained from Address Based Population Registration System are used while constructing this life table. Firstly, age specific death rates has been calculated for each age group in a certain time period. Next, age specific death, life probabilities and the expected lifetime have been obtained for the population of Turkey. As a result, the expectation of life at birth is observed to be 79.4 for females and 74.1 for males in 2011. As a result of these studies, it is observed that the population of Turkey is getting older, the life expectancy for females is on the average 5.3 years more than the one for males, and this difference is expeditiously decreasing as time passes. Finally, the results are evaluated and discussed.

Keywords: Life expectancy, Abridged period life table, Age specific death probability, Life/mortality table.

Ek 1. Türkiye Özetlenmiş Dönem Yaşam Tablosu, 2009

Age	$_nm_x$	$_na_x$	$_nq_x$	$_np_x$	l_x	$_nd_x$	$_nL_x$	T_x	e_x^0
0	0,0139		0,0139	0,9861	1.000.000	13.895	986.105	75.759.422	75,8
1	0,0010	1,1726	0,0039	0,9961	986.105	3.810	3.933.647	74.773.317	75,8
5	0,0006	1,8642	0,0028	0,9972	982.295	2.723	4.902.937	70.839.671	72,1
10	0,0004	2,0484	0,0020	0,9980	979.572	1.934	4.892.154	65.936.734	67,3
15	0,0006	2,1131	0,0028	0,9972	977.639	2.702	4.880.393	61.044.580	62,4
20	0,0006	2,0526	0,0029	0,9971	974.937	2.791	4.866.457	56.164.187	57,6
25	0,0006	2,0506	0,0032	0,9968	972.146	3.118	4.851.531	51.297.730	52,8
30	0,0007	2,0019	0,0036	0,9964	969.027	3.527	4.834.560	46.446.199	47,9
35	0,0010	2,0381	0,0051	0,9949	965.500	4.942	4.812.860	41.611.638	43,1
40	0,0017	2,3448	0,0082	0,9918	960.558	7.914	4.781.775	36.798.778	38,3
45	0,0027	2,1651	0,0135	0,9865	952.644	12.897	4.726.657	32.017.003	33,6
50	0,0045	2,2712	0,0224	0,9776	939.747	21.072	4.641.236	27.290.346	29,0
55	0,0074	2,1823	0,0364	0,9636	918.675	33.402	4.499.262	22.649.110	24,7
60	0,0115	2,0985	0,0558	0,9442	885.274	49.400	4.283.033	18.149.848	20,5
65	0,0195	2,1604	0,0924	0,9076	835.874	77.265	3.959.963	13.866.815	16,6
70	0,0320	1,9882	0,1458	0,8542	758.609	110.616	3.459.890	9.906.852	13,1
75	0,0548	1,9660	0,2349	0,7651	647.993	152.198	2.778.192	6.446.961	9,9
80	0,0910	1,9126	0,3552	0,6448	495.794	176.098	1.935.287	3.668.769	7,4
85	0,1465	1,6235	0,4900	0,5100	319.697	156.648	1.069.559	1.733.482	5,4
90+	0,2456	3,3810	1,0000	0,0000	163.048	163.048	663.923	663.923	4,1

Ek 2. Türkiye Özetlenmiş Dönem Yaşam Tablosu, Erkek 2009

Age	$_nm_x$	$_na_x$	$_nq_x$	$_np_x$	l_x	$_nd_x$	$_nL_x$	T_x	e_x^0
0	0,0146		0,0146	0,9854	1.000.000	14.609	985.391	73.125.091	73,1
1	0,0010	1,1806	0,0039	0,9961	985.391	3.883	3.930.615	72.139.700	73,2
5	0,0006	1,8955	0,0028	0,9972	981.508	2.758	4.898.975	68.209.085	69,5
10	0,0004	2,0924	0,0022	0,9978	978.749	2.143	4.887.516	63.310.109	64,7
15	0,0007	2,1871	0,0035	0,9965	976.606	3.448	4.873.333	58.422.593	59,8
20	0,0008	2,0652	0,0038	0,9962	973.159	3.682	4.854.986	53.549.260	55,0
25	0,0008	2,0286	0,0042	0,9958	969.476	4.091	4.835.227	48.694.274	50,2
30	0,0009	1,9779	0,0046	0,9954	965.386	4.473	4.813.410	43.859.047	45,4
35	0,0013	2,0333	0,0064	0,9936	960.912	6.177	4.786.236	39.045.638	40,6
40	0,0021	2,3291	0,0106	0,9894	954.735	10.090	4.746.726	34.259.402	35,9
45	0,0036	2,1797	0,0179	0,9821	944.645	16.909	4.675.538	29.512.675	31,2
50	0,0063	2,2578	0,0309	0,9691	927.737	28.622	4.560.195	24.837.137	26,8
55	0,0104	2,1793	0,0503	0,9497	899.114	45.218	4.368.026	20.276.942	22,6
60	0,0159	2,0830	0,0760	0,9240	853.896	64.869	4.080.258	15.908.916	18,6
65	0,0260	2,1248	0,1210	0,8790	789.028	95.445	3.670.719	11.828.659	15,0
70	0,0406	1,9519	0,1808	0,8192	693.582	125.429	3.085.591	8.157.940	11,8
75	0,0659	1,8503	0,2727	0,7273	568.154	154.948	2.352.736	5.072.349	8,9
80	0,1074	1,8567	0,4014	0,5986	413.206	165.847	1.544.722	2.719.613	6,6
85	0,1698	1,5561	0,5358	0,4642	247.359	132.525	780.392	1.174.891	4,7
90+	0,2911	3,0861	1,0000	0,0000	114.835	114.835	394.499	394.499	3,4

Ek 3. Türkiye Özetlenmiş Dönem Yaşam Tablosu, Kadın, 2009

Age	$_nm_x$	$_na_x$	$_nq_x$	$_np_x$	l_x	$_nd_x$	$_nL_x$	T_x	e_x^0
0	0,0131		0,0131	0,9869	1.000.000	13.057	986.943	78.455.301	78,5
1	0,0009	1,1636	0,0037	0,9963	986.943	3.655	3.937.406	77.468.358	78,5
5	0,0005	1,8298	0,0027	0,9973	983.288	2.650	4.908.040	73.530.952	74,8
10	0,0003	1,9900	0,0017	0,9983	980.638	1.703	4.898.064	68.622.912	70,0
15	0,0004	1,9721	0,0019	0,9981	978.935	1.899	4.888.925	63.724.848	65,1
20	0,0004	2,0265	0,0019	0,9981	977.036	1.854	4.879.669	58.835.923	60,2
25	0,0004	2,0949	0,0022	0,9978	975.183	2.098	4.869.818	53.956.254	55,3
30	0,0005	2,0455	0,0026	0,9974	973.085	2.544	4.857.908	49.086.436	50,4
35	0,0008	2,0464	0,0038	0,9962	970.541	3.676	4.841.847	44.228.528	45,6
40	0,0012	2,3744	0,0058	0,9942	966.865	5.624	4.819.557	39.386.681	40,7
45	0,0018	2,1360	0,0091	0,9909	961.240	8.756	4.781.125	34.567.124	36,0
50	0,0028	2,3018	0,0138	0,9862	952.484	13.147	4.726.949	29.785.999	31,3
55	0,0045	2,1890	0,0224	0,9776	939.337	21.074	4.637.448	25.059.051	26,7
60	0,0076	2,1278	0,0371	0,9629	918.264	34.072	4.493.456	20.421.603	22,2
65	0,0139	2,2177	0,0668	0,9332	884.192	59.107	4.256.504	15.928.146	18,0
70	0,0249	2,0361	0,1160	0,8840	825.085	95.682	3.841.836	11.671.643	14,1
75	0,0457	2,1005	0,2020	0,7980	729.403	147.308	3.219.895	7.829.807	10,7
80	0,0811	1,9568	0,3251	0,6749	582.095	189.246	2.334.554	4.609.912	7,9
85	0,1343	1,6666	0,4638	0,5362	392.848	182.213	1.356.847	2.275.358	5,8
90+	0,2293	3,5081	1,0000	0,0000	210.635	210.635	918.511	918.511	4,4

Ek 4. Türkiye Özetlenmiş Dönem Yaşam Tablosu, 2010

Age	$_nm_x$	$_na_x$	$_nq_x$	$_np_x$	l_x	$_nd_x$	$_nL_x$	T_x	e_x^0
0	0,0122		0,0122	0,9878	1.000.000	12.175	987.825	76.492.416	76,5
1	0,0009	1,1365	0,0034	0,9966	987.825	3.381	3.941.617	75.504.592	76,4
5	0,0005	1,7157	0,0024	0,9976	984.444	2.315	4.914.615	71.562.974	72,7
10	0,0004	2,0373	0,0018	0,9982	982.128	1.732	4.905.512	66.648.360	67,9
15	0,0005	2,0947	0,0026	0,9974	980.397	2.595	4.894.444	61.742.848	63,0
20	0,0005	2,0317	0,0026	0,9974	977.802	2.576	4.881.362	56.848.403	58,1
25	0,0006	2,1163	0,0029	0,9971	975.226	2.871	4.867.849	51.967.041	53,3
30	0,0007	1,9497	0,0034	0,9966	972.355	3.346	4.851.566	47.099.192	48,4
35	0,0009	2,1159	0,0047	0,9953	969.008	4.543	4.831.939	42.247.626	43,6
40	0,0015	2,1704	0,0075	0,9925	964.465	7.186	4.801.991	37.415.687	38,8
45	0,0025	2,0603	0,0124	0,9876	957.279	11.879	4.751.472	32.613.696	34,1
50	0,0043	2,1919	0,0213	0,9787	945.399	20.169	4.670.359	27.862.224	29,5
55	0,0067	1,9699	0,0328	0,9672	925.230	30.374	4.534.114	23.191.865	25,1
60	0,0111	1,9683	0,0539	0,9461	894.856	48.225	4.328.074	18.657.752	20,9
65	0,0184	2,0616	0,0874	0,9126	846.631	73.983	4.015.761	14.329.678	16,9
70	0,0312	2,0057	0,1425	0,8575	772.648	110.111	3.533.532	10.313.917	13,3
75	0,0522	2,1130	0,2267	0,7733	662.537	150.216	2.879.011	6.780.385	10,2
80	0,0881	1,9427	0,3471	0,6529	512.321	177.850	2.017.868	3.901.374	7,6
85	0,1416	1,6205	0,4788	0,5212	334.472	160.149	1.131.134	1.883.507	5,6
90+	0,2317	3,2246	1,0000	0,0000	174.323	174.323	752.373	752.373	4,3

Ek 5. Türkiye Özetlenmiş Dönem Yaşam Tablosu Erkek, 2010

Age	$_nm_x$	$_na_x$	$_nq_x$	$_np_x$	l_x	$_nd_x$	$_nL_x$	T_x	e_x^0
0	0,0129		0,0129	0,9871	1.000.000	12.853	987.147	73.903.554	73,9
1	0,0009	1,1509	0,0035	0,9965	987.147	3.479	3.938.675	72.916.407	73,9
5	0,0005	1,7769	0,0023	0,9977	983.668	2.311	4.910.890	68.977.732	70,1
10	0,0004	2,0467	0,0020	0,9980	981.357	1.935	4.901.068	64.066.843	65,3
15	0,0007	2,1508	0,0034	0,9966	979.422	3.367	4.887.515	59.165.774	60,4
20	0,0007	2,0451	0,0036	0,9964	976.055	3.500	4.869.931	54.278.259	55,6
25	0,0008	2,0785	0,0039	0,9961	972.555	3.789	4.851.704	49.408.328	50,8
30	0,0009	1,9388	0,0045	0,9955	968.766	4.312	4.830.628	44.556.624	46,0
35	0,0012	2,1314	0,0059	0,9941	964.454	5.703	4.805.909	39.725.996	41,2
40	0,0019	2,1784	0,0093	0,9907	958.751	8.946	4.768.513	34.920.087	36,4
45	0,0033	2,0870	0,0165	0,9835	949.805	15.697	4.703.302	30.151.574	31,7
50	0,0060	2,2004	0,0293	0,9707	934.109	27.385	4.593.876	25.448.272	27,2
55	0,0093	1,9758	0,0453	0,9547	906.724	41.046	4.409.488	20.854.396	23,0
60	0,0153	1,9468	0,0730	0,9270	865.678	63.224	4.135.353	16.444.908	19,0
65	0,0243	2,0427	0,1135	0,8865	802.454	91.058	3.742.986	12.309.555	15,3
70	0,0396	1,9676	0,1767	0,8233	711.396	125.703	3.175.797	8.566.569	12,0
75	0,0628	2,0501	0,2650	0,7350	585.693	155.233	2.470.543	5.390.772	9,2
80	0,1048	1,8875	0,3951	0,6049	430.460	170.069	1.622.967	2.920.229	6,8
85	0,1626	1,5819	0,5226	0,4774	260.391	136.078	836.825	1.297.263	5,0
90+	0,2700	2,9682	1,0000	0,0000	124.313	124.313	460.438	460.438	3,7

Ek 6. Türkiye Özetlenmiş Dönem Yaşam Tablosu, Kadın, 2010

Age	$_nm_x$	$_na_x$	$_nq_x$	$_np_x$	l_x	$_nd_x$	$_nL_x$	T_x	e_x^0
0	0,0114		0,0114	0,9886	1.000.000	11.399	988.601	79.121.556	79,1
1	0,0008	1,1201	0,0033	0,9967	988.601	3.217	3.945.141	78.132.955	79,0
5	0,0005	1,6506	0,0023	0,9977	985.385	2.290	4.919.251	74.187.814	75,3
10	0,0003	2,0245	0,0015	0,9985	983.094	1.502	4.911.002	69.268.563	70,5
15	0,0004	1,9819	0,0018	0,9982	981.592	1.761	4.902.646	64.357.561	65,6
20	0,0003	2,0010	0,0016	0,9984	979.831	1.602	4.894.351	59.454.916	60,7
25	0,0004	2,1944	0,0019	0,9981	978.229	1.904	4.885.805	54.560.565	55,8
30	0,0005	1,9708	0,0024	0,9976	976.326	2.337	4.874.550	49.674.759	50,9
35	0,0007	2,0889	0,0034	0,9966	973.989	3.353	4.860.183	44.800.210	46,0
40	0,0011	2,1562	0,0055	0,9945	970.636	5.316	4.838.061	39.940.026	41,1
45	0,0017	2,0061	0,0082	0,9918	965.320	7.945	4.802.811	35.101.965	36,4
50	0,0026	2,1722	0,0131	0,9869	957.375	12.554	4.751.372	30.299.154	31,6
55	0,0041	1,9569	0,0204	0,9796	944.820	19.272	4.665.452	25.547.782	27,0
60	0,0073	2,0090	0,0360	0,9640	925.548	33.275	4.528.213	20.882.329	22,6
65	0,0132	2,0920	0,0636	0,9364	892.273	56.743	4.296.355	16.354.117	18,3
70	0,0244	2,0550	0,1138	0,8862	835.529	95.052	3.897.724	12.057.761	14,4
75	0,0434	2,1875	0,1932	0,8068	740.478	143.074	3.299.996	8.160.037	11,0
80	0,0781	1,9866	0,3160	0,6840	597.404	188.807	2.418.071	4.860.041	8,1
85	0,1305	1,6450	0,4539	0,5461	408.597	185.449	1.420.801	2.441.970	6,0
90+	0,2185	3,3281	1,0000	0,0000	223.148	223.148	1.021.170	1.021.170	4,6

Ek 7. Türkiye Özetlenmiş Dönem Yaşam Tablosu, 2011

Age	${}_nm_x$	${}_na_x$	${}_nq_x$	${}_np_x$	l_x	${}_nd_x$	${}_nL_x$	T_x	e_x^0
0	0,0117		0,0117	0,9883	1.000.000	11.720	988.280	76.755.744	76,8
1	0,0008	1,1269	0,0032	0,9968	988.280	3.152	3.944.064	75.767.464	76,7
5	0,0004	1,6902	0,0021	0,9979	985.128	2.064	4.918.809	71.823.400	72,9
10	0,0003	2,0768	0,0016	0,9984	983.064	1.525	4.910.863	66.904.591	68,1
15	0,0005	2,1151	0,0027	0,9973	981.539	2.629	4.900.111	61.993.728	63,2
20	0,0005	2,0568	0,0027	0,9973	978.910	2.657	4.886.730	57.093.618	58,3
25	0,0006	2,0218	0,0029	0,9971	976.253	2.839	4.872.811	52.206.888	53,5
30	0,0007	1,9993	0,0033	0,9967	973.414	3.259	4.857.291	47.334.077	48,6
35	0,0009	2,2168	0,0047	0,9953	970.155	4.555	4.838.097	42.476.786	43,8
40	0,0014	2,0620	0,0071	0,9929	965.600	6.852	4.807.868	37.638.689	39,0
45	0,0024	2,0649	0,0119	0,9881	958.748	11.419	4.760.223	32.830.821	34,2
50	0,0040	2,0285	0,0198	0,9802	947.329	18.798	4.680.783	28.070.598	29,6
55	0,0067	1,9424	0,0327	0,9673	928.530	30.386	4.549.744	23.389.815	25,2
60	0,0111	2,0380	0,0536	0,9464	898.144	48.182	4.348.009	18.840.071	21,0
65	0,0182	2,0949	0,0864	0,9136	849.963	73.415	4.036.535	14.492.063	17,1
70	0,0308	2,0790	0,1413	0,8587	776.547	109.716	3.562.258	10.455.527	13,5
75	0,0524	2,2384	0,2288	0,7712	666.831	152.593	2.912.749	6.893.269	10,3
80	0,0861	1,9454	0,3408	0,6592	514.238	175.254	2.035.866	3.980.520	7,7
85	0,1412	1,5985	0,4769	0,5231	338.984	161.664	1.145.025	1.944.654	5,7
90+	0,2218	3,0890	1,0000	0,0000	177.321	177.321	799.629	799.629	4,5

Ek 8. Türkiye Özetlenmiş Dönem Yaşam Tablosu, Erkek, 2011

Age	$_nm_x$	$_na_x$	$_nq_x$	$_np_x$	l_x	$_nd_x$	$_nL_x$	T_x	e_x^0
0	0,0123		0,0123	0,9877	1.000.000	12.284	987.716	74.091.496	74,1
1	0,0008	1,1333	0,0033	0,9967	987.716	3.232	3.941.598	73.103.780	74,0
5	0,0004	1,7252	0,0021	0,9979	984.484	2.099	4.915.544	69.162.182	70,3
10	0,0003	2,1797	0,0017	0,9983	982.385	1.714	4.907.090	64.246.638	65,4
15	0,0007	2,1384	0,0035	0,9965	980.671	3.461	4.893.451	59.339.548	60,5
20	0,0007	2,0744	0,0036	0,9964	977.210	3.555	4.875.650	54.446.097	55,7
25	0,0008	1,9811	0,0038	0,9962	973.655	3.653	4.857.246	49.570.447	50,9
30	0,0009	1,9931	0,0043	0,9957	970.002	4.163	4.837.491	44.713.201	46,1
35	0,0012	2,2068	0,0058	0,9942	965.839	5.558	4.813.667	39.875.710	41,3
40	0,0018	2,0954	0,0091	0,9909	960.280	8.766	4.775.940	35.062.043	36,5
45	0,0031	2,0968	0,0156	0,9844	951.515	14.845	4.714.473	30.286.103	31,8
50	0,0055	2,0392	0,0270	0,9730	936.669	25.303	4.608.429	25.571.630	27,3
55	0,0092	1,9501	0,0449	0,9551	911.366	40.887	4.432.129	20.963.201	23,0
60	0,0153	2,0138	0,0732	0,9268	870.479	63.743	4.162.044	16.531.072	19,0
65	0,0244	2,0591	0,1138	0,8862	806.736	91.792	3.763.728	12.369.029	15,3
70	0,0397	2,0351	0,1775	0,8225	714.944	126.885	3.198.522	8.605.300	12,0
75	0,0639	2,2307	0,2713	0,7287	588.059	159.546	2.498.466	5.406.778	9,2
80	0,1044	1,8739	0,3935	0,6065	428.512	168.618	1.615.445	2.908.312	6,8
85	0,1664	1,5586	0,5290	0,4710	259.894	137.496	826.284	1.292.867	5,0
90+	0,2623	2,7725	1,0000	0,0000	122.397	122.397	466.583	466.583	3,8

Ek 9: Türkiye Özetlenmiş Dönem Yaşam Tablosu, Kadın, 2011

Age	$_nm_x$	$_na_x$	$_nq_x$	$_np_x$	l_x	$_nd_x$	$_nL_x$	T_x	e_x^0
0	0,0111		0,0111	0,9889	1.000.000	11.124	988.876	79.433.927	79,4
1	0,0008	1,1197	0,0031	0,9969	988.876	3.067	3.946.668	78.445.052	79,3
5	0,0004	1,6520	0,0021	0,9979	985.808	2.027	4.922.254	74.498.384	75,6
10	0,0003	1,9364	0,0013	0,9987	983.781	1.326	4.914.844	69.576.130	70,7
15	0,0004	2,0667	0,0018	0,9982	982.455	1.752	4.907.139	64.661.287	65,8
20	0,0004	2,0187	0,0018	0,9982	980.704	1.717	4.898.401	59.754.148	60,9
25	0,0004	2,0998	0,0020	0,9980	978.987	1.989	4.889.167	54.855.747	56,0
30	0,0005	2,0107	0,0024	0,9976	976.998	2.325	4.878.040	49.966.579	51,1
35	0,0007	2,2329	0,0036	0,9964	974.673	3.529	4.863.601	45.088.539	46,3
40	0,0010	1,9983	0,0050	0,9950	971.144	4.833	4.841.213	40.224.938	41,4
45	0,0016	2,0034	0,0082	0,9918	966.311	7.895	4.807.895	35.383.726	36,6
50	0,0025	2,0047	0,0125	0,9875	958.416	11.968	4.756.230	30.575.831	31,9
55	0,0042	1,9259	0,0206	0,9794	946.448	19.506	4.672.273	25.819.600	27,3
60	0,0072	2,0855	0,0352	0,9648	926.941	32.650	4.539.550	21.147.327	22,8
65	0,0127	2,1558	0,0613	0,9387	894.292	54.817	4.315.545	16.607.778	18,6
70	0,0236	2,1384	0,1107	0,8893	839.475	92.943	3.931.414	12.292.233	14,6
75	0,0431	2,2475	0,1927	0,8073	746.532	143.891	3.336.601	8.360.819	11,2
80	0,0751	2,0048	0,3067	0,6933	602.642	184.836	2.459.590	5.024.218	8,3
85	0,1285	1,6246	0,4480	0,5520	417.806	187.196	1.457.176	2.564.628	6,1
90+	0,2082	3,2219	1,0000	0,0000	230.610	230.610	1.107.452	1.107.452	4,8

Volume: 10 Number: 02 Category: 02 Page: 25-41 ISSN No: 1303-6319 Cilt: 10 Sayı: 02 Kategori: 02 Sayfa: 25-41

BAYES FAKTÖRÜ, BAYESCİ BİLGİ ÖLÇÜTÜ VE SAPMA BİLGİ ÖLÇÜTÜ KULLANIMIYLA BAYESCİ MODEL SEÇİMİNİN BİR UYGULAMASI

Mutlu KAYA* **Emel ÇANKAYA****

ÖZET

İstatistiksel modelleme çalışmalarında, artan ileri teknoloji ve metodolojik gelişmeler sayesinde veriyi ürettiği varsayılan alternatif modeller oluşturabilmek mümkün olmaktadır. Dolayısıyla, mevcut rakip modeller arasından "en iyi" olanı seçme işlemi, modelleme sürecine dahil edilmesi gereken önemli aşamalardan biri olarak ortaya çıkmaktadır. Bu çalışmada, istatistiksel model seçimi probleminin Bayesci yaklaşımla çözümünde tercih edilen Bayes faktörü tanıtılmış, analitik olarak hesaplanmasının mümkün olmadığı durumlarda kullanılabilen Bayesci Bilgi Ölçütü (BIC) yanı sıra Markov Zinciri Monte Carlo (MCMC) simülasyonuna dayalı Carlin ve Chib yöntemi açıklanmıştır. Ayrıca Bayes faktöründen tamamen farklı prensipte çalışan ve son yıllarda model seçimi uygulamalarında sıklıkla kullanılan Sapma Bilgi Ölçütü (DIC) ayrıntılı olarak anlatılmıştır. Bir yarı-parametrik modelleme örneği olan kuantal modellemenin, literatürdeki bir uygulaması sonucu ortaya çıkan alternatif iki model Bayes faktörü, BIC ve DIC kullanılarak kıyaslanmıştır.

Anahtar Kelimeler: Bayes faktörü, Carlin ve Chib yöntemi, DIC, MCMC, BIC.

1. GİRİŞ

İstatistiksel bir modelin oluşturulması, değişkenler arasındaki ilişkinin matematiksel eşitlikler şeklinde formülasyonu olarak tanımlanabilir. Gerçek hayatta gözlenen bir olayı betimlemek için oluşturulan bir model, gelecekle ilgili tahmin yapmada önemli bir rol oynamaktadır. Bu yüzden, modelleme amaçlı yapılan istatistiksel bir çalışma; oluşturulan modelin yeterliliğinin, duyarlılığının ve alternatif modellerin varlığının test edilmesi işlemlerini içermelidir.

Günümüzde artan teknolojik ve metodolojik gelişmeler sayesinde daha karmaşık modeller oluşturulabildiğinden, aynı problemin çözümüne alternatif olabilecek model sayısında büyük bir artış olmuştur. Bu çalışmada, Box (1979, 202)'ın belirttiği "Aslında tüm modeller eksiktir ancak bazıları kullanışlıdır" ifadesini akılda tutarak, birbirine rakip modeller arasından "en kullanışlı" ya da "en iyi" olanı belirleme işlemi olan model seçimi problemi incelenecektir. Doğru modelin var olduğu varsayımı altında, model seçimi ifadesiyle, veriyi ürettiği varsayılan modelin oluşturulma işlemi değil, alternatif modellerin kıyaslanması işlemi kastedilmektedir.

Model seçimi problemi, regresyon modellemesi, karma modellemesi, çoklu değişim noktası problemi, değişkenlerin dağılımsal formunu belirlemek, hipotez testleri gibi istatistiksel çalışmalarda sıklıkla ortaya çıkmaktadır. Model seçimine klasik yaklaşımda, bilgi teorisine dayalı Akaike Bilgi Ölçütü (AIC), Bayesci Bilgi Ölçütü (BIC) ve Mallow'un Cp Ölçütü, tercih edilen yöntemlerden sadece birkaçıdır (Ucal, 2006).

*Arş. Gör., Sinop Üniversitesi, Fen-Edebiyat Fakültesi, İstatistik Bölümü, Sinop, e-posta: mutlu.alt@gmail.com

**Yrd. Doç. Dr., Sinop Üniversitesi, Fen-Edebiyat Fakültesi, İstatistik Bölümü, Sinop, e-posta: ecankaya@sinop.edu.tr

Çeşitli kaynaklardan elde edilen ön bilginin tahmin sürecine dahil edilmesine olanak sağlayan Bayesci yaklaşımda da model seçimi problemi popüler bir çalışma alanı olarak ortaya çıkmaktadır. Bayesci modelleme yapan araştırmacılar, bu amaç için sıklıkla Bayes faktörünü kullanmayı tercih etmişlerdir (Kass ve Raftery, 1995). İki rakip model kıyaslanmak istendiğinde kullanılan Bayes faktörü, modeller hakkındaki ön bilgiden son bilgiye geçişte modellerden biri lehine oddslarda ne kadarlık bir değişim olduğunun ölçüsüdür. Modeller hakkında bir ön bilginin olmaması durumunda dahi, modellerin doğruluğuna dair bir olasılık hesaplanmasına olanak tanıması ve bunların oranlanmasından elde edilmesi yorumu kolay, tutarlı ve otomatik olarak modelde "basitlik prensibine" sahip olması açısından tercih edilir bir yöntem olmuştur. Ayrıca $BF_{ij} = BF_{ik}\, BF_{kj}$ formülüyle çoklu model karşılaştırması da yapılabilmektedir (Martino, 2007).

Bayes faktörünün hipotez testlerindeki kullanımı Jeffreys (1961) tarafından, model seçimindeki kullanımı ise ilk olarak Schwarz (1978) ve Raftery (1986) tarafından ortaya konulmuştur. Mccullogh ve Rossi (1992) eşlenik önselleri; Kass ve Wasserman (1995) ise referans önselleri kullanarak, model karşılaştırmasında Bayes faktörü uygulamaları sunmuşlardır. Han ve Carlin (2001) Bayes faktörü hesabı için Markov Zinciri Monte Carlo (MCMC) metodunu, Sinharay ve Stern (2002), Bayes faktörünün önsel dağılımlara duyarlılığını incelemiş olup, Bayes faktörü ve alternatif çeşitlerinin kullanımı Araujo ve Pereira (2007)'nın çalışmasında bazı simülasyon uygulamalarıyla gösterilmiştir.

Marjinal olabilirlikler oranı olan Bayes faktörünün analitik olarak hesaplanmasının mümkün olmadığı durumlarda, Laplace yaklaşım yöntemi, önem örneklemesi, Gauss kareleştirme ve MCMC simülasyonu en çok kullanılan yöntemler arasındadır (Rosenkranz ve Raftery, 1994). Bileşik model-parametre uzayı tarama yöntemlerinden Carlin ve Chib (1995) yöntemi ise model belirsizliğini bir model göstergesi yardımıyla belirleyip, MCMC yöntemini kullanarak modellerin karşılaştırılmasında Bayes faktörünün elde edilmesini göstermiştir. Ayrıca belli şartlar altında, Schwarz (1978) tarafından geliştirilen ve model parametrelerine önsel dağılım belirlemeyi gerektirmeyen BIC (Schwarz ölçütü) yardımıyla, Bayes faktörünün kabaca bir tahmini elde edilebilmektedir (Kass ve Wasserman, 1995).

Bayes faktörü ile model karşılaştırmasının, model parametre sayısını bilmeyi gerektirmesi, parametre sayısının gözlemlerden sayıca fazla olduğu karmaşık hiyerarşik modellerde hesaplanmasının mümkün olmaması gibi problemler (Gelfand ve Dey., 1994; Kass ve Raftery, 1995), AIC'ye benzer prensipte çalışan yeni bir ölçütün, Sapma Bilgi Ölçütü (DIC) adıyla geliştirilmesine neden olmuştur (Spiegelhalter vd., 2002). Kullanımı son dönemlerde oldukça artan ve halen aktif araştırma konusu olan bu ölçütün özellikleri ve uygulamaları (Da Silva vd., 2004), Mislevy (2006), Spiegelhalter (2006a), Martino (2007), Asseburg (2007) ve Wilberg ve Bence (2008)'de bulunabilir.

Bu çalışmada, yarı parametrik modellemenin özel bir formu olan kuantal modellemenin, literatürdeki bir uygulaması sonucu ortaya çıkan alternatif iki modelin kıyaslaması, Bayes faktörü ve DIC hesaplanarak yapılacaktır. Bayes faktörü hesabı kullanım kolaylığı açısından Carlin ve Chib yöntemi ile yapılmıştır. İki yöntemin kullanım amaçlarındaki farklılıktan dolayı (Gelman vd., 2004; Spiegelhalter vd., 2002), sonuçların kıyaslanabilirliğini sağlamak açısından modeller için hesaplanan BIC

değerleri de çalışmada sunulmuştur. Tüm işlemler için Winbugs 1.4.3 paket programı kullanılmıştır.

2. YÖNTEM

2.1 Bayes Faktörü

Sonlu sayıda modeller arasından "en iyi" modeli seçme problemiyle ilgilenelim. Rakip modeller kümesi $\mathbb{M}$ olmak üzere $\mathbf{y}$ gözlenen veri setinin, M_j ($j \in \mathbb{M}$) modeli tarafından üretildiği varsayılsın. Her modelin farklı bilinmeyen parametreleri θ_j vektörü ile gösterilsin.

Model j için, $P(M_j)$, modelin önsel olasılığı $(\sum_j P(M_j) = 1)$; $P(\theta_j/M_j)$, model parametre vektörünün önsel dağılımı ve $P(y/\theta_j, M_j)$, olabilirlik fonksiyonu olarak tanımlanırsa bileşik dağılım,

$$P(y, \theta_j, M_j) = P(M_j)P(\theta_j/M_j)P(y/\theta_j, M_j) \quad (1)$$

olur.

Bileşik dağılımın θ_j parametre vektörüne göre marjinallenmesi ve gözlenen veri üzerine koşullandırılması sonucu her bir modele ilişkin sonsal model olasılıkları,

$$P(M_j/y) = \frac{P(y, \theta_j, M_j)}{P(y)} = \frac{P(M_j)P(y/M_j)}{P(y)}$$

$$P(M_j/y) \propto P(M_j)P(y/M_j) \quad (2)$$

ile hesaplanabilir (Martino, 2007). Burada $P(y/M_j)$, marjinal olabilirlik olup

$$P(y/M_j) = \int P(\theta_j/M_j)P(y/\theta_j, M_j)d\theta_j \quad (3)$$

şeklinde hesaplanır.

Alternatif model sayısının iki olduğu durumda, kıyaslama işlemi sonsal model olasılıklarının oranlanması yoluyla yapılır. Birbirine rakip M_1 ve M_2 modelleri için bu oran;

$$\frac{P(M_1/y)}{P(M_2/y)} = \frac{P(M_1)}{P(M_2)}\frac{P(y/M_1)}{P(y/M_2)}$$

dır. Burada marjinal olabilirlikler oranı Bayes faktörüdür. Bu eşitlik sözel olarak,

[Sonsal Odds] = [Önsel Odds] x [Bayes faktörü]

ile ifade edilirse,

$$\mathrm{BF}_{12} = \frac{\text{Sonsal Odds}}{\text{Önsel Odds}} = \frac{\mathrm{P}(\mathrm{M}_1/\mathrm{y})\mathrm{P}(\mathrm{M}_2)}{\mathrm{P}(\mathrm{M}_2/\mathrm{y})\mathrm{P}(\mathrm{M}_1)} \quad (4)$$

şeklinde odds oranı olarak tanımlanan BF_{12} oranına M_2'ye kıyasla M_1 lehine Bayes faktörü adı verilir (Raftery, 1995). Bir başka deyişle BF_{12}, önselden sonsala geçişte odds'larda M_1 lehine ne kadarlık bir değişim olduğunun bir ölçüsüdür. Modellerin tercih edilebilirliği konusunda bir ön bilginin olmaması durumunda, önsel olasılıklar $\mathrm{P}(\mathrm{M}_1) = \mathrm{P}(\mathrm{M}_2) = 0.5$ olarak seçilir ve dolayısıyla Bayes faktörü sonsal odds'a eşit olur. Sonsal model olasılıklarının toplamı bire eşit olduğundan formül 4,

$$\mathrm{BF}_{12} = \frac{\mathrm{P}(\mathrm{M}_1/\mathrm{y})}{\mathrm{P}(\mathrm{M}_2/\mathrm{y})} = \frac{\mathrm{P}(\mathrm{M}_1/\mathrm{y})}{1 - \mathrm{P}(\mathrm{M}_1/\mathrm{y})} \quad (5)$$

şeklinde yeniden ifade edilebilir. Hesaplanan Bayes faktörünün yorumu, Tablo 1 kullanılarak yapılır.

Tablo 1. Bayes faktörü değerinin model seçimindeki yorumu (Jeffreys, 1961)

Bayes faktörü değeri (BF_{ij}^*)	Yorum	Ok yönünde artan derecede
$BF_{ij} < 0.1$	M_j lehine güçlü kanıt	↑ M_j tercih edilir
$0.1 < BF_{ij} < 0.3$	M_j lehine makul kanıt	
$0.3 < BF_{ij} < 1$	M_j lehine zayıf kanıt	
$1 < BF_{ij} < 3$	M_i lehine zayıf kanıt	↓ M_i tercih edilir
$3 < BF_{ij} < 10$	M_i lehine makul kanıt	
$BF_{ij} > 10$	M_i lehine güçlü kanıt	

$^*BF_{ij}$: j. modele kıyasla i. model lehine hesaplanan Bayes faktörü değeri

Bayes faktörünün analitik olarak hesaplanabilmesi, eşitlik 3'de verilen integral işlemlerinin yapılmasını gerektirir. Modellerde yer alan bilinmeyen parametre sayısı arttığında, bu işlemleri gerçekleştirmek oldukça güçleşir hatta bazı durumlarda mümkün değildir. Karmaşık integrasyon teknikleri uygulamak yerine, koşullu dağılımlardan örneklem çekmek yoluyla parametre tahminlerinin elde edilmesini sağlayan MCMC yöntemi, model seçimi problemlerinde de etkin bir şekilde kullanılmaktadır (Gilks vd., 1996). Bu yaklaşıma sahip Carlin ve Chib yöntemi bir sonraki bölümde ayrıntılı olarak açıklanmıştır.

Kıyaslanmak istenen modellerin ön olasılıklarının eşit olduğu durumlarda, Bayes faktörünün yaklaşık hesabı BIC yardımıyla aşağıdaki formül kullanılarak da yapılabilmektedir (Kass ve Wasserman, 1995):

$$-2\ln BF_{ij} \cong BIC_i - BIC_j \quad \text{ya da} \quad 2\ln BF_{ji} \cong BIC_i - BIC_j \tag{6}$$

Burada n= örnek genişliği, p= modeldeki parametre sayısı ve $\hat{\theta}$= parametre vektörünün en çok olabilirlik tahmin edicisi olmak üzere, BIC_i ve BIC_j;

$$BIC = -2\log P\left(\frac{y}{\hat{\theta}}\right) + p\log(n) \tag{7}$$

formülü kullanılarak, i. ve j. modeller için ayrı ayrı hesaplanmış değerlerdir. Modeldeki parametre sayısının artması, bu ölçütün değerinin log(n) oranında büyümesine neden olacağından, BIC de basit ya da boyutu küçük modelleri tercih etme eğilimindedir. Doğru modelin olduğu varsayımıyla, en küçük BIC değerli modelin en iyi model olduğu yönünde yorumlanması önerilmektedir (Burnham, 2004).

2.2 Carlin ve Chib Yöntemi

Bayes faktörünün elde edilmesinde gerekli olan sonsal model olasılıklarının hesaplanmasına olanak sağlayan yaklaşımlardan biri, MCMC prensiplerini kullanan Carlin ve Chib (1995) yöntemidir. Bu yaklaşımda, model belirsizliği bir parametre olarak tanımlanır ve aldığı değerler bir gösterge değişkeni ile belirlenir. Bu parametre yardımıyla birbirine rakip k modelin örneklem uzayları birbirine bağlanır ve böylece MCMC örneklemesinin bu bileşik model uzayının bir formundan örneklemler alması sağlanır.

Model parametresi, tamsayı değerli M ile ve tüm modellere ait θ_j vektörlerinin birleşimi ise θ ile gösterilsin. $M = j$ olduğunda, k rakip model için bileşik dağılım,

$$P(y, \theta, M_j) = P(y/\theta_j, M_j) \prod_{j=1}^{k} P(\theta_j / M_j) P(M_j) \tag{8}$$

olur. Gösterge değişken M, hangi θ_j vektörünün y ile ilişkili olduğunu belirlediğinden, $M = j$ verildiğinde y, diğer modellerin $\{\theta_{i\neq j}\}$ vektöründen bağımsızdır. Ayrıca, modellere ait θ_j vektörlerinin de birbirinden bağımsız olduğu varsayılır. Bu durumda, j. model altında $P(\theta_j/M \neq j)$ için kullanılan dağılımsal form Bayesci model tanımlamasında önemli olmadığından, bunlar için "sözde (pseudo) önsel" seçimi yapılabilir.

Burada sözde önsel gerçek bir önsel olmayıp, bileşik model-parametre uzayını tanımlayabilmek için uygun şekilde seçilmiş bir bağlantı dağılımıdır. Gibbs örneklemesini yürütmek için gerekli θ_j'nin tam koşullu dağılımı,

$$P(\theta_j/\theta_{i\neq j}, M, y) \propto \begin{cases} P(y/\theta_j, M_j)P(\theta_j/M_j) & ; \quad M = j \\ P(\theta_j/M \neq j) & ; \quad M \neq j \end{cases} \tag{9}$$

olarak tanımlanır. Burada, $M = j$ olduğunda tüm koşullu olasılıklar geçerli olan model j'den, $M \neq j$ olduğunda ise sözde önselinden üretilir.

Sonuçta Gibbs örneklemesinin j. modeli ziyaret sayısı, tüm örneklemler sayısına orantılanarak sonsal model $P(M_j/y)$ olasılıklarının tahminleri,

$$\hat{P}(\mathrm{M_j/y}) = \frac{\mathrm{M^{(g)} = j}}{\mathrm{M^{(g)}'nin\ toplam\ sayısı}} \quad , \quad \mathrm{j = 1,2,\dots,k} \tag{10}$$

şeklinde elde edilebilir. Burada $\mathrm{M^{(g)}}$, Gibbs iterasyon sayısıdır (Carlin ve Chib, 1995). Bu tahmin değerlerinden elde edilen sonsal odds, önsel odds ile eşitlik 4'de ifade edildiği gibi birleştirilirse, modellerin ikişerli kıyaslamaları amaçlı Bayes faktörü hesaplanabilir.

2.3 Sapma Bilgi Ölçütü (DIC)

Klasik bir modelleme çalışmasında model kıyaslaması yapılmak istendiğinde, verinin modele uyumunu ölçen bir sapma istatistiği ile modeldeki parametre sayısının belirlediği model karmaşıklığı arasında denge kurmaya dayalı ölçütler kullanılır. Bunlardan bazılarının modeldeki parametre sayısını belirlemeyi gerektirmesi ve bazılarının ise parametreleri gözlemlerden sayıca üstün olan karmaşık hiyerarşik modellerde doğrudan uygulanamaması (Gelfand vd., 1992), son zamanlarda kullanımı yaygın olan alternatif bir Bayesci model seçim yöntemi Sapma Bilgi Ölçütünün geliştirilmesine sebep olmuştur.

DIC, iki bileşenden oluşmaktadır:

$$\begin{array}{ccccc} \mathrm{DIC} & = & \underset{\substack{\downarrow \\ \text{Kuramsal bir model ile örneklem verisi arasındaki uyumun bir ölçüsü}}}{\text{Uyum iyiliği}} & + & \underset{\substack{\downarrow \\ \text{Artan model parametre sayısı için bir sınırlama terimi}}}{\text{Karmaşıklık}} \end{array} \tag{11}$$

DIC'nin birinci bileşeni model yeterliliğinin bir tahmini olup, Dempster (1974) tarafından klasik sapma ifadesine dayalıdır.

$$\mathrm{Sapma} = \mathrm{D(\theta)} = -2\mathrm{logP(y/\theta)} + 2\mathrm{logP(y)} \tag{12}$$

Burada $\mathrm{P(y/\theta)}$; θ model parametre vektörüne dayalı olabilirlik fonksiyonu olup $\mathrm{P(y)}$ ise $\mathrm{P(y)} = \int \mathrm{P(\theta)P(y/\theta)d\theta}$ şeklinde sadece verinin fonksiyonu olan sabit bir terimdir. Bu terim model karşılaştırmasında tüm modeller için 1 alınarak aşağıda verilen formüllerin tümünde ihmal edilmesi sağlanır (Celeux vd., 2006).

Sapmanın sonsal dağılımının ortalaması uyum iyiliği ölçüsü olarak kullanılır. Bu ifade Dbar şeklinde adlandırılır ve $\overline{\mathrm{D(\theta)}}$ ile gösterilir.

$$\begin{aligned} \overline{\mathrm{D(\theta)}} &= \mathrm{E_{\theta/y}[D(\theta)]} \\ &= \mathrm{E_{\theta/y}[-2logP(y/\theta)]} \\ &= \frac{1}{\mathrm{C}}\sum_{\mathrm{c=1}}^{\mathrm{C}} -2\log_{\mathrm{e}} \mathrm{P(y/\theta_c)} \end{aligned} \tag{13}$$

Burada C, burn-in (yakınsama) periyodu çıkarılmış MCMC simülasyonlarının sayısı olup $\log_e P(y/\theta_c)$ ise olabilirlik fonksiyonunun doğal logaritmasıdır (Spiegelhalter vd., 2002). Bu eşitlikten görüldüğü gibi Dbar, Gibbs örneklemesinin bir iterasyonunun sonunda hesaplanmış log-olabilirliklerin ortalamasıdır.

DIC eşitliğinde yer alan ikinci bileşen ise, θ'nın sonsal ortalaması $(\hat{\theta})$ kullanılarak hesaplanan sapmaya dayalıdır. Dhat diye adlandırılan ve $D(\hat{\theta})$ ile gösterilen bu ifade,

$$D(\hat{\theta}) = D(E_{\theta/y}[\theta]) = -2\log P(y/\hat{\theta}) \qquad (14)$$

şeklinde tanımlıdır. Bir başka deyişle θ'nın sonsal ortalamasını kullanarak hesaplanmış log-olabilirliktir.

Veriyle en iyi uyumu sağlayan bir modelde yer alması gereken etkin parametre sayısı olarak ölçülen model karmaşıklığı, pD ile gösterilir ve

$$pD = \overline{D(\theta)} - D(\hat{\theta}) \qquad (15)$$

eşitliği ile elde edilir. Burada pD, modeldeki parametre sayısı için bir sınırlama terimidir ve model parametre sayısının yaklaşık değerini verir.

Bu tanımları kullanarak DIC ölçütü,

$$DIC = \overline{D(\theta)} + pD = D(\hat{\theta}) + 2pD \qquad (16)$$

şeklinde ifade edilir (Spiegelhalter vd., 2002).

Literatürde, modeldeki parametre sayısı aynı kalmak koşuluyla model parametrelerinin yeniden ifadelendirilmesi sonucu pD'nin değerinde büyük değişiklikler olabileceği konusunda yapılan uyarılar nedeniyle; Spiegelhalter ve Bull (1997), Gelman vd. (2004) tarafından pD yerine sapmanın sonsal varyansının yarısı şeklinde tanımlı pV'nin kullanımı önerilmiştir.

Matematiksel ifadeyle,

$$pV = 0.5\mathrm{Var}_{\theta/y}(D(\theta)) \qquad (17)$$

olarak gösterilir.

Veriyi ürettiği varsayılan iyi bir modelin olabilirliğinin büyük değerli olması, $\overline{D(\theta)}$'nın daha küçük değerlere ulaşmasını ve dolayısıyla DIC'nin küçük değerli olmasını sağlayacaktır. Sonuç olarak, en küçük DIC değerine sahip model, rakip modeller arasından en iyi model olarak seçilebilir.

DIC değeri daima pozitif olmayıp negatif değerler de verebilmektedir. Standart sapması küçük olan bir $P(y/\theta)$ olabilirliğinin sebep olduğu böylesi durumlarda negatif değerler de dikkate alınarak en küçük DIC'ye sahip model lehinde seçim yapılır (Spiegelhalter, 2006b).

Bayes faktörünün aksine DIC'nin mutlak büyüklüğünün model karşılaştırmasında bir önemi yoktur. Önemli olan sadece DIC değerleri arasındaki farkın mutlak büyüklüğüdür. (Spiegelhalter vd., 2002), minimum değerli DIC değeri ile arasında 2'den daha az fark olan modellerin eşit derecede iyi model olarak dikkate alınması gerektiğini, 2−7 arasında fark değerli modellerin ise daha az desteğe sahip modeller olarak değerlendirilmesini önermiştir.

3. BULGULAR

3.1 Kuantum Modelleme

İstatistiksel modellemenin özel bir formu, kuantal modelleme adı altında pek çok farklı disiplinde karşımıza çıkmaktadır. Bir yarı-parametrik modelleme örneği olan kuantal modellemede; verilen herhangi bir örneklem verisinin bir temel birim olan kuantumun tamsayı katları ($\pm$ rastgele hata) olarak ifade edilip edilemeyeceği sorusuna cevap aranmaktadır. Bu sorgudaki en önemli nokta böyle bir kuantum değerinin var olup olmadığının önceden bilinmemesidir. Ancak böyle bir değer varsa veriden tahmin edilmek yoluyla hesaplanmalıdır (Acar, 2000).

$\{Y_i\}_{i=1}^{n}$ örneklem verisi olmak üzere matematiksel olarak basit kuantum modeli,

$$Y_i = m_i q + \varepsilon_i \quad , \quad i = 1,2,\dots,n \quad ; \quad m_i \epsilon\, N \;\; \text{ve} \;\; -\frac{q}{2} \leq \varepsilon_i < \frac{q}{2} \qquad (18)$$

olarak ifade edilir. Burada q = kuantum, m_i = round (Y_i/q) şeklinde hesaplanan tamsayı değerler (round = bölümün en yakın tamsayıya yuvarlanması), ε_i hata değerleri sıfır civarında ve $\frac{q}{2}$ 'ye kıyasla küçük değerler alacaktır.

Kuantum modelleme ilk olarak, İngiltere, İskoçya ve Galler'de pek çok sayıda bulunan Stonehenge isimli, dairesel formdaki tarihi taş dikitlerin yarıçaplarının (Thom, 1955) tarafından ölçülmesi sonucunda ortaya çıkmıştır. Megalitik dönemi insanlarının bu yapıtları oluştururken, temel bir ölçü birimi ve onun katlarını kullanmış olabilecekleri savı ortaya atılmış ve klasik istatistiksel yöntemlerle analiz edilmesi sonucunda q = 5.44 inç değerinin kuantum ya da literatürdeki ünlü adıyla "Megalitik Yard" olabileceğine dair önemli bulgular elde edilmiştir (Kendall, 1974).

Bu problemin Bayesci yaklaşımla ilk analizi ise Freeman (1976) tarafından yapılmış ve kuantum için 5.44 yanı sıra 4 ve 7.5 değerlerinin de alternatif olabileceği çıkarsamasında bulunulmuştur. Modern Bayesci yöntemlerden MCMC kullanılarak problemin yeniden analizi (Acar, 2000) sonucunda elde edilen iki alternatif kuantum tahmin değeri q = 5.439 ve q = 7.480, Freeman (1976)'ın sonuçları ile tutarlıdır.

Bu çalışmada yukarıda özetlediğimiz çalışmalar sonucu elde edilmiş farklı kuantum tahmin değerlerinin oluşturduğu iki basit kuantum modelinin kıyaslanması Bayes faktörü, BIC ve DIC kullanılarak yapılmıştır. Literatürde, kuantal olma özelliğine en fazla sahip olduğu sıklıkla atfedilen good rings (n=16) veri seti (Thom, 1967), bu çalışma uygulaması için tercih edilmiştir.

Modeller matematiksel olarak,

$$\text{Model 1}: \; y_i = m_{1i}q_1 + \varepsilon_i \quad , \quad \varepsilon_i \sim N(0, \tau_1) \quad , \quad i = 1,2,\dots,16$$

$$\text{Model 2}: \; y_i = m_{2i}q_2 + {}_i \quad , \quad {}_i \sim N(0, \tau_2) \quad , \quad i = 1,2,\dots,16 \qquad (19)$$

şeklinde tanımlanır. Burada $\tau_1 = \sigma_1^{-2}$, $\tau_2 = \sigma_2^{-2}$ olup q_1 = 7.480 ve q_2 = 5.439 'dur.

3.2 Kuantum Modellerinin Carlin ve Chib Yöntemiyle Kıyaslanması

Model göstergesi M, Gibbs örneklemesi boyunca zincirin ürettiği j değerlerini (j = 1,2) kullanarak bileşik modellemeyi gerçekleştiren bir parametre olup kuantum modellerinden model 1 ve model 2'ye ilişkin bilinmeyen parametreler ise sırasıyla $\theta_1=(q_1, \tau_1)$ ve $\theta_2=(q_2, \tau_2)$'dir.

Basit kuantum modelleme çalışmalarında, sıfıra yakın çok küçük değerlerin gerçek kuantum olamayacağı ve ayrıca maximum gözlemden daha büyük değerli bir tahminin mümkün olmadığı vurgulanmış, "Megalitik Yard" örneği verisinin klasik yöntemlerle analizinde [q_{min}=2, q_{max}=10] aralığında tarama yapılarak bir tahmin sonucuna ulaşılmıştır (Kendal, 1974). Ancak daha güncel bir çalışmada (Acar, 2000), t=1/q değişkeninin Uniform dağılıma sahip olduğu ispatlanmış ve taramanın [$t_{min}=1/q_{max}$, $t_{max}=1/q_{min}$] aralığında yapılması önerilmiştir. Aynı verinin kullanıldığı bu çalışmada, bu parametre için önsel dağılım olarak 0.1 ve 0.5 parametreli Uniform dağılım seçilmiştir.

Herhangi bir verinin kuantal olduğundan bahsedilebilmek için; model hata teriminin (ε_i) eşitlik 18 ile tanımlı değerlerinin $\pm \frac{q}{2}$ 'ye kıyasla küçük ve "0" civarında yoğunlaşması, bir başka deyişle hata varyansının (σ^2), test edilen kuantum değerine kıyasla küçük olması gerekmektedir. Bu çalışmada kullanılan verinin, Freeman (1976) tarafından Bayesci yaklaşımla yapılan ilk çözümlenmesinde, hata varyansının "2" den büyük olması durumunda, veride var olabilecek bir kuantal yapının ortaya çıkarılamayacağından bahsedilmiş ve bu parametre için $\chi^2_{(2)}$ önsel dağılımı kullanılmıştır. Dolayısıyla çalışmanın ilerleyen bölümlerinde, σ^2 için bu önsel ile benzer değerler üretebilecek şekilde parametrize edilmiş bir Gamma önseli, Winbugs model formülasyonunda $\tau = \sigma^{-2}$ parametresi için kullanılmıştır.

Bu bilgiler ışığında, iki kuantal model kıyaslaması için gerekli önsel dağılımlar, $P(\theta_j/M_j)$;

$$P(t_1/M_1) = P(t_2/M_2) \sim U(0.1\,, 0.5) \qquad (20)$$

$$P(\tau_1/M_1) = P(\tau_2/M_2) \sim Ga(1\,, 0.5) \qquad (21)$$

şeklinde belirlenmiştir. Burada $t_1 = {}^1/_{q_1}$ ve $t_2 = {}^1/_{q_2}$'dir.

Bileşik model tanımlamasını sağlamak için gerekli sözde önseller, aynı veri setinin MCMC yaklaşımı ile çözümlenmesinden elde edilen (Acar, 2000) ve Tablo 2'de sunulan parametre tahminlerinin %95 Bayesci güven aralıklarından faydalanılarak elde edilmiştir.

Tablo 2. Yakınsaklık testlerini geçen parametrelerin farklı başlangıç değerleri için sonsal tahminleri

Zincir	Parametre	Başlangıç Değeri	Sonsal Ortalama	%95 Güven Aralığı
1	t	0.13414	0.134	[0.1323, 0.1352]
	q	---	**7.480**	[7.395, 7.559]
	τ	0.33128	1.371	[1.034, 1.707]
	σ	---	0.586	[0.3437, 0.9346]
2	t	0.16003	0.184	[0.183, 0.185]
	q	---	**5.439**	[5.410, 5.470]
	τ	0.67310	2.190	[0.949, 3.870]
	σ	---	0.709	[0.509, 1.030]

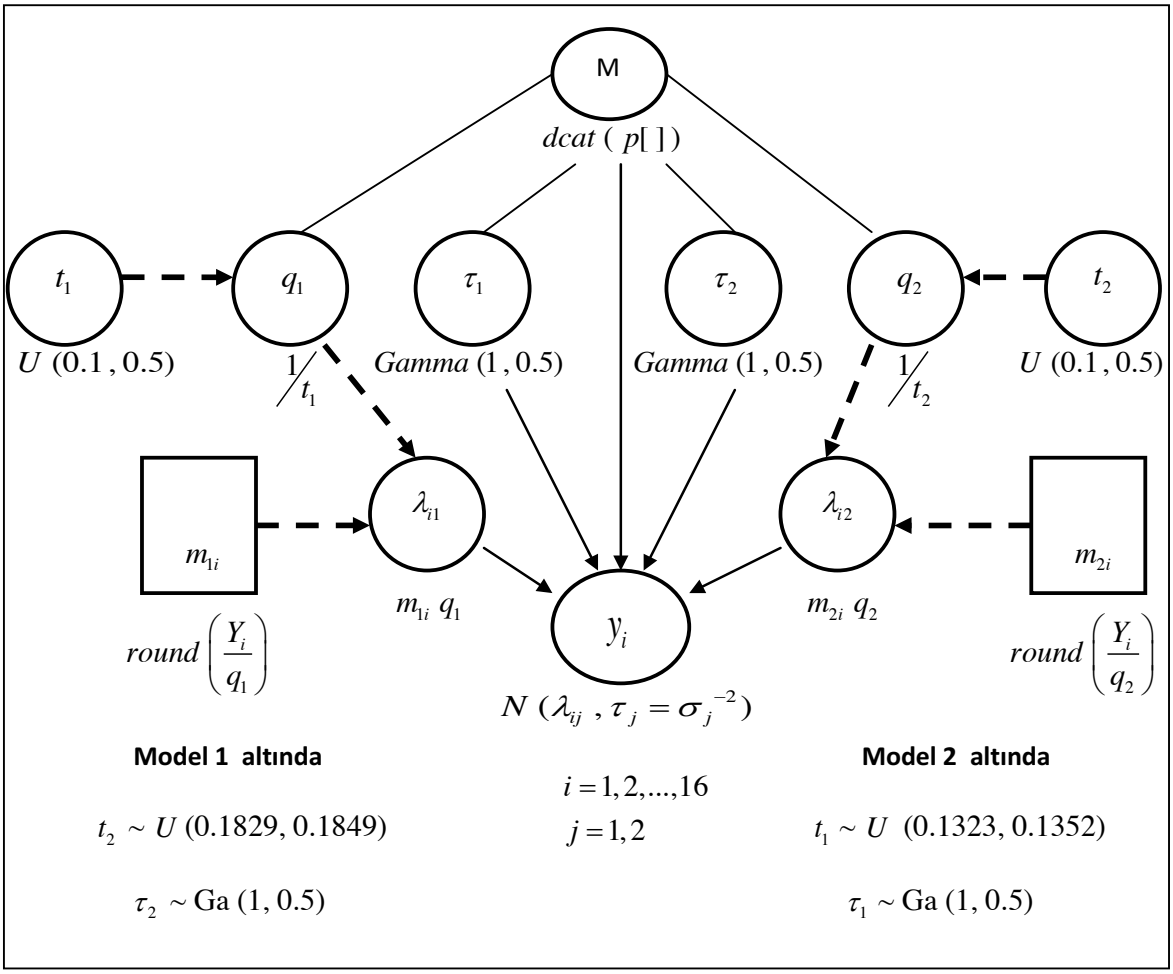
Not: Gibbs örneklemesi başlangıç değerleri ataması, sadece önsel dağılım tanımlanan model parametreleri t ve τ için gereklidir. Tahmin edilmek istenen q ve σ parametreleri için zincir değerleri, t=1/q ve $\tau = \sigma^{-2}$ deterministik ilişkiler kullanılarak oluşturulmaktadır.

Bu durumda pseudo önselleri $P\ (\theta_j / M \neq j)$,

Model 1 için (j=1), $t_1 \sim$ U (0.1323 , 0.1352) , $\tau_1 \sim$ Ga (1 , 0.5)
Model 2 için (j=2), $t_2 \sim$ U (0.1829 , 0.1849) , $\tau_2 \sim$ Ga (1 , 0.5)

olarak alınmıştır.

Formül 19'da verilen iki kuantum modelini, model göstergesi (M) yardımıyla kıyaslamak ve bileşik modellemeyi belirlemek amacıyla, modellerin Winbugs programı içindeki grafiksel gösterimi Şekil 1'de görüldüğü gibidir. Model parametresi M'nin alabileceği "1" ve "2" değerlerini üretmek amacıyla, Winbugs programında tanımlı *dcat* kategorik dağılımı kullanılmıştır. Dağılım parametresi $p[j]$; j= 1, 2 ise modellerin önsel olasılıklarına karşılık gelip, $\sum_j p[j] = 1$ sağlamaktadır. Şekildeki kesikli çizgili oklar ise deterministik ilişkileri göstermek amacıyla kullanılmaktadır.



Şekil 1. Carlin ve Chib yöntemi ile iki Kuantum modelinin kıyaslanmasının grafiksel gösterimi

Elde edilen bu bilgiler kullanılarak Carlin ve Chib yöntemi ile yazılan programın Winbugs Paket Programı'nda çalıştırılması sonucu 2. kuantum modelinin sonsal model olasılığının tahmini $P(M_2/y) = 0.5005$ olarak bulunur. (MC hata = 0.01259)

Bu durumda model 1'e kıyasla model 2 lehine Bayes faktörü formül 4'den,

$$BF_{21} = \frac{0.5005 \;\; 0.999995}{(1 - 0.5005)\; 0.000005} = 200399 \qquad (22)$$

olarak hesaplanır. Bu hesaplamada modellerin önsel olasılıkları, zincirin j değerini eşit sayıda üretmesini sağlamak amacıyla $P(M_1) = 0.999995$ ve $P(M_2) = 0.000005$ olarak seçilmiştir.

Elde edilen BF_{21} değerinin, Tablo 1'e göre 10'dan oldukça büyük bir değer olması sebebiyle, veriyi ürettiği varsayılan en iyi kuantum modelinin M_2 modeli; dolayısıyla en olası kuantum değerinin 5.439 olması yönünde güçlü bir kanıtın bulunduğu söylenebilir.

3.3 Kuantum Modellerinin BIC ve DIC Kullanılarak Kıyaslanması

DIC hesabı (formül 16) için gerekli olan sapma; $D(\theta)$ (formül 12), Dbar (formül 13), Dhat (formül 14) ve pD (formül 15) kullanılarak Winbugs paket programında yazılan programların çalıştırılması sonucunda her iki kuantum modeli için ayrı ayrı hesaplanan DIC değerleri, Tablo 3'te görüldüğü gibi elde edilmiştir. Bilgi ölçütü temelinde kıyaslamaya olanak tanıması açısından formül 7 ile tanımlanan BIC hesap değerleri de aynı tabloda sunulmuştur.

Tablo 3. M_1 ve M_2 modellerinin BIC ve DIC hesap değerleri

Model	Dbar	Dhat	p	pD	pV	$Var_{\theta/y}(D(\theta))$	DIC	BIC
M_1	57.662	55.706	2	1.957	1.930	3.8612	59.619	60.058
M_2	33.613	31.648	2	1.965	1.919	3.8377	35.578	36.018

Sonuçlar yorumlandığında, en küçük BIC ve DIC değerlerinin aynı modeli işaret ettiği, bir başka deyişle veri setini en iyi temsil eden kuantum modelinin M_2 modeli olduğu görülmektedir.

Model 1'e karşı model 2 lehine Bayes faktörü değeri ise formül 6 kullanılarak,

$$2lnBF_{21} \cong 60.058 - 36.018$$
$$\mathrm{BF}_{21} \cong 166042$$

şeklinde bulunur. Bu değerin, formül 22'de Carlin ve Chib yöntemiyle tam olarak hesaplanan Bayes faktörünün kaba bir tahmini olduğunu hatırlatarak, $BF_{21} = 166042 > 10$ olması sebebiyle M_1 modeline karşı M_2 modeli lehine kesin kanıt olduğu sonucuna bir kere daha ulaşılabilir.

4. TARTIŞMA ve SONUÇ

Günümüz istatistiksel modelleme çalışmalarında, ne kadar kompleks olursa olsun hemen hemen her problemin modellenmesinin mümkün olduğu görülmektedir. Mevcut modeller arasından en uygun modeli seçmek için bir araç gereksinimi gitgide arttığından halen pek çok metot geliştirilmektedir. Metotların çeşitliliği ve sayıca fazla olması problem çeşitliliğinden, gerekli analitik işlemlerin zorluk derecesinden ya da modelin oluşturulma amacından kaynaklıdır.

Bu çalışmada, model seçimi problemine Bayesci yaklaşımlardan Bayes faktörü ve Sapma Bilgi Ölçütü (DIC), ilkinin literatürde sıklıkla kullanılması ve diğerinin son dönemlerde geliştirilmiş olması sebebiyle tanıtılmış ve bir uygulaması gösterilmiştir. Bayes faktöründen tamamen farklı yapıda ve prensipte çalışan DIC'nin kıyaslanabilirliğini sağlamak amacıyla Bayesci Bilgi Ölçütü (BIC) de çalışmaya dahil edilmiştir. Bayes faktörünün analitik işlemleri için, MCMC simülasyonuna dayalı

Carlin ve Chib yöntemi uygulama kolaylığı açısından tercih edilmiştir. Bayes faktörünün BIC yardımıyla yaklaşık olarak hesabı da kısaca açıklanmıştır.

İki model karşılaştırmasında ve özellikle hipotez testlerinde sıklıkla tercih edilen Bayes faktörü oldukça kullanışlı olmasına rağmen, önsel dağılım seçimine ve modelin yeniden parametrize edilmesine karşı duyarsız değildir. Özelikle kıyaslanacak modellerin farklı boyutlarda olması, uygunsuz ya da muğlak önsel seçimi Bayes faktörünün kullanımını geçersiz kılmaktadır (Hall, 2012). Ayrıca parametre sayısı gözlem sayısından fazla olan hiyerarşik modellerin kıyaslanmasında Bayes faktörü uygulanamamaktadır. Bu dezavantajlardan bazılarının üstesinden gelmek için Bayes faktörünün versiyonları (sonsal, kısmi, içsel ve kesirli Bayes faktörü) geliştirilmiş olmasına rağmen, bunların kullanım koşulları da halen tartışma konusudur (Araujo ve Pereira, 2007).

BIC, farklı sayıda parametre içeren farklı modeller arasından seçim yapmak için kullanılması önerilen ölçütlerden biridir (Burnham, 2004). Artan değişken sayısının uyumda sağladığı iyileşmeyi, artan parametre sayısı ile log(n) oranında cezalandırdığından, modelde basitlik prensibine dayanır. Örneklem büyüklüğü arttığında tutarlı bir metottur. Bir başka deyişle, eğer alternatif modeller arasında doğru model var ise onu bulur. Parametreler için önsel dağılım belirtmeyi gerektirmediğinden, ön bilgi yetersizliğinde tercih edilebilmektedir.

Son dönemlerde, özellikle kompleks hiyerarşik modellemede karşılaşılan problemleri gidermek amacıyla geliştirilen DIC ölçütü, Bayes faktöründen tamamen farklı bir yapıda olup, AIC ve BIC ile benzer prensipte çalışmaktadır. Bu yaklaşım, veriye uygunluğu test edilen modelde yer alması gereken etkin parametre sayısının tahminini sağlaması, MCMC yöntemi ile kolaylıkla hesaplanabilmesi, önsel seçimine duyarlı olmaması gibi avantajlara sahip olmasına rağmen, her modelleme türüne uygulanamaması ve büyük örneklemlerde gereksiz sayıda parametreye sahip büyük modelleri tercih etmesi gibi dezavantajlara da sahiptir.

Bir modelleme çalışmasında hangi metot ya da metotlarla model seçimi yapılacağı, modellerin oluşturulma amacına göre belirlenmelidir. Eğer amaç öngörü yapmak ise, Bayes faktörü ve DIC (ya da BIC) kullanımıyla model kıyaslamasının aynı sonuca işaret etmesi beklenmemelidir. Çünkü ikisi de farklı amaca hizmet etmektedir. Bayes faktörü, sadece bir modelin doğru olduğu varsayımıyla, alternatif modeller arasından bu doğru modelin seçilmesi amacıyla kullanılır. DIC kullanımında ise amaç, diğer bilgi ölçütleri (AIC, BIC, AIC_c, vs) kullanımlarında olduğu gibi, doğru model olmasa bile veriyle en uyumlu modeli seçmektir. Özellikle hiyerarşik modellemede, öngörünün hiyerarşinin hangi seviyesi için yapılacağına bağlı olarak model seçim ölçütü tercihinin yapılması gerektiği önemle vurgulanmıştır (Spiegelhalter, 2006b).

İstatistiksel modelleme alanında farklı bir yere sahip kuantal modellemede ise amaç, gözlemlenen veride olması muhtemel bir yapının (kuantal olma) ortaya çıkarılmasına yönelik tanımlayıcı bir model oluşturmaktır. Literatürde "Megalitik Yard" adıyla

bilinen bir örneğinde, antik dönem insanlarının yapı inşaatlarında, günümüz metrik sistemindeki gibi yerleşik bir temel uzunluk ölçüsü (kuantum) kullanıp kullanmadıkları sorusuna cevap aranmaktadır. Burada model parametresi kuantumun birden fazla tahmini mümkün olduğundan, parametrenin hangi değerinin verinin quantal yapısını tanımlayan en iyi/en doğru değer olduğu tespiti Bayes faktörü, DIC ve BIC ile yapılabilir. Bu çalışmada, "Megalitik Yard" probleminin bir veri setine üç yöntemin uygulanması sonucu, Megalit yapı inşaatında kullanılmış olabilecek temel ölçü birimi için q=7.480'e kıyasla q=5.439 inç değerini destekler yönde ortak bulguya ulaşılmıştır.

Bu çalışma için seçilen yöntemlerden aynı doğrultuda sonuç elde edilmiştir. Ancak model seçim ölçütlerinin farklı sonuçlar verdiği bir başka uygulama çalışmasında; modelleme alanına ve amacına, örneklem genişliğine, modeldeki parametre sayısına, modellerin içiçe geçmiş olup olmadığına vs. bakılarak sonuçlar yorumlanmalıdır. Örneğin, büyük örneklemler için modelin gereksiz derecede büyük olma ihtimaline önlem olarak BIC tercih edilirken, küçük örneklemlerde AIC prensibinde çalışan DIC daha iyi sonuçlar vermektedir. Bayes Faktörü, BIC, AIC ve DIC model seçim yöntemlerinin simülasyona dayalı bir gözden geçirme çalışması ve daha ayrıntılı kıyaslamaları Ward (2008)'da bulunabilir.

5. KAYNAKLAR

Acar, E., 2000. Extensions of Quantal Problems. PhD. Thesis. University of Sheffield Department of Probability and Statistics, UK. (in English).

Araujo, M. I., Pereira, B.B., 2007. A Comparison of Bayes factors for Separated Models: Some Simulation Results. Communications in Statistics, Simulation and Computation, 36, 297-309.

Asseburg, C., 2007. An Introduction to Using WinBUGS for Cost-Effectiveness Analyses in Health Economics Centre for Health Economics, University of York, UK.

Box, G. E. P., 1979. Robustness in the Strategy of Scientific Model Building. In R.L.Launer & G. N. Wilkinson, (Eds.) Robustness in Statistics New York: Academic Press, 201-236.

Burnham, K. P. 2004. Multimodel Inference: Understanding AIC and BIC in Model Selection. Colorado Cooperative F&W Research Unit, Coloroda State University, Amsterdam Workshop on model selection, USA.

Carlin, B., Chib, S., 1995. Bayesian Model Choice via Markov Chain Monte Carlo Methods. J. Royal Statist. Society Series B, 57(3), 473-484.

Celeux, G., Forbes, F., Robert, C. P., Titterington, D. M., 2006. Deviance Information Criteria for Missing Data Models. Bayesian Analysis, 4, 651-674.

Da Silva, S. A., Melo, L. L. M., Ehlers, R., 2004. Spatial Analysis of Incidence Rates: A Bayesian Approach. Biostatistics, 1-17.

Dempster, A. P., 1974. The Direct Use of Likelihood for Significance Testing in Proceedings of Conference on Foundational Questions in Statistical Inference. University of Aarhus, 335-352.

Freeman, P. R., 1976. A Bayesian Analysis of the Megalithic Yard. J. R. Statist. Soc. A, 139, 20-55.

Gelfand, A. E., Dey, D. K., Chang, H., 1992. Model Determination Using Predictive Distributions with Implementation via Sampling-based Methods (with discussion). In Bayesian Statistics , Oxford University Press, 4, 147-167.

Gelfand, A. E., Dey, D. K., 1994. Bayesian Model Choice: Asymptotics and Exact Calculations. Journal Royal Statistics Soc. B., 56.

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

Gelman, A., Carlin, J. B., Stern, H. S., Rubin, D. B., 2004. Bayesian Data Analysis. Second Edition, Chapman and Hall / CRC, Boca Raton, FL.

Hall, B., 2012. Bayesian Inference. Statisticat, http://www.statisticat.com/laplacesdemon.html

Han, C., Carlin, B. P., 2001. MCMC Methods for Computing Bayes Factors: A Comparative Review. Journal of the American Statistical Association, 96, 1122-1132.

Jeffreys, H., 1961. Theory of Probability. Oxford University Press, Oxford, U.K.

Kass, R. E., Raftery, A. E., 1995. Bayes Factors. Journal of the American Statistical Association, 90, 773-795.

Kass, R. E., Wasserman, L., 1995. A Reference Bayesian Test for Nested Hypotheses and Its Relationship to the Schwarz Criterion. Journal of the American Statistical Association, 90(431), 928-934.

Kendall, D. G., 1974. Hunting Quanta. Phil. Trans. R. Soc. A, 276, 231-266.

Martino, S., 2007. Recent Methods for Bayesian Model Comparison. Department of Mathematical Science, NTNU.

Mccullogh, R., Rossi, P. E., 1992. A Bayesian Approach To Testing The Arbitrage Pricing Theory. Journal of Econometrics, 49, 141-168.

Mislevy, R. J., 2006. An Introduction to the DIC Index. University of Maryland.

Raftery, A. E., 1986. Choosing Modeles for Cross-Classiffications. American Sociological Review, 51, 145-146.

Raftery, A. E., 1995. Bayesian Model Selection in Social Research. Sociological Methodology, Marsden, P. V. Cambridge, Mass., Blackwells, 111-196.

Rosenkranz, S. L., Raftery, A. E., 1994. Covariate Selection in Hierarchical Models of Hospital Admission Counts: A Bayes Factor Approach No.268 Department of Statistics, GN 22 University of Washington Seattle, Washington, 98195 USA.

Schwarz, G., 1978. Estimating the Dimension of a Model. Annals of Statistics, 6(2), 461-464.

Sinharay, S., Stern, H. S., 2002. On the Sensitivity of Bayes Factors to the Prior Distributions. The American Statistician, 56(3), 196-201.

Spiegelhalter, D. J., Bull, K., 1997. Tutorial in Biostatistics Survival Analysis in Observational Studies. Statistics in Medicine, 16(9), 1041-1074.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., Van der Linde, A., 2002. Bayesian Measures of Model Complexity and Fit. Journal of the Royal Statistical Society, Series B, Methodological, 64(4), 583-616.

Spiegelhalter, D. J., 2006a. Two Brief Topics on Modelling with WinBUGS. MRC Biostatistics Unit, Cambridge.

Spiegelhalter, D. J., 2006b. Some DIC Slides. MRC Biostatistics Unit, Cambridge.

Thom, A., 1955. A Statistical Examination of the Megalithic Sites in Britain. J. R. Statist. Soc. A., 118, 275-295.

Thom, A., 1967. Megalithic Sites in Britain. Clarendon Press, Oxford.

Ucal, M. Ş., 2006. Ekonometrik Model Seçim Kriterleri Üzerine Kısa Bir İnceleme. C.Ü. İktisadi ve İdari Bilimler Dergisi, 7(2), 41-57.

Ward, E. J., 2008. A Review and Comparison of Four Commonly Used Bayesian and Maximum Likelihood Selection Tools. Ecological Modelling, 211, 1-10.

Wilberg, M. J., Bence, J. R., 2008. Performance of Deviance Information Criterion Model Selection in Statistical Catch-at-age Analysis. Fisheries Research, 93, 212-221.

AN APPLICATION OF THE BAYESIAN MODEL SELECTION BY USING BAYES FACTOR, BAYESIAN INFORMATION CRITERION AND DEVIANCE INFORMATION CRITERION

ABSTRACT

In statistical modelling studies, due to the advanced technology and methodological developments, it is possible to construct alternative models assumed to generate the data. Therefore, the process of choosing "the best model" among available competing models appears to be one of the crucial steps that has to be included in the modelling process. In this study, Bayes factor, which is a preferred Bayesian approach to the solution of statistical model selection problem, is introduced. For the cases when analytical computation of Bayes factor is not possible, in addition to Bayesian Information Criterion (BIC), Carlin and Chib method based on Markov Chain Monte Carlo (MCMC) simulation is explained. Besides, a frequently used criteria in the recent years of model selection applications, namely Deviance Information Criterion (DIC), which has a completely different working principle than Bayes factor, is described in detail. Two models appeared in the literature as a result of an application of quantal modelling, which is an example of a semi-parametric modelling, are compared by means of Bayes factor, BIC and DIC.

Keywords: Bayes factor, Carlin and Chib method, DIC, MCMC, BIC.

Cilt: 10 Sayı: 02 Kategori: 02 Sayfa: 42-55 ISSN No: 1303-6319 Volume: 10 Number: 02 Category: 02 Page: 42-55

MARMARA ÜNİVERSİTESİ ÖĞRENCİLERİNİN KREDİ KARTI SAHİBİ OLMALARINI ETKİLEYEN FAKTÖRLERİN BAYESCİ LOJİSTİK REGRESYON YARDIMIYLA İNCELENMESİ

Esin AVCI*

ÖZET

Bayesci yaklaşım, verilerden elde edilen yeni bilgi ile önceden bilinen bilginin derlenmesi ile oluşan bir yöntemdir. Bu çalışmada klasik yaklaşıma alternatif olan Bayesci yaklaşım, ikili sonuç değişkeni ile etki eden değişken(ler) arasındaki sebep-sonuç ilişkisini ortaya çıkaran Lojistik regresyona uygulanmıştır. Bu amaçla Marmara üniversite öğrencilerinin kredi kartına sahip olma durumlarına etki eden sosyoekonomik ve demografik faktörler incelenmiştir.

Anahtar Kelimeler: Kredi kartı sahipliliği, Bayesci lojistik regresyon, WinBUGS.

1. GİRİŞ

Kredi kartı; mülkiyeti kendilerine ait olmak üzere banka ya da finansal kuruluşların müşterilerine önceden belirlenen limitlerde, anlaşmalı işyerlerinden yurtiçi ve yurtdışında mal ve hizmet satın alma ile nakit ödeme birimleri veya otomatik ödeme makinelerinden nakit çekimlerde kullanılmak amacıyla verilen karttır. Artık modern dünyada çağdaş bir ödeme sistemi olan ve "plastik para" olarak adlandırılan kredi kartı; kredi kartını veren banka veya kuruluşun açtığı krediye istinaden kart sahibinin gereksinim duyduğu mal veya hizmeti o anda bir ödeme yapmadan satın alınmasına ve bedelini daha sonra herhangi ek bir mali külfet yüklenmeksizin ödeme yapmasına imkan veren bir ödeme aracıdır (Çavuş, 2006).

Kredi kartlarının, nakit dolaşım ihtiyacını azaltması, ekonominin kayıt altına tutulmasını kolaylaştırarak kayıt dışı hareketlerin önlenmesi, genel ekonominin kartlar sayesinde kağıt yükünden kurtulması ve tasarruf-yatırım akışının hızlanması sonucu ticari faaliyetlerin canlanması gibi ekonomik faydaları da söz konusudur (Turgay ve Başgül, 2007) .

Yeni gelişmeler paralelinde, kredi kartı sektörü de Türkiye'de hızlı bir büyüme trendi içerisine girmiştir. Türkiye'de hızla büyüyen kredi kartı sektörü, üniversite öğrencilerini de müşterileri arasına dahil edebilmek için yarışmaktadır. Üniversite öğrencilerinin kredi kartı kulanım tercihlerinin, sorunlarının, bu konudaki tutum ve davranışlarının tespit edilmesi bankalar açısından büyük önem taşımaktadır. Çünkü teknolojik gelişmelerle birlikte öğrencilerin kredi kartı sahipliği üzerinde etkili olan faktörler de değişiklik göstermektedir (Keskin ve Koparan, 2010).

Sosyal bilimlerde özellikle sosyo-ekonomik araştırmalarda, incelenen değişkenlerin bazıları hassas ölçekle ölçülmekle beraber, bazıları da olumlu-olumsuz, başarılı-başarısız, evet-hayır gibi iki şıklı verilerden oluşmaktadır. İki şıklı veriler, kategorik verilerin en yaygın olarak kullanılan şeklidir. Bağımlı değişkenin iki şıklı kategorik veriler olması durumunda bağımsız değişkenle (veya değişkenlerle) bağımlı değişken

*Yrd. Doç. Dr., Giresun Üniversitesi, Fen Edebiyat Fakültesi, İstatistik Bölümü, Giresun, e-posta: esin.avci@giresun.edu.tr

arasındaki sebep-sonuç ilişkisini incelerken Lojistik regresyon analizi kullanılır (Oktay vd., 2009). Lojistik regresyon modeli ilk olarak 1944 yılında Berkson tarafından kullanılmıştır. Cox (1970) bu modeli gözden geçirerek çeşitli uygulamalarını yapmış, özet gelişmeler ise ilk Anderson (1979, 1983) tarafından verilmiştir. Pregibon (1981) iki grup Lojistik modelde etkin (influential) aykırı (outlier) gözlemleri ve belirleme ölçütlerini, Lesaffre (1986), Lesaffre ve Albert (1989) ise çoklu grup Lojistik modellerde etkin ve aykırı gözlemlerle belirleme ölçülerini incelemişlerdir.

Lojistik regresyon modellerinin yaygın bir biçimde kullanılır hale gelmesi, katsayı tahmin yöntemlerinin geliştirilmesi ve Lojistik regresyon modellerinin daha ayrıntılı incelenmesine sebep olmuştur. Cornfield (1962), Lojistik regresyondaki katsayı tahmin işlemlerinde diskriminant fonksiyonu yaklaşımını ilk kez kullanarak popüler hale getirmiştir. Robert vd. (1987) Lojistik regresyonda standart Ki-kare, olabilirlik oran (G^2), "pseudo" en çok olabilirlik tahminleri, uyum mükemmelliği ve hipotez testleri üzerine araştırmalar yapmışlardır. Duffy (1990) Lojistik regresyonda hata terimlerinin dağılışı ve parametre değerlerinin gerçek değerlere yaklaşmasını incelemiştir. Başarır (1990) klinik verilerde çok değişkenli Lojistik regresyon analizi ve ayrımsama sorunu üzerine çalışmıştır.

Son yıllarda klasik yöntemlerdeki kısıtlamalar Bayesci yaklaşıma olan ilgiyi arttırmıştır. Bayesci yaklaşım subjektif düşüncenin temel taşı olarak kabul edilen Bayes teoremine dayanarak geliştirilmiştir ve yaklaşıma göre parametreler klasik yaklaşımdaki gibi sabit olarak değil, olasılığa bağlı olarak tanımlanır. Dolayısı ile her bir parametreye ilişkin bir dağılım söz konusudur. Bu olasılıklar "kanaat derecesi (degrees of belief)" olarak tanımlamaktadır. Bir başka deyişle parametreler rasgele değişken olarak ele alınmaktadır. "Bayesci" kelimesi de, parametre tahminleri için yapılan çıkarsamalarda Thomas Bayes'in teoreminden faydalanılmasından doğmuştur (Ibrahim vd., 2001).

Bayesci yaklaşım, karmaşık veriyi modellemede önsel (prior) bilgiye başvurma esnekliği nedeniyle klasik yöntemlere göre oldukça avantajlıdır (Ibrahim vd., 2001; Wong vd., 2005). Önsel bilginin elde edilmesi, Bayesci çıkarsamada önemli rol oynar. Önsel dağılımın sonuçları ne kadar değiştireceği model seçim kriterleri ile tespit edilmelidir. Önsel dağılımlar, açıklayıcı olan (informative) ve açıklayıcı olmayan (noninformative) olmak üzere iki temel gruba ayrılır. Açıklayıcı önsel bilgiler, daha önce yapılmış çalışmalardan elde edilen bilgiler, geçmiş deneyimler olarak belirlenirken, açıklayıcı olmayan önsel bilgiler ise parametrenin tanımlı olduğu aralık bilgisi dışında herhangi bir bilginin olmaması olarak tanımlanmaktadır. Bayesci yaklaşım, önsel bilgiler ışığında, gözlenen verinin subjektif yorumuna dayanır ve buradan hareketle elde edilen yeni bilginin bileşimi olan sonsal dağılımla açıklanır (Congdon, 2006; Wong vd., 2005). Hsu ve Leonard (1995) Lojistik regresyon fonksiyonlarında Bayes tahminlerinin elde edilmesi işlemleri üzerine çalışmışlar ve Bayesci Lojistik regresyonda Monte Carlo dönüşümünün kullanılabileceğini göstermişlerdir.

Çalışmanın amacı, Bayes tanımından yola çıkarak var olan bilginin yeni bilgi ile güncellenmesini göstermektir. Bu amaçla 2011 yılında Gaziosmanpaşa ve İnönü üniversite öğrencilerinin kredi kartı sahipliğini etkileyen faktörleri belirleme çalışmasında uygulanan 24 soruluk anket, Marmara üniversitesinde okuyan 200 öğrenciye uygulanmıştır. Marmara üniversitesi öğrencilerinden derlenen veriler, önsel bilgi olarak kullanılan Gaziosmanpaşa ve İnönü üniversite öğrencilerinin verileri ile

birleştirilerek Bayesci Lojistik regresyon uygulanmış ve kredi kartı sahipliğine etki eden faktörler belirlenmiştir.

2. YÖNTEM

2.1 Lojistik regresyon

İstatistiksel modellerin kullanıldığı birçok bilimsel araştırmada sonuçların analiz edilmesinde en çok lineer olmayan modeller kullanılmaktadır. Lojistik regresyon modeli lineer-olmayan modellerden en önemlilerindendir. Lojistik regresyon modeli bağımlı değişkenin kesikli; bağımsız değişkenlerin ise kesikli veya sürekli olması durumunda bağımlı değişkenle bağımsız değişkenler arsındaki sebep-sonuç ilişkisinin ortaya konulmasında kullanılmaktadır (Eskandari ve Meshkani, 2006).

Lojistik regresyon analizi, diskriminant analizi ve çoklu regresyon analizinden farklı olarak bağımsız değişkenlerin dağılımına ilişkin araştırmacılarca karşılanması gereken varsayımlar gerektirmez (Tabachnick ve Fidell, 1996). Bir başka deyişle bağımsız değişkenlerin normal dağılması, doğrusallık ve varyans-kovaryans matrislerinin eşitliği gibi varsayımların karşılanması gerekmez. Dolayısıyla da Lojistik regresyonun diğer iki teknikten çok daha esnek olduğu ifade edilebilir. Lojistik regresyonun yansız ve sapmasız istatistikler ortaya koyması için büyük örneklemler gerektirdiği bildirilmektedir. Özellikle bağımlı değişkenin ikiden fazla kategorisinin olduğu durumlarda, geçerli bir hipotez testi için, her bağımsız değişkende en az 50 kişilik bir grup büyüklüğüne ihtiyaç vardır. Bazı kaynaklarda bu sayının her bağımsız değişken için minimum 20, toplamda minimum 60 olması gerektiği vurgulanmaktadır. Diğer yandan örneklem büyüklüklerinin aynı olması durumunda, bağımlı değişkenin her bir kategorisinde bağımsız değişkenlerin çok değişkenli normalliğe sahip olması, her bir kategori için varyans ve kovaryansların eşitliği varsayımlarının karşılanması durumunda, daha önce de değinildiği gibi diskriminant analizi, Lojistik regresyon analizine tercih edilmelidir. Bununla birlikte, Lojistik regresyon analizi ile yapılan çözümlemeden elde edilen matematiksel modelin yorumlanmasının daha kolay olduğunu belirtmekte yarar vardır (Akkuş ve Çelik, 2004; Grimm ve Yarnold, 1995; Kalaycı, 2005; Leech vd., 2005; Poulsen ve French, 2008; Tabachnick ve Fidell, 1996; Tatlıdil, 1996).

Lojistik regresyon analizi adını, bağımlı değişkene uygulanan logit dönüşümünden (logit transformation) almaktadır. Bu durum aynı zamanda hem kestirim, hem de yorumlama sürecinde bazı farklılıklara neden olur (Hair vd., 2006). Lojistik regresyon analizi, bağımlı değişkenin ölçüldüğü ölçek türüne ve bağımlı değişkenin seçenek sayısına göre üçe ayrılmaktadır. Eğer bağımlı değişken iki seçenekli bir kategorik değişken ise "İkili Lojistik Regresyon Analizi (Binary Logistic Regression Analysis)" adını alır. Örneğin bir akademik programı bitirme durumuna göre öğrencilerin başarılı ve başarısız olarak nitelendirilmesi durumunda ikili Lojistik regresyon uygulanır. Eğer bağımlı değişken ikiden çok kategorili (düzeyli) sınıflamalı bir değişken ise "Çok Kategorili/Düzeyli İsimsel Lojistik Regresyon Analizi (Multinominal Logistic Regression Analysis)" adını alır. Örneğin üç farklı akademik programda öğrenim görmekte olan öğrencilerden oluşan bir bağımlı değişkenin olması durumunda, çok düzeyli isimsel Lojistik regresyon uygulanır. Eğer bağımlı değişken sıralama ölçeğiyle elde edilmiş ise, bu durumda da "Sıralı Lojistik Regresyon Analizi (Ordinal Logistic Regression Analysis)" kullanılır. Örneğin öğrencilerin öğrenim gördükleri akademik

programdaki başarılarının "düşük", "orta" ve "yüksek" olarak gruplandığı durumda sıralı Lojistik regresyon uygulanır. Lojistik regresyon, "tek değişkenli Lojistik regresyon (bağımsız değişkenin tek olduğu durum)" ve "çok değişkenli Lojistik regresyon (bağımsız değişkenin iki veya daha fazla olduğu durum)" olarak sınıflandırma yapılmaktadır (Stephenson, 2008).

Doğrusal regresyon analizinde bağımlı değişkenin değeri kestirilirken, Lojistik regresyon analizinde bağımlı değişkenin alacağı değerlerden birinin gerçekleşme olasılığı kestirilir. Bu olasılık değeri aşağıdaki model kullanılarak elde edilir.

$$\pi(x) = \frac{\exp(\beta_0 + \beta_1 x_1 + ... + \beta_p x_p)}{1 + \exp(\beta_0 + \beta_1 x_1 + ... + \beta_p x_p)} \quad (1)$$

Burada $x = (x_1, x_2, ..., x_p)$ bağımsız değişkenler vektörü ve $\beta = (\beta_1, \beta_2, ..., \beta_p)$ bağımsız değişkenlere ait parametre vektörünü göstermektedir.

İkili değer alan bağımlı değişken y(0,1) için Eşitlik 1'de verilen ifade, verilen x değeri için y 'nin 1'e eşit olma koşullu olasılığını vermektedir. Bu olasılık,

$$\pi(x) = P(y = 1|x) \quad (2)$$

eşitliği ile sağlanır. Benzer biçimde,

$$1 - \pi(x) = P(y = 0|x) \quad (3)$$

eşitliği y'nin 0'ı alma koşullu olasılığını göstermektedir.

Lojistik regresyon modelinde parametre tahmini yapılabilmesi için olabilirlik fonksiyonu öncelikle oluşturulmalıdır. y(0,1) bağımlı değişkeni $\pi(x)$ başarı olasılığı ile Bernoulli dağılmaktadır. Yukarıdaki eşitlik 1 ve 2'den *i*. birim için $y_i = 1$ olduğunda olabilirlik fonksiyonuna katkısı $\pi(x_i)$ ve $y_i = 0$ olduğunda olabilirlik fonksiyonuna katkısı $1 - \pi(x_i)$ kadardır. Buna göre *i*. birimin olabilirlik fonksiyonuna katkısı,

$$L(\beta_i) = \pi(x_i)^{y_i} (1 - \pi(x_i))^{(1-y_i)} \qquad i=1,2,\ldots,n \quad (4)$$

eşitliği ile ifade edilir. Gözlemlerin birbirinden bağımsız olduğu varsayıldığında olabilirlik fonksiyonu eşitlik 4'teki her bir birimin çarpılmasıyla elde edilir:

$$L(y|\beta) = \prod_{i=1}^{n} \pi(x_i)^{y_i} (1 - \pi(x_i))^{(1-y_i)} \qquad i=1,2,\ldots,n \quad (5)$$

Burada *n* birim sayısını göstermektedir.

Lojistik regresyon modelinin parametre tahmininde klasik yöntemler olarak; En Çok Olabilirlik, Yeniden Ağırlıklandırılmış Tekrarlı En Küçük Kareler ve tekrarlı veri

durumunda Minimum lojit Ki-Kare yöntemleri kullanılmaktadır (Murat ve Işığıçok, 2007). Klasik yönteme alternatif olarak Bayesci yöntemde kullanılmaktadır.

2.2 Lojistik Regresyon için Bayesci Yaklaşım

Bayesci yaklaşım, verilerden elde edilen yeni bilgi ile önceden bilinen bilginin derlenmesi ile oluşan bir yöntemdir (Wong vd., 2005). Klasik yaklaşımda tahmin yöntemi sadece veriden derlenen en çok olabilirlik fonksiyonuna dayanırken, Bayesci yaklaşımda klasik yöntemde elde edilen olabilirlik, parametre hakkında bilinen önsel bilgiyi ($p(\beta)$) değiştirmek için kullanılmaktadır. Buna göre Bayesci yaklaşıma göre parametre tahmini;

$$p(\beta / y) \propto L(y|\beta) \times p(\beta) \tag{6}$$

eşitliği ile verilen sonsal dağılım ile elde edilmektedir. Burada sonsal dağılım, parametre hakkındaki önsel bilgi ile veriden elde edilen olabilirlik fonksiyonunun birleşmesinden oluşan güncel bilgidir.

Bayesci yaklaşım, teorik çatı altında verinin önsel bilgi ile birleşiminin temel ve doğal bir yol sağlaması, asimptotik yaklaşım olmaksızın, veriden şartlı ve kesin çıkarımlar vermesi, küçük örnek büyüklüklerinde kullanılabilmesi, parametre tahminleri ve hipotez testlerinde doğrudan çıkarımlar yapması, kayıp veri ve hiyerarşik modeller için kolaylıkla uygulanabilmesi gibi birçok avantajının yanı sıra; önsel bilginin seçiminde kesin bir yöntem olmaması ve özellikle parametre sayısının fazla olduğu durumlarda hesaplama zorluğu dezavantajları arasında sayılabilmektedir (Cengiz vd., 2012).

Bir parametrenin önsel bilgisi, veri elde edilmeden önce parametre hakkındaki bilgileri ifade eder. Bu bilgiler bir dağılım ile ifade edilir. Bu nedenle Bayesci yaklaşıma göre parametreler klasik yaklaşımdaki gibi sabit olarak değil, olasılığa bağlı olarak tanımlanır. Önsel bilgi olmaksızın Bayesci çıkarsama yapılamaz. Önsel bilgiler; bilgi verici (informative) ve bilgi verici olmayan (non-informative) olmak üzere sınıflandırılırlar. Bilgi verici olmayan önseller, sonsal dağılım üzerinde minimum etkiye sahiptir. Bu önseller daha objektiftir olduklarından birçok istatistikçi tarafından kullanılırlar. Ancak, parametre hakkındaki toplam belirsizliğin objektif önselle verilmesi her zaman uygun değildir. Bazı durumlarda, objektif önseller kullanıcıyı yanlış sonsallara yönlendirebilir. Bilgi verici önseller, olabilirlik fonksiyonu tarafından etkisi azaltılmayan önsel dağılımdır. Bu tip önseller deneyimlerden, benzer geçmiş çalışmalardan ve uzman görüşlerinden elde edilen bilgi çerçevesinde belirlenirler. Sonsal dağılım üzerinde oldukça etkili oldukları için önsel dağılım belirlenirken çok dikkatli olunmalıdır. Genel olarak parametre hakkındaki belirsizliği en iyi açıklayacak dağılım;

$$\beta_j \sim N(\mu_j, \sigma_j^2) \qquad \text{j=1,2,...,p} \tag{7}$$

Bilgi olamaması halinde en çok, sıfır ortalamalı ve olabildiğince büyük varyans seçilmelidir. Varyans için 100 ile 10000 arasında bir aralık tercih edilmektedir (Rashwan vd., 2012).

Lojistik regresyon için Bayesci yaklaşım (6) Eşitliğine göre ifade edilirse,
n gözlem için olabilirlik fonksiyonu,

$$L(y|\beta)=\left[\left(\frac{\exp(\beta_0+\beta_1x_1+...+\beta_px_p)}{1+\exp(\beta_0+\beta_1x_1+...+\beta_px_p)}\right)^{y_i}\left(1-\frac{\exp(\beta_0+\beta_1x_1+...+\beta_px_p)}{1+\exp(\beta_0+\beta_1x_1+...+\beta_px_p)}\right)^{(1-y_i)}\right] \quad (8)$$

Bilinmeyen β parametrelerine ait önsel bilginin $\beta_j \sim N(\mu_j,\sigma_j^2)$ dağılımlı olduğu varsayıldığında sonsal dağılım,

$$p(\beta|y)=\prod_{i=1}^{n}\left[\left(\frac{\exp(\beta_0+\beta_1x_1+...+\beta_px_p)}{1+\exp(\beta_0+\beta_1x_1+...+\beta_px_p)}\right)^{y_i}\left(1-\frac{\exp(\beta_0+\beta_1x_1+...+\beta_px_p)}{1+\exp(\beta_0+\beta_1x_1+...+\beta_px_p)}\right)^{(1-y_i)}\right] \times\prod_{j=1}^{p}\frac{1}{\sqrt{2\pi\sigma_i}}\exp\left\{-\frac{1}{2}\left(\frac{\beta_j-\mu_j}{\sigma_j}\right)^2\right\} \quad (9)$$

elde edilir (Acquah, 2013). Eşitlik (9)'un analitik çözümü bulunmamaktadır. İstatistiksel çıkarım yapabilmek için her bir parametrenin marjinal dağılımlarının elde edilmesi gerekmektedir. Bunun için çoklu integrallerle işlem yapılması gerekmektedir. Uygulamada sonsal dağılımlardan istatistiksel çıkarımlar yapmak için simülasyon yöntemlerinden Markov Zinciri Monte Carlo (MCMC) yönteminin kullanımını yaygınlaştırmıştır. MCMC yöntemi, ilgilenilen sonsal dağılımdan başarılı bir şekilde bir öncekine bağlı örneklemler seçmektedir. MCMC yöntemi ücretsiz olarak indirilebilen WinBUGS paket programı yardımıyla kolaylıkla uygulanmaktadır. BUGS (Spiegelhalter vd., 1996), özel olarak MCMC yönteminin tam olasılık modellerine uygulanması için kullanılan ve kod yazımına dayanan bir programdır. Bu program Bayesci analizi sağladığından tüm parametreler rasgele değişken olarak ele alınmaktadır.

3. BULGULAR

Bu çalışmada, Marmara Üniversitesi'nde 2013 yılında eğitim görmekte olan öğrencilerin kredi kartı sahipliliği belirlemede etkili olan sosyo-ekonomik ve demografik faktörler Bayesci Lojistik regresyon yardımıyla belirlenmeye çalışılmıştır. Bu amaçla, 2011 yılında Gaziosmanpaşa ve İnönü üniversite öğrencilerinin kredi kartı sahipliğini etkileyen faktörleri belirleme çalışmasında uygulanan 24 soruluk anket, Marmara Üniversitesi Haydarpaşa ve Göztepe kampüsü'nde okuyan 200 öğrenciye uygulanmış ve veriler derlenmiştir. Anketin genel güvenirlilik katsayısı Cronbach Alfa katsayısı 0.733 olarak elde edilmiştir. Bu değer, kullanılan anketin oldukça güvenilir olduğunu göstermektedir.

Marmara Üniversitesi Haydarpaşa ve Göztepe kampüsü'nde okuyan 200 öğrencinin kredi kartı sahipliliğini belirlemede etkili olabilecek değişkenler; sınıf düzeyi, cinsiyet, yaş, kardeş sayısı, öğretim türü, kredi-burs durumu, anne ve babanın eğitim durumları, annenin çalışma durumu, ailenin aylık geliri, öğrencinin aylık geliri ve harcama tutarı gibi sosyo-ekonomik ve demografik göstergeler ele alınmıştır. Araştırmada yer verilen kredi kart sahipliliği (Bağımlı değişken) ve etki eden sosyo-ekonomik ve demografik

göstergelere (bağımsız değişkenlere) ait belirleyici istatistikler Tablo 1'de verilmektedir.

Tablo 1. Belirleyici istatistikler

Değişkenler	*Sayı*	*Yüzde*
KKS (Kredi Kartı Sahipliliği)		
1=Evet	92	46
2=Hayır	108	54
SNF (Sınıf Düzeyi)		
1=1. Sınıf	35	17.5
2=2.Sınıf	63	31.5
3=3. Sınıf	31	15.5
4=4. Sınıf	71	35.5
CNS (Cinsiyet)		
0=Bayan	117	58.5
1=Erkek	83	41.5
YAS (Yaş)		
1=17-20	66	33
2=21-23	111	55.5
3=24 ve üzeri	23	11.5
KRDS (Kardeş Sayısı)		
1=1-2	134	67
2=3-4	66	33
3=5 ve üzeri kardeş	0	0
OGT (Öğretim Türü)		
1=1.Öğretim	178	89
2=2.Öğretim	22	11
KRDB (Kredi veya Burs)		
1=Hayır	80	40
2=Evet	120	60
BEGT (Baba Eğitim Durumu)		
1=İlköğretim	73	36.5
2=Lise	65	32.5
3=Önlisans	13	6.5
4=Lisans	45	22.5
5=Yüksek Lisans	4	2
AEGT (Anne Eğitim Durumu)		
1=İlköğretim	114	57
2=Lise	58	29
3=Önlisans	4	2
4=Lisans	22	11
5=Yüksek Lisans	2	1
ACLS (Anne Çalışma Durumu)		
1=Hayır	149	74.5
2=Evet	51	25.5
AGLR (Ailenin Aylık Geliri)		
1=500 TL'den az	10	5
2=501-1000 TL	39	19.5
3=1001-1500 TL	43	21.5
4=1501-2000 TL	38	19

5=2001-2500 TL	26	13
6=2501 TL'den fazla	44	22
OGLR (Öğrencinin Aylık Geliri) TL		
1=300 TL'den az	74	37
2=301-600 TL	50	25
3=601 TL'den fazla	76	38
OHRC (Öğrencinin Aylık Harcama Tutarı) TL		
1=300 TL'den az	58	63
2=301-600 TL	23	25
3=601 TL'den fazla	11	12

Bayesci yaklaşımın uygulanmasında, ele alınan sosyo-ekonomik ve demografik göstergelere ait parametrelerin dağılımlarının (önsel bilginin) belirlenmesi amacıyla, 2011 yılında Yayar ve arkadaşları tarafından Gaziosmanpaşa ve İnönü Üniversitelerinde okuyan öğrencilerin kredi kart sahipliliğine etki eden faktörlerin saptanması çalışmasından yararlanılmıştır. Bu çalışmada, her iki üniversite verilerini birleştiren (Model-3) klasik Lojistik regresyon analizi sonucunda elde edilen parametre tahmin ve bu tahminlere ait standart hatalar kullanılmış ($\beta \sim N(\hat{\beta}_{2011}, \sigma^2_{2011})$) ve Tablo 2'de verilmiştir.

Tablo 2. Açıklayıcı değişkenlerin parametreleri için önsel bilgiler

Değişkenler	$\hat{\beta}_{2011}$	$\sigma_{\hat{\beta}_{2011}}$
Sabit	-4.359	0.687
SNF	0.022	0.089
CNS	0.051	0.163
YAS	0.635	0.160
KRDS	-0.173	0.132
OGT	0.277	0.169
KRDB	0.291	0.185
BEGT	-0.006	0.093
AEGT	0.050	0.150
ACLS	0.670	0.279
AGLR	0.122	0.074
OGLR	0.366	0.211
OHRC	0.367	0.214

Tablo 2'de verilen önsel bilgiler kullanılarak, Bayesci Lojistik regresyon yöntemi WinBUGS programında macro yazılarak uygulanmıştır. Açıklayıcı değişkenlerin parametrelerine ait sonsal dağılımın sonuçları Tablo 3'te verilmiştir.

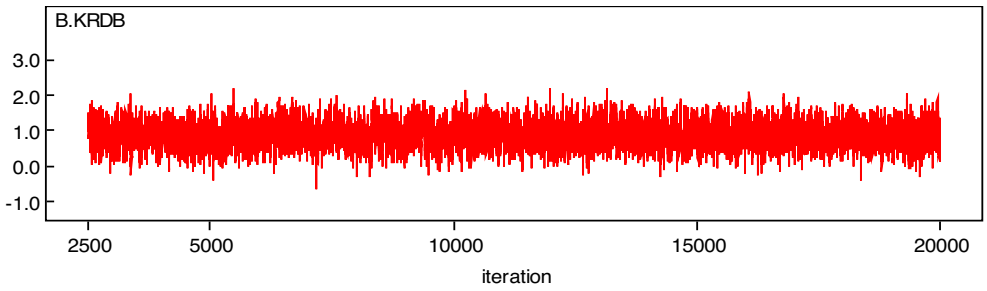
Tablo 3. Bayesci Lojistik regresyon analiz sonuçları

Parametreler	$\hat{\beta}$	$\sigma_{\hat{\beta}}$	*MC HATA*	*MC Hata/* $\sigma_{\hat{\beta}}$	*Alt sınır %2.5*	*Üst sınır %97.5*	*Başlama*	*Örnek*	*Exp(* $\hat{\beta}$ *)*
B.Sabit	-3.591	0.8186	0.0311	0.0380	**-5.206**	**-1.964**	2500	17501	0.0276*
B.SNF	0.499	0.1965	0.0063	0.0319	**0.1207**	**0.8976**	2500	17501	1.6471*
B.CNS	0.687	0.3487	0.0060	0.0173	**0.0053**	**1.373**	2500	17501	1.9877*
B.YAS	-0.344	0.346	0.0117	0.0338	-1.04	0.313	2500	17501	0.7089
B.KRDS	0.0091	0.3434	0.0106	0.03078	-0.6644	0.6762	2500	17501	1.0091
B.OGT	0.6514	0.5493	0.0060	0.0109	-0.4143	1.748	2500	17501	1.9182
B.KRDB	0.8901	0.3501	0.0062	0.0177	**0.2048**	**1.577**	2500	17501	2.4354*
B.BEGT	0.1316	0.1679	0.0041	0.0243	-0.1956	0.4618	2500	17501	1.1407
B.AEGT	-0.258	0.2279	0.0055	0.0243	-0.713	0.1885	2500	17501	0.7726
B.ACLS	0.5046	0.4372	0.0076	0.0173	-0.3485	1.368	2500	17501	1.6563
B.AGLR	0.1061	0.1308	0.0036	0.0271	-0.1419	0.3696	2500	17501	1.1119
B.OGLR	0.6807	0.2	0.0047	0.0235	**0.2911**	**1.085**	2500	17501	1.9753*
B.OHRC	0.2989	0.3430	0.0086	0.0247	-0.3734	0.9712	2500	17501	1.3484

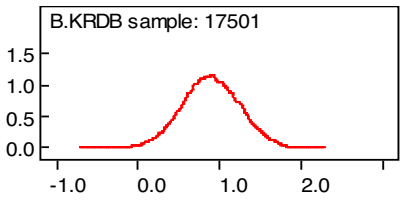
Sonsal dağılımın yakınsama sağlaması ve başlangıç değerlerinin etkisinin en aza indirilmesi için, Markov zincirinde elde edilen örneklemin başlangıç kısmı çıkarılır. Uygulamada 20000 iterasyonun ilk 2500 iterasyonu çıkarılmış ve elde edilen Bayesci Lojistik regresyon sonuçları Tablo 3'te verilmiştir. Tablo 3 incelendiğinde MC hatanın, sonsal standart hataya oranı %5'ten küçük olduğu için, Thumb kuralına göre yeterli iterasyon sayısına ulaşıldığı söylenir. %2.5 ve %97.5 güven aralıkları incelendiğinde, "0" değerini içeren aralığa sahip olan parametrelerin kredi kart sahipliliği üzerine bir etkisi olmadığı söylenir. Buradan hareketle, kredi kart sahipliliğine yaş, kardeş sayısı, öğretim türü, baba eğitim durumu, anne eğitim durumu, anne çalışma durumu, ailenin gelir durumunun önemli etkilerinin olmadığı saptanmıştır.

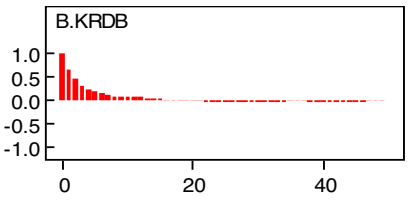
Sınıf düzeyi, cinsiyet, kredi veya burs alma durumu, öğrencinin gelirinin kredi kart sahipliliğine %5 önem seviyesinde önemli etkisi olduğu saptanmıştır. Sınıf düzeyindeki bir birimlik artış 1.65 kat veya %65 oranında, erkek öğrencilerin bayan öğrencilere göre yaklaşık 2 kat veya % 99 oranında, kredi ve burs alanların almayanlara göre yaklaşık 2.5 kat veya %104 oranında ve öğrencinin gelirindeki bir birimlik artışın yaklaşık 2 kat veya %98 oranında kredi kart sahipliliğini arttırdığı görülmektedir.

Bayesci yaklaşımda kullanılan MCMC yönteminin doğru sonuçlar verildiğine karar verebilmek için sonsal dağılıma yakınsamanın gerçekleştiğinin saptanması gerekir bu amaçla; iz, sonsal yoğunluk ve otokorelasyon grafiklerinden yararlanılır. Kredi kart sahipliliğine etki eden değişkenlerden kredi veya burs alma durumuna ait ilgili sırasıyla iz (a), Kernel yoğunluk (b) ve otokorelasyon (c) grafikleri Şekil 1'de verilmiştir.



(a)





(b) (c)

Şekil 1. İz, Kernel yoğunluk ve otokorelasyon grafikleri

İz grafiğinde salınım fazlalığı ve sıklığı sonsal dağılıma yakınsamanın hızlı bir şekilde gerçekleştiğini, Kernel yoğunluk grafiğinde ise, çan şeklinde oluşan görünümün sonsal dağılıma ulaşıldığını, Otokorelasyon grafiği ise Markov zinciri örnekleri arasındaki bağımlılığı ölçmekte, bu nedenle düşük korelasyon yakınsamanın sağlandığını belirtmektedir.

Klasik ve Bayesci yaklaşımın karşılaştırılması amacıyla Akaike Bilgi Kriteri (AIC) ve Bayesci Bilgi Kriteri (BIC) değerleri bütün veriler hesaplanmış ve Tablo 4'te verilmiştir.

Tablo 4. Klasik ve Bayesci Lojistik regresyon için AIC ve BIC değerleri

Kriter	Klasik	Bayesci
AIC	301.9775	267.44
BIC	344.8556	310.3181

Tablo 4'te yer alan sonuçlar incelendiğinde AIC ve BIC değerleri içinde en küçük değerin Bayesci Lojistik regresyon yaklaşımına ait olduğu görülür.

4. SONUÇ

Bu çalışmada, var olan bilginin güncellenmesini sağlayan Bayes yaklaşımının kullanılması amaçlanmıştır. Bu amaçla, 2011 yılında Gaziosmanpaşa ve İnönü üniversite öğrencilerine kredi kartı sahipliliğine etki eden faktörlerin belirlenmesi için uygulanan anket formu Marmara Üniversitesi öğrencilerine uygulanmıştır. Kredi kartı sahipliliğine etki eden faktörler, Bayesci Lojistik regresyon yaklaşımı ile analiz edilmiştir. Önsel bilgiler, 2011 yılında Gaziosmanpaşa ve İnönü Üniversiteleri için uygulanan klasik Lojistik regresyon analizi sonuçlarından elde edilmiştir. Bayesci Lojistik regresyon yaklaşımının yeterli yakınsama sağlaması ile ilgili Thumb kuralı, iz

grafiği, Kernel yoğunluk grafiği ve otokorelasyon grafikleri verilmiştir. Klasik yöntem ile karşılaştırılmasında AIC ve BIC kriterleri kullanılmıştır.

2011 yılında Yayar ve arkadaşları tarafından Gaziosmanpaşa ve İnönü Üniversitelerinde okuyan öğrencilerin kredi kart sahipliliğine etki eden faktörlerin saptanması çalışmasında her iki üniversite verilerini birleştiren (Model-3) klasik Lojistik regresyon analizi sonucunda elde edilen parametre tahmin ve bu tahminlere ait standart hatalar kullanılarak belirlenen önsel bilgiler yardımıyla uygulanan Bayesci Lojistik regresyon sonucunda yeterli yakınsamanın Thumb kuralı, iz grafiği, Kernel yoğunluk grafiği ve otokorelasyon grafikleri ile sağlandığı görülmüştür.

Kredi kartı sahipliliğine etki eden faktörlerin belirlenmesinde sınıf düzeyi, cinsiyet, yaş, kardeş sayısı, öğretim türü, kredi-burs durumu, anne ve babanın eğitim durumları, annenin çalışma durumu, ailenin aylık geliri, öğrencinin aylık geliri ve harcama tutarı gibi sosyo-ekonomik ve demografik göstergeler ele alınmıştır.

%5 önem düzeyine göre belirlenen güven aralığında sıfır'ı içermeyen (anlamlı bulunan) faktörlerin; sınıf düzeyi, cinsiyet, kredi veya burs alma durumu, öğrencinin gelirinin kredi kart sahipliliğine önemli etkisi olduğu saptanmıştır. Sınıf düzeyindeki artış 1.65 kat, erkek öğrencilerin bayan öğrencilere göre yaklaşık 2 kat, kredi ve burs alanların almayanlara göre yaklaşık 2.5 kat ve öğrencinin gelirindeki bir birimlik artışın yaklaşık 2 kat veya kredi kart sahipliliğini arttırdığı görülmektedir. Öğrencinin kredi kartı sahip olmalarında kendi gelirinin etkili ancak aile gelir durumun etkili olmamasının nedeni, öğrencilerin yarı zamanlı çalışmasına bağlanabilir.

Klasik ve Bayesci yaklaşımın karşılaştırılması amacıyla AIC ve BIC değerleri hesaplanmış ve en küçük AIC ve BIC değerlerinin Bayesci yaklaşıma ait olduğu saptanmıştır. Dolayısıyla bu çalışma için Bayesci yaklaşımın tercih edilebileceği göstermiştir.

Önsel bilgiye başvurma esnekliği nedeniyle Bayesci yaklaşım, klasik yöntemlere göre oldukça avantajlı olduğu yorumu yapılabilmektedir.

5. KAYNAKLAR

Acquah, H. D., 2013. Bayesian Lojistic Regression Modelling Via Markov Chain Monte Carlo Algorithm, Journal of Social and Development Sciences, 4(4), 193-197.

Akkuş, Z., Çelik, M. Y., 2004. Lojistik Regresyon ve Diskriminant Analizi Yöntemlerinde Önemli Ölçütler. VII. Ulusal Biyoistatistik Kongresinde sunulan bildiri, Mersin Üniversitesi, Tıp Fakültesi, Biyoistatistik Anabilim Dalı, Mersin.

Anderson, H. A., 1983. Robust Inference Using Lojistic Models, Bulletin of International Statistical Institute, 48, 25-53.

Anderson, J. A., 1979. Multivariate Lojistic Compounds. Biometrika, 66, 17-191.

Başarır, G., 1990. Çok Değişkenli Verilerde Ayrımsama Sorunu ve Lojistik Regresyon Analizi, (Uygulamalı İstatistik Doktora Tezi). Hacettepe Üniversitesi., 1-36, Ankara.

Cengiz, M. A., Terzi, E., Şenel, T., Murat, N., 2012. Lojistik Regresyonda Parametre Tahmininde Bayesci bir Yaklaşım, Afyon Kocatepe Üniversitesi Fen Bilimleri Dergisi, 12, 15-22.

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

Cornfield, J., 1962. Joint Dependence of the Risk of Coronay Heart Disease on Serum Cholestrol and Sistolic Blood Pressure: A Diskrimant Function Analysis, Federation Proceedings, 21, 58-61.

Cox, D. R., 1970. Analysis of Binary Data. London: Chapman and Hall (2nd ed. 1989 with E.J. Snell).

Çavuş, M. F., 2006. Bireysel Finansmanın Temininde Kredi Kartları: Türkiye'de Kredi Kartı Kullanımı Üzerine bir Araştırma, Selçuk Üniversitesi Sosyal Bilimler Enstitüsü Dergisi, 15, 173-187.

Duffy, D. E., 1990. On Continuity-corrected Residuals in Lojistic Regression, Biometrika, 77, 287-293.

Eskandari, F., Meshkanı M. R., 2006. Bayesian Logistic Regression Model Choice Via Laplace-Metropolis Algorithm, Journal of the Iranian Statistical Society (JIRSS), 5(1-2), 9-24.

Grimm, L. G., Yarnold, P. R., 1995. Reading and Understanding Multivariate Statistics, Washington D.C.: American Psychological Association.

Hair, J. F., Black, W. C., Babin, B., Anderson, R. E., Tatham, R. L., 2006. Multivariate Data Analysis (6th ed), Upper Saddle River, NJ: Prentice-Hall.

Hsu, J. S., Leonard, T., 1995. Hierarchical Bayesian Semiparametric Procedures for Lojistic Regression, Biometrika, 84, 85-93.

Ibrahim, J. G., Chen, M. H., Sinha, D., 2001. Bayesian Survival Analysis, Springer-Verlag, New York.

Kalaycı, Ş., 2005. SPSS Uygulamalı Çok Değişkenli İstatistik Teknikleri, Ankara: Asil Yayın Dağıtım.

Keskin, D., Koparan, E., 2010. Üniversite Öğrencilerinin Kredi Kartı Sahipliliğini Belirleyen Faktörler, Eskişehir Osmangazi Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi, 5(1), 111-129.

Leech, N.L., Barrett, K.C., Morgan, G.A., 2005. SPSS for Intermediate Statistics: Use and Interpretation (2nd ed), Mahwah, NJ: Lawrence Erlbaum Associates.

Lesaffre, E., 1986. Lojistic Discriminant Analysis with Applications in Electrocardiography, Phd Thesis, Katholieke Universiteit Leuven, Belgium, 354. Unpublished.

Lesaffre, E., Albert, A., 1989. A Multiple Group Lojistic Regression Diagnostics, Applied Statistics, 38 (3), 425-440.

Murat, D., Işığıçok, E., 2007. 2007 Seçim Döneminde Ekonomik ve Siyasi Duruma İlişkin Beklentiler: Bursa uygulaması, 8. Türkiye Ekonometri ve İstatistik Kongresi 24-25 Mayıs 2007. İnönü Üniversitesi, Malatya.

Oktay, E., Özen, Ü., Alkan, Ö., 2009. Kredi Kart Sahipliğinde Etkili Olan Faktörlerin Araştırılması: Erzurum örneği, Dokuz Eylül Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi, 24(2), 1-22.

Poulsen, J., French, A., 2008. Discriminant Function Analysis. http://userwww.sfsu.edu/~efc/classes/biol710/discrim/discrim.pdf adresinden 22 Kasım 2008 tarihinde edinilmiştir.

Pregibon, D., 1981. Lojistic Regression Diagnostics, The Annals of Statistics, 9, 705-724.

Rashwan, N. I., El Dereny, M., 2012. The Comparision Between Results of Application Bayesian and Maximum Likelihood Approaches on Lojistic Regression Model for Prostate Cancer Data, Applied Mathematical Sciences, 6 (23), 1143-1158.

Robert, G., Rao, N. K., Kumar, S., 1987. Lojistic Regression Analysis of Sample Data, Biometrika, 35, 58-79.

Spiegelhalter, D., Thomas, A., N. Best, N. Gilks, W., 1996. BUGS 0.5: Examples Volume 1, MRC Biostatistics Unit, Institute of Public Health, Cambridge.

Stephenson, B., 2008. Binary Response and Logistic Regression Analysis, www.public.iastate.edu/~stat415/stephenson/stat415_chapter3.pdf adresinden 22 Kasım 2008 tarihinde edinilmiştir. Şekercioğlu, G. (2008). Fatalizm.

Tabachnick, B. G., Fidell, L. S., 1996. Using Multivariate Statistics (3rd ed.), New York, USA: Harper Collins College Publishers.

Tatlıdil, H., 1996. Uygulamalı Çok Değişkenli İstatistiksel Analiz, Ankara: Engin Yayınları.

Turgay, O., Başgül, N., 2007. Önemli Bir Finanasman Kaynağı Olarak Kredi Kartları: Kredi Kartlarının Kart Sahiplerinin Harcamaları Üzerindeki Etkilerini Belirlemeye Yönelik Burdur İlinde Bir Araştırma, Süleyman Demirel Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi, 12, 215-226.

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.

Yayar, R., Karaca, S. S., Turkut, A., 2011. Üniversite Öğrencilerinin Kredi Kart Sahibi Olmaları Üzerinde Etkili Olan Faktörler: Gaziosmanpaşa ve İnönü Üniversitelerinden Ampirik Bulgular, Akademik Yaklaşımlar Dergisi, 2(1),152-169.

INVESTIGATION OF FACTORS EFFECTIVE ON CREDIT CARD OWNERSHIP OF MARMARA UNIVERSITY STUDENTS BY BAYESIAN LOGISTIC REGRESSION

ABSTRACT

Bayesian approach is a method which combines new information obtained from data and previous knowledge. In this study, Bayesian approach, which is an alternative to classical approach, is applied to logistic regression, which explores the effects of covariate(s) on binary dependent variable. To this aim, socioeconomic and demographic factors that are effective on credit card ownership of Marmara University students are examined.

Keywords: Credit card ownership, Bayesian logistic regression, WinBUGS

Cilt: 10 Sayı: 02 Kategori: 02 Sayfa: 56-72 ISSN No: 1303-6319 Volume: 10 Number: 02 Category: 02 Page: 56-72

YENİ DOĞAN BEBEKLERİN DÜŞÜK DOĞUM AĞIRLIĞININ MARS YÖNTEMİNE DAYALI İKİLİ LOJİSTİK REGRESYONLA MODELLENMESİ

Soner ÖZTÜRK* **Volkan SEVİNÇ****

ÖZET

Düşük doğum ağırlıklı bebekler ilerleyen yıllarda sağlık açısından bazı sorunlarla karşılaşmaktadır. Bu yüzden bir bebeğin doğmadan önce düşük doğum ağırlıklı olup olmayacağının tahmini önemlidir. Bu tahminin elde edilebilmesi için ihtiyaç duyulan bir modelin geliştirilmesinde lojistik regresyon modeli uygun bir seçimdir. Lojistik regresyon analizi, bağımlı değişkenin kategorik olduğu durumlarda kullanılan ve kolay yorumlanabilen modelleme tekniklerinden birisidir. Bağımlı değişkenin iki düzeyli olduğu lojistik regresyon analizi ikili lojistik regresyon analizi olarak adlandırılır. Lojistik regresyon analizinin parametrik ve parametrik olmayan çözümleri bulunmaktadır. MARS yöntemi parametrik olmayan ve lojistik regresyon analizinde parametrik çözümlere alternatif olarak kullanılabilecek bir çözüm yöntemidir. Parametrik olmayan modeller, parametrik modellere göre daha az varsayım gerektirir ve daha esnektirler. Uygulama çalışmasında doğacak bebeklerin düşük doğum ağırlıklı olup olmayacağının tahmin edilmesini sağlayacak bir ikili lojistik regresyon modeli oluşturulmuştur. Model MARS yöntemine dayalı olarak tahmin edilmiştir. Analizde 982 bireye ait veri, MARS paket programı kullanılarak incelenmiştir. Sonuç kısmında elde edilen bulgular yorumlanmıştır.

Anahtar Kelimeler: Düşük doğum ağırlığı, Lojistik regresyon, En çok olabilirlik, MARS.

1. GİRİŞ

Eldeki örnek verilerden hareket ederek yorumlama, genelleme ve tahmin yapmak istatistiğin temel konularıdır. İstatistikte bu amaçlara yönelik yöntemlerden biri regresyon analizidir. Regresyon analizi, bir bağımlı değişken ile bir ya da birden fazla bağımsız değişken arasındaki ilişkiyi ifade etmekte kullanılan bir istatistiksel yöntemdir. Regresyon analizi ilişkinin türü, gücü ve yapısını araştırmaktadır. Regresyon analizinin en temel türü, doğrusal regresyon analizidir. Doğrusal regresyon analizinde bağımlı değişken kesikli veya sürekli değerler alabilir. Ancak bazı durumlarda bağımlı değişken kategorik değerler de alabilmektedir. Bağımlı değişkenin kategorik olduğu durumlarda kullanılan regresyon türü lojistik regresyondur. Lojistik regresyonda bağımsız değişkenler sayısal veya kategorik değerler alabilirler. Bağımlı değişkenin sadece iki değer aldığı lojistik regresyon analizi iki düzeyli, ikiden fazla değer alabildiği analiz, çoklu lojistik regresyon analizi olarak adlandırılır.

Lojistik regresyonun parametrik çözümünde kullanılan en yaygın yöntem en çok olabilirlik yöntemidir. Parametrik yöntemlere alternatif olarak parametrik olmayan yöntemler de bulunmaktadır. Parametrik olmayan yöntemler, veri sayısının ve değişken sayısının çok olduğu, kayıp verilerin olduğu durumlarda iyi sonuçlar vermektedir. Bu

*Muğla Sıtkı Koçman Üniversitesi, Fen Bilimleri Enstitüsü, Muğla, e-posta: soner985@hotmail.com
**Yrd. Doç. Dr., Muğla Sıtkı Koçman Üniversitesi, Fen Fakültesi, İstatistik Bölümü, Muğla, e-posta: vsevinc@mu.edu.tr

yöntemlerden biri çok değişkenli uyarlayıcı regresyon uzanımları: MARS (Multivariate Adaptive Splines) yöntemidir.

Düşük doğum ağırlıklı olarak dünyaya gelen bebekler ilerleyen yıllarda sağlık açısından bazı sorunlarla karşılaşmaktadır. Bu nedenle, bir bebeğin doğum ağırlığının düşük değerde olup olmayacağının tahmini için uygun bir modele ihtiyaç vardır. Bu modelin oluşturulmasını içeren uygulama çalışmasında, bebeğin doğum ağırlığını etkileyebilecek faktörler, MARS yöntemine dayalı lojistik regresyon modeli oluşturmak için kullanılmıştır. Çalışmada Danimarka Ulusal Doğum Grubu'nun hamilelikte ateş ve buna bağlı ölü doğumlarla ilgili çalışmasına ait verilerin bir kısmı kullanılmıştır. Bebek doğum ağırlıkları bağımlı değişken olarak seçilmiştir. Çalışmada doğum yapacak kadınlardan elde edilen veriler kullanılarak, bebeğin düşük doğum ağırlıklı olup olmayacağına ilişkin model oluşturulmuştur. Oluşturulan model yardımıyla doğum öncesi bebek doğum ağırlığı için tahminler yapılabilmesi ve gerekli önlemlerin alınabilmesi amaçlanmıştır.

2. MARS YÖNTEMİ

Çok değişkenli uyarlayıcı regresyon uzanımları (MARS), Friedman tarafından 90'lı yılların başında geliştirilmiş, parametrik olmayan bir regresyon yöntemidir. MARS kelimesi açılımı aşağıdaki kavramların baş harflerinden oluşturulmuştur.

Multivariate (çok değişkenli): Çok boyutlu veriler üzerinde işlem yapılabilir, özellikle bağımsız değişken sayısı fazla olduğu durumlarda tercih sebebidir.

Adaptive (uyarlayıcı): Yöntemin basamakları final modeline ulaşana kadar eleme ve seçme aşamalarından oluşur.

Regression (regresyon): Bağımlı ve bağımsız değişkenler arasındaki fonksiyonel ilişkiyi ifade etmektedir.

Splines (uzanımlar): Regresyon eşitliği, düz bir regresyon doğrusu yerine, bükülmüş bir yapıya sahiptir.

MARS yöntemi bankacılık, sigortacılık, ekonominin yanı sıra, yaşam analizi, sosyal bilimler gibi birçok alanda kullanılmaktadır. Literatürde MARS yöntemi ile ilgili olarak, Kim (2000) gençlerin uyuşturucu kullanımı ile ilgili yaptığı çalışma sonucunda, bağımlı değişkenin kategorik olduğu durumlarda da MARS'ın iyi sonuçlar verdiğini göstermiştir. Kuhnert, Do vd. (2000) parametrik olmayan CART ve MARS yöntemlerini, parametrik lojistik regresyonla karşılaştırmıştır. Motor kazalarındaki yaralanma verilerine uygulanmış olan bu çalışma için MARS modelinin diğer ikisine göre daha iyi performans gösterdiği belirtilmiştir. Amerikan Çevre Koruma Kuruluşu (EPA)'ndan Nash ve Bradford (2001)'un yaptığı çalışmada belirli bir bölgedeki bir kurbağa türünün varlığı lojistik regresyon ve MARS yöntemiyle tahmin edilmiş ve iki yöntemin sonuçları değerlendirilmiştir. Kolyshkina ve Brookes (2002) sigorta riskini veri madenciliği yaklaşımlarını MARS ve klasik lojistik regresyonla tahmin etmeye çalışmıştır. Dieterle (2003) zamana bağlı analitik veriler üzerine hazırladığı doktora tezinde yapay sinir ağları, genetik algoritmalar, CART ve MARS'ı karşılaştırmıştır. Lee, Chiu vd. (2004) kredi skorlama ile ilgili çalışmalarında, diskriminant analizi,

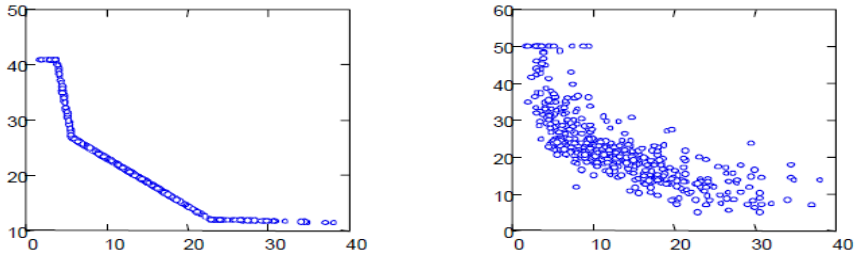
lojistik regresyon, CART ve MARS'ın doğru sınıflama oranlarını ve hatalarını karşılaştırmışlardır. Stokes ve Lattyak (2005) MARS yöntemini ekonometrik bazı sistem ve yazılımlar ile geliştirmiş ve kullanmışlardır. Kriner (2007), yaşam analizini MARS yöntemini kullanarak yapmıştır. Quiros, Felicisimo vd. (2009) MARS yöntemini, arazi örtüsünün uydu görüntülerinden yararlanarak sınıflandırılması için kullanmışlardır. Mina (2009), yoksulluk profilinin belirlenmesinde MARS yöntemini kullanmış ve bazı koşullar altında parametrik lojistik regresyondan daha etkili olduğunu belirtmiştir. Mina (2010) özürlü kişilerin iş seçimi ile ilgili yaptığı çalışmada, parametrik lojistik regresyon ve MARS yöntemlerini kullanmış ve sonuçları değerlendirmiştir. Samui ve Kothari (2011) depolardaki buharlaşma kayıplarının tahminini MARS ile yapmış ve sonuçları yapay sinir ağları ile kıyaslamıştır. Türkiye'de ise Tunay (2001), Türkiye'de paranın gelir dolaşım hızlarının MARS yöntemiyle tahmini çalışmasını gerçekleştirmiştir. Yerlikaya (2008) MARS üzerinde bir takım düzenlemeler yaparak, oluşturduğu yeni modeli veri madenciliği uygulamaları için kullanmıştır. Kan ve Yazıcı (2010) yakıt tüketimi için, faktöriyel deneyleri, regresyon ağaçları ve MARS yönteminin sonuçlarını karşılaştırmışlardır. Kayri (2010) internet bağımlılığı ölçeğini MARS yöntemini kullanarak analiz etmiştir. Tunay (2010) bankacılık krizlerini ve Türkiye'deki durgunlukları MARS yöntemi ile tahmin etmiştir. Topak (2011) Türkiye'de kurumsal başarısızlığı modellemek için MARS yöntemini kullanmıştır.

2.1 Mars Yöntemi ile Tahmin

Parametrik olmayan regresyon yöntemleri, Kernel tahmini, yerel polinom regresyonu veya düzleştirme uzanımları yöntemlerini kullanır. MARS yöntemi, bağımlı değişken ve bağımsız değişken kümesi arasındaki ilişkiyi düzleştirme uzanımlarını kullanarak belirleyen bir yöntemdir (Friedman, 1991).

MARS yöntemi, her bağımsız değişkenin bağımlı değişkenle olan ilişkisini incelemenin yanı sıra, bağımsız değişkenlerin birbirleri arasındaki etkileşimlerini de belirler ve bu etkileşimlerin bağımlı değişken üzerindeki etkisini ortaya koyar (Tunay, 2001). Bu nedenle gözlem sayısının ve bağımsız değişken sayısının çok olduğu durumlarda MARS yöntemi iyi sonuç vermektedir.

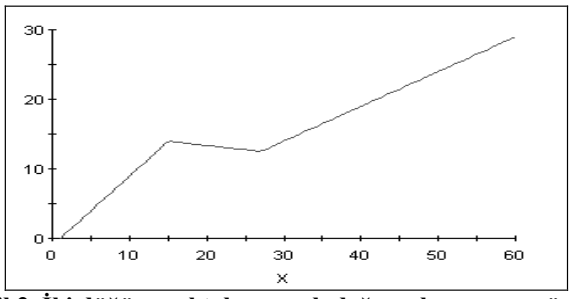
Matematikte iki tür eğri bulunmaktadır. Birinci tür interpolasyon eğrileridir. Bu eğrilerde eğri tüm veri noktalarından geçmektedir. Bu klasik eğri çizimidir. Diğeri ise düzleştirme eğrileridir. Düzleştirme eğrilerinde ise eğri, veri noktalarına yakın olmaktadır. Tam olarak bu noktalardan geçmesi gerekmez. Bu anlamda en basit düzleştirme eğrilerine parçalı doğrusal regresyon eğrisidir.



Şekil 1. Parçalı doğrusal regresyon eğrisi

Şekil 1'de sağdaki resim orijinal veriye aittir. Modelleme yapılırken veriye ait düz bir doğru oluşturmak yerine regresyon doğrusunun bükülmesi sağlanmıştır. Soldaki resimdeki doğru, üç noktadan bükülmüştür ve bir MARS uzanımı haline gelmiştir.

MARS, parçalı doğrusal regresyon uygulayarak esnek modeller oluşturur. Bağımsız değişkenin farklı aralıklarında ayrı regresyon eğim değerleri kullanarak doğrusallığı korur. Regresyon doğrusunun eğiminin değiştiği ve bir aralıktan diğerine geçildiği noktalara düğüm denir. Kullanılacak değişkenler ve her değişken için aralıkların bitiş noktası araştırma sonucu bulunur (Lee, Chiu vd., 2004).



Şekil 2. İki düğüm noktalı, parçalı doğrusal regresyon örneği

Örneğin, Şekil 2'de yer alan ve aralıkları birbirinden ayıran iki tane düğüm noktası olduğu görülmektedir. Regresyon doğrusunun eğimi bir aralıktan diğerine geçtiğinde değişmektedir.

Modeldeki değişkenler, etkileşimleri ve düğüm noktalarının konumu, kaba kuvvet yaklaşımıyla (brute force aproach), katsayılar ise en küçük kareler (EKK) yöntemiyle bulunur. Kaba kuvvet yaklaşımı, tüm olası çözümler bulunduktan sonra en iyi olana karar verilen bir yöntemdir (Dieterle, 2003).

MARS yönteminde diğer bir önemli konu, temel fonksiyonlardır. Temel fonksiyonlar, değişken aralıklarının tanımlandığı bölgesel modellerdir. Temel fonksiyonlar, tek bir eğri fonksiyonu ya da birden fazla değişkenin etkileşim terimi olabilir. Temel fonksiyonlar düğüm noktalarının belirlenmesi açısından önemlidir (Put, Massart vd., 2003). Temel fonksiyonlar bağımlı değişken (Y) ve bağımsız değişkenler $(X_1, X_2, ..., X_m)$ arasındaki ilişkiyi temsil eden fonksiyonlar olarak görev yapar. Bu fonksiyonlar, β_0 sabit parametresi ve diğer temel fonksiyonların ağırlıklandırılmış toplamından oluşur.

$$\hat{Y} = f(x) = \beta_0 + \sum_{m=1}^{M} \beta_m B_m(X) \qquad (1)$$

β_0 : sabit terim

$B_m(X)$: m. temel fonksiyon

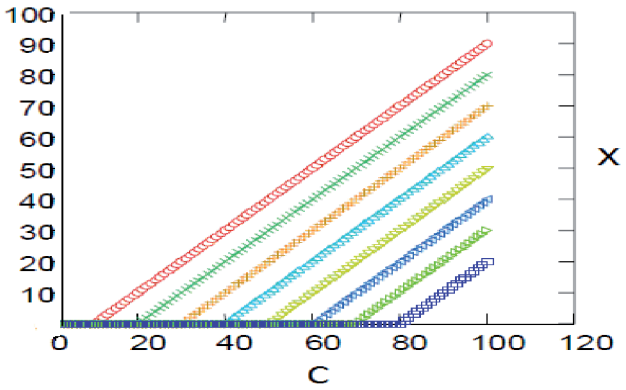
β_m : m. temel fonksiyonun katsayısı

M : Temel fonksiyon sayısı

Hokey sopası temel fonksiyonları - HSTF (hockey stick basis functions - HSBF), MARS modelinde önemli bir unsurdur. HSTF'ler sürekli X değişkeninin dönüşmüş hali olan X^* değerini,

$$X^* = \max(0, X - c), \text{veya} \quad \max(0, c - X) \tag{2}$$

biçiminde tanımlar. X^*, X'in eşik değeri olarak tanımlanan c değerinden küçük tüm değerleri için 0 değerini, c'den büyük X değerleri içinse X değerlerini alır. Örnek olarak X, 0 ve 100 arasında değerler alan bağımsız bir değişken olsun. $c = 10, 20, 30, ..., 70, 80$ için X^*'in aldığı değerler Şekil 3'de gösterilmiştir.



Şekil 3. Farklı eşik değerleri (c) için oluşturulan temel fonksiyonlar

MARS, veri setinin tüm değerleri için, değişik c değerlerine karşılık gelen çok sayıda temel fonksiyon oluşturabilir.

MARS ilk olarak sürekli bağımlı değişkenlerin tahmini için tasarlanmıştır. Daha sonra ikili kategorik bağımlı değişkenler için de kullanılmıştır (Salford Systems, 2001). Tunay (2011) çalışmasında, MARS yönteminin geleneksel regresyon yöntemleri ile parametrik olmayan modellerin üstünlüklerini başarıyla birleştiren ve ekonomik durgunluklar gibi ikili yapıdaki kategorik bağımlı değişkenlere rahatlıkla uygulanabilen yapısının önemli bir avantaj olduğunu belirtmiştir.

Klasik modellemede kategorik değişkenin her bir kategorisi için kukla değişkenler kümesi oluşturulur. Bu küme, regresyon analizinde girdiler olarak kullanılır. Bağımsız değişkenlerin kategorik olduğu durumlarda, MARS da aynı şekilde kukla değişkenler oluşturur. Ancak bu kukla değişkenler, ilgili değişkenin düzeylerinin bütünüdür.

Hastie ve Tibshirani (1990) MARS yöntemini, bağımlı değişkenin ikili olması durumunda, aşağıdaki formül ile logistik regresyon analizine uyarlamıştır.

$$\text{lojit } P(Y = 1) = f(x) + \varepsilon \tag{3}$$

Bu eşitlikte $f(x)$ MARS yöntemi tarafından tahmin edilen temel fonksiyonlar kümesini ifade eder. MARS yöntemi ile formül (3) birlikte ele alınırsa aşağıdaki formül elde edilir.

$$\text{lojit}(P_i) = \beta_0 + \sum_{m=1}^{M} \beta_m B_m(X) \tag{4}$$

Takip eden alt başlıkta MARS yöntemiyle modellemede yer alan basamaklar incelenmiştir. İyi bir model bu üç basamak sonucunda oluşur.

2.2 MARS Yönteminin Basamakları

Put, Massart vd. (2003)'e göre MARS yönteminde en iyi model üç basamaktan oluşan bir süreç sonunda elde edilir. Bu basamaklar yapım, budama ve yumuşatma aşamaları olarak adlandırılır.

2.2.1 Yapım aşaması (constructive phase)

Bu aşamada sabit terimle başlayarak ve sürekli olarak temel fonksiyonlar eklenir. Temel fonksiyonlar eklendikçe karışık ve de esnek bir model oluşur. Bu işlem kullanıcı tarafından sürecin başında belirlenen maksimum terim sayısına ulaşana kadar devam eder.

2.2.2 Budama aşaması (pruning phase)

İkinci aşama geri doğru eleme aşamasıdır. Birinci aşamada fazla sayıda temel fonksiyon eklendiği için oluşturulan model aşırı tahminleme problemiyle karşı karşıya kalacaktır. Aşırı tahminlemeyi ortadan kaldırmak için geri doğru eleme işlemine başlanır. Modelin karmaşıklığının azaltıldığı aşamadır. Modele en az katkısı olan temel fonksiyonlar atılır. Bu aşamada Craven ve Vahba (1979) tarafından geliştirilen Genelleştirilmiş Çapraz Geçerlilik (GÇG) (Generalized Cross-Validation - GCV) ölçütü kullanılır. GÇG aynı zamanda uyum eksikliği kriteridir.

$$\text{GÇG}(M) = \frac{1}{N} \frac{\sum_{i=1}^{N} \left(Y_i - \hat{f}_M(X_i)\right)^2}{\left(1 - C(M)/N\right)^2} \tag{5}$$

Formül (5)'de,
N: örneklem veri sayısı
$C(M)$: modelde uydurulan geçerli parametre sayısı
Eğer modelde M tane doğrusal bağımsız temel fonksiyon varsa,

$$C(M) = M + dc \tag{6}$$

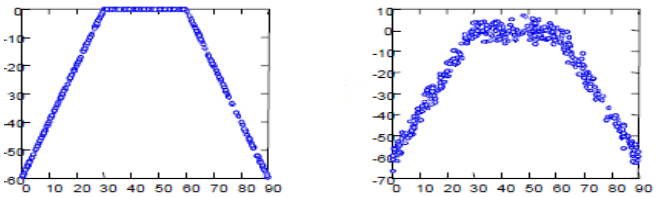
eşitliği sağlanır. Formül (6)'da, c ileri doğru olan süreçte seçilen düğüm sayısı, d ise her bir temel fonksiyon optimizasyonundaki maliyeti ifade eden bir değerdir. MARS

modellerinde genellikle d=3 alınır. Friedman (1991) tüm yapılan çalışmalar sonucunda d için en iyi değerlerin $2 \leq d \leq 4$ aralığında olduğunu belirtmiştir.

2.2.3 Yumuşatma aşaması (smoothing phase)

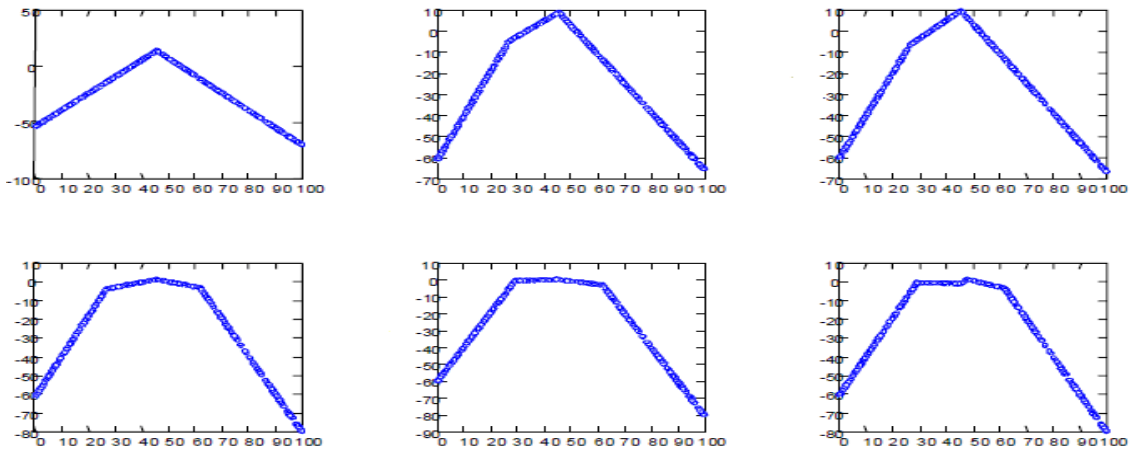
Son olarak bölgesel sınırlar içindeki süreksizliğin giderilmesi ve birincil ve ikincil türevlerin sürekliliğinin sağlanması için yumuşatma gereklidir (Quiros vd., 2009).

MARS'ın belirtilen bu basamaklarının özeti ve modellenmesi Şekil 4 ve Şekil 5'de gösterilmiştir. X ekseni bağımsız değişkeni, Y ekseni bağımlı değişkeni göstermektedir.



Şekil 4. Orijinal verinin dağılım grafiği

Şekil 5'de yer alan soldaki fonksiyon, düz tepeli fonksiyon olarak bilinir ve X=30 ile x=60 düğüm noktalarına sahiptir. Sağdaki fonksiyon ise gözlenen verinin dağılımını göstermektedir.



Şekil 5. MARS modeli ve oluşturduğu düğümler

MARS ilk olarak tek bir düğüm noktasını (X=45) belirlemiştir. İleri aşamada düğüm sayısı arttıkça MARS, Şekil 5'deki yaklaşımla modeli tahmin etmiştir (Salford Systems, 2001).

3. DÜŞÜK DOĞUM AĞIRLIĞINA İLİŞKİN MARS YÖNTEMİNE DAYALI İKİLİ LOJİSTİK REGRESYON MODELİ

Doğum ağırlığı bir bebeğin doğduğu anda ölçülen ağırlığıdır. Doğum ağırlığı yeni doğan ölümleri ve bebeklik dönemi hastalıklarını etkileyen faktörlerden biridir. Bu nedenle, doğum ağırlığı ve bunu etkileyen faktörler her zaman önemli bir klinik araştırma konusu olmuştur.

Düşük doğum ağırlığı, Dünya Sağlık Örgütü tarafından 2500 gramdan daha az olan doğum ağırlığı olarak tanımlanmıştır. Doğum ağırlığı ve gebelik haftası cenin veya yeni doğan ölümlerinin etkenlerinden biridir. Bu nedenle düşük doğum ağırlıkları veya prematüre doğumlar (37 haftadan önce gerçekleşen doğum) bu konudaki araştırmalar için önemli birer veridirler. Bebeğin düşük doğum ağırlıklı olarak doğmasında etkisi olan etkenler, Kramer (1987) tarafından aşağıdaki maddelerle verilmiştir.

- Doğum kusuru da denilen bebeğin doğumsal genetik ve yapısal anomalileri
- Annenin doğum geçmişi: önceki ölü doğum sayısı, düşük sayısı
- Annenin yüksek kan basıncı, şeker, kalp, ciğer ve böbrek hastalıkları
- Annenin alkol uyuşturucu ve sigara alışkanlığı
- Annenin geçirdiği enfeksiyonlar
- Plasentada görülen sorunlar (bebeğe kan ve besin ulaşımında sorunlara sebep olmaktadır).
- Annenin yeterince beslenememesi ve uygun kiloda olmaması
- Sosyoekonomik faktörler (az geliri olan, eğitim düzeyi düşük, 17 yaşından küçük veya 35 yaşından büyük annelerin bu konuda artan riske sahip olduğu saptanmıştır.)

Uygulama çalışması, yeni doğan bebeklerin düşük doğum ağırlığına sahip olma olasılıklarının, ikili lojistik regresyon analizi ile MARS yöntemine dayalı olarak model oluşturulmasını kapsamaktadır.

Bu çalışma sonucunda oluşturulacak model bir annenin bebeğinin doğum ağırlığının düşük olması ya da olmaması ile ilgili tahminde bulunulabilecektir. Böylece düşük doğum ağırlıklı olarak meydana gelme olasılığı bulunan bebek için önlemler alınılabilecektir.

Çalışmada Danimarka Ulusal Doğum Grubu (DUDG)'nun hamilelikte ateş ve buna bağlı ölü doğumlarla ilgili çalışmasına ait verilerin bir kısmı kullanılmıştır. DUDG, 1997 ve 2002 yılları arasında 100.000 kadın ile hamileliğin ilk 12–16 haftasında ve sonunda görüşüp gerekli verileri derlemiştir. Andersen, Vastrup vd. (2002) derlenmiş olan bu verilerden 31 Mart 1999 öncesine ait olanları kullanarak, ölü doğum bağımlı değişkenini etkileyen, bağımsız değişkenleri saptama çalışması gerçekleştirmiştir.

Çalışmamızda Andersen, Vastrup vd. (2002)'nin elde ettiği bağımsız değişkenlerden yararlanılmıştır. Ancak bu çalışmadan farklı biçimde, bağımlı değişken, bebeklerin doğum ağırlığı olarak seçilmiştir. Yeni doğan bir bebeğin ağılığı ortalama olarak 2500–4000 gr. arasındadır. Ağırlığı 2500 gramdan daha az olan bir bebek düşük doğum ağırlıklı olarak nitelendirilmektedir. Bu nedenle 2500 gramdan düşük olan bağımlı değişkenler 1 diğerleri 0 olarak kodlanmıştır. Çalışmamızda yer alan değişkenlerden annede ilk 17 haftada gözlenen yüksek ateş sayısı, annenin yaşı, önceki ölü doğum sayısı, önceki canlı doğum sayısı, gebelik haftası, sigara, alkol ve kahve kullanımı ve bebeğin doğum boyu değişkenlerinin doğum ağırlığını etkileyebileceği düşüncesiyle değerlendirmeye alınmıştır. 982 bireye ait veri kullanılmıştır. Kullanılan aday değişkenler ve kodlama işlemi Tablo 1'de verilmiştir.

Tablo 1. Uygulama verileri ve kodlama işlemi

Bağımlı değişken		Aldığı değerler ve kodları	Kısaltması
Y	Doğum ağırlığı	0=DA $\geq$ 2500 gr 1=DA<2500 gr	DA
Bağımsız değişkenler			
x_1	İlk 17 haftadaki yüksek ateş sayısı	0,1,2,.....	ateş
x_2	Annenin yaşı	0=yas<35 1=yas $\geq$ 35	yer
x_3	Geçmişteki düşük sayısı	0=düşük yapmamışsa 1=diğer d.	düşük
x_4	Önceki canlı doğum sayısı	0=canlı doğum yapmamışsa 1=diğer d.	canlı
x_5	Doğumda gebelik haftası	Sürekli değişkendir	hafta
x_6	Sigara kullanımı	0=Yok 1=Var	sigara
x_7	Alkol tüketimi	0=Yok 1=Var	alkol
x_8	Kahve tüketimi	0=Yok 1=Var	kahve
x_9	Boy	Sürekli değişkendir (cm)	boy

3.1 Verilerin MARS Yöntemi ile Analizi

Verilerin çok değişkenli uyarlayıcı regresyon uzanımları (MARS) yöntemiyle analizi aynı adı taşıyan MARS paket programı ile yapılmıştır. Bu programın MARS 6.6 sürümü, Salford Systems tarafından geliştirilen Salford Predictive Modeler Builder adlı paket program içinde yer almaktadır. Bu program CART, MARS, Combine, Randomforest, Treenet gibi değişik parametrik olmayan analiz yöntemlerini içeren paket programdır. MARS programında veriler girildikten sonra aynı verilerle fazla sayıda model oluşturulabilir. Bu modellerden en iyi olanı en az uyum eksikliğine sahip olandır. Uyum eksikliği en az olan model ise en düşük GÇG değerine sahip olandır. MARS program çıktılarında modele ait GÇG değeri verilmektedir.

Maksimum temel fonksiyon sayısı, maksimum etkileşim, düğümler arasındaki en az gözlem sayısı, düğüm optimizasyonu için serbestlik derecesi gibi değerler oluşturulan modeli belirler (Tablo 2). Bu değerler tamamen uygulayıcıya bırakılmıştır.

Tablo 2. MARS'ta model oluşumunda etkili olan değerler ve kısaltmaları

	MARS'taki ilk değeri	Kısaltması
Maksimum temel fonksiyon sayısı	15	TFmax
Maksimum etkileşim sayısı	1	Emax
Düğümler arasındaki en az gözlem sayısı	0	Omin
Düğüm optimizasyonu için serbestlik derecesi	3	SD

Model tahmininde öncelikle serbestlik derecesi 3 olarak belirlenmiştir. Farklı Tfmax, Emax ve Omin değerleri için GÇG değerleri kaydedilmiştir (Tablo 3-6).

Tablo 3. Emax=1 için GÇG değerleri

	TFmax		
Omin	15	25	50
0	0.04499	0.04506	0.04542
20	0.04499	0.04512	0.04580
30	0.04508	0.04508	0.04553
40	0.04719	0.04734	0.04802
50	0.04719	0.04734	0.04802
100	0.04912	0.04931	0.04966

Tablo 4. Emax=2 için GÇG değerleri

	TFmax		
Omin	15	25	50
0	0.04522	0.04476	0.04432
20	0.04522	0.04517	0.04527
30	0.04504	0.04521	0.04531
40	0.04735	0.04734	0.04742
50	0.04735	0.04735	0.04748
100	0.04914	0.04920	0.04938

Tablo 5. Emax=3 için GÇG değerleri

	TFmax		
Omin	15	25	50
0	0.04522	0.04476	0.04369
20	0.04522	0.04517	0.04546
30	0.04504	0.04521	0.04513
40	0.04735	0.04734	0.04749
50	0.04735	0.04734	0.04754
100	0.04914	0.04918	0.04922

Tablo 6. Emax=5 için GÇG değerleri

	TFmax		
Omin	15	25	50
0	0.04522	0.04476	0.04320
20	0.04522	0.04517	0.04530
30	0.04504	0.04506	0.04525
40	0.04735	0.04734	0.04753
50	0.04735	0.04734	0.04753
100	0.04914	0.04918	0.04922

Yapılan uygulamalar sonucunda GÇG değerinin, TFmax'ın 50 ve Omin'in 0'a yakın olduğunda en küçük değerlere sahip olduğu görülmüştür. Emax değerinin 5 olduğu durumda ise bu değer en küçük olmaktadır (Tablo 6).

Tfmax=50, Omin=0, Emax=5 olduğunda değişik SD değerleri için hesaplanan GÇG değerleri incelendiğinde SD değeri azaldıkça GÇG değerinin azaldığı ve optimum modele yaklaşıldığı görülmüştür (Tablo7).

Tablo7. Değişik SD değerleri için hesaplanan GÇG değerleri

SD	Genelleştirilmiş Çapraz Geçerlilik (GÇG) Değeri
0	0.04018
1	0.04141
2	0.04234
3	0.04320

Oluşturulan bu modellerden en küçük GÇG değerine sahip olan model uygun model olarak seçilmiş ve bu modele ait çıktılar aşağıda yorumlanmıştır:

Tfmax=50, Omin=0, Emax=5 ve SD=0 değerleri ile MARS uygulanmıştır.

MARS yönteminde sıradan en küçük kareler (EKK) yönteminde olduğu gibi temel fonksiyonlar ile bağımlı değişken arasındaki ilişkinin derecesini gösteren üç tür R-kare değeri kullanılmaktadır:

R-Kare: Bağımlı değişken ve temel fonksiyonlar arasındaki ilişkinin derecesi. Bu değer basit sıradan en küçük kareler regresyon analizi için hesaplanan R^2 ile aynı şekilde hesaplanır.

Düzeltilmiş R-Kare: Temel fonksiyonların sayısı için düzeltilmiştir.

Merkezsel olmayan R-Kare: Uyum iyiliğini test etmek için kullanılmamalıdır (Nash ve Bradford, 2001).

Modele ait R-kare değeri yaklaşık olarak 0.65 bulunmuştur. Düzeltilmiş R-kare 0.64 ve merkezsel olmayan R-kare 0.69 olarak bulunmuştur.

Uygulama sonucunda ileri ve geriye doğru işleyen yöntemle temel fonksiyonlardan 17 tanesinin model oluşumuna katkı sağladığı görülmektedir. Bu fonksiyonların katsayı tahminleri, standart hataları, t ve p değerleri görülmektedir (Tablo 8).

Tablo 8. Modele katkısı olan temel fonksiyonlar ve değerleri

Parametre	Tahmin	S.H	t-Değeri	p-Değeri
Sabit	1.00630	0.03693	27.25161	0.00000
Temel Fonksiyon 5	0.60325	0.08294	7.27300	0.00000
Temel Fonksiyon 9	-0.37637	0.04868	-7.73197	0.00000
Temel Fonksiyon 11	0.42711	0.06110	6.99032	0.00000
Temel Fonksiyon 13	-0.04797	0.01628	-2.94576	0.00330
Temel Fonksiyon 20	-0.03822	0.01477	-2.58753	0.00981
Temel Fonksiyon 21	-0.57432	0.08996	-6.38400	0.00000
Temel Fonksiyon 23	-0.61113	0.08769	-6.96930	0.00000
Temel Fonksiyon 27	-0.92835	0.13938	-6.66081	0.00000
Temel Fonksiyon 29	0.44079	0.09552	4.61449	0.00000
Temel Fonksiyon 31	0.60825	0.09022	6.74164	0.00000
Temel Fonksiyon 33	-0.14866	0.06755	-2.20068	0.02800
Temel Fonksiyon 35	0.04525	0.01951	2.31965	0.02057
Temel Fonksiyon 39	0.02204	0.00688	3.20159	0.00141
Temel Fonksiyon 41	0.04300	0.01453	2.95862	0.00317
Temel Fonksiyon 43	-0.03130	0.01698	-1.84370	0.06553
Temel Fonksiyon 45	-0.02329	0.00655	-3.55531	0.00040
Temel Fonksiyon 49	-0.08965	0.01298	-6.90723	0.00000

Model değerlendirmesi için kullanılan F istatistiği 104.85 ve p değeri 0.00000 olarak hesaplanmıştır.

Tahmin edilen model sonucunda bebeğin doğum ağırlığını, hafta değişkeninin %100 etkilediği, sırasıyla canlı, alkol, yaş ve sigara değişkenlerinin çoktan aza doğru etkisinin olduğu görülmektedir (Tablo 9). Tablodaki son sütun ilgili değişkenin yokluğunda modelin genel uyum iyiliğinde ne kadarlık azalma olacağı yer almaktadır.

Tablo 9. İlişkili değişkenlerin önemi

Değişken	Önemi	Genelleştirilmiş Çapraz Geçerlilik (GÇG) Değeri
Hafta	100.00000	0.11055
Canlı	18.24723	0.04253
Alkol	17.01658	0.04253
Yaş	16.85816	0.04253
Sigara	9.55560	0.04253

Modelin oluşumunda etkisi olan temel fonksiyonlar ve açıklamaları Tablo 10'da verilmiştir.

Tablo 10. Temel fonksiyonlar

BF5 = max(0, HAFTA - 38)
BF7 = (SIGARA in (0))
BF8 = (SIGARA in (1))
BF9 = max(0, HAFTA - 35)
BF11 = max(0, HAFTA - 36)
BF13 = max(0, HAFTA - 37) * BF8
BF15 = (CANLI in (0))
BF17 = (YAS in (0))
BF18 = (YAS in (1))
BF20 = max(0, 35 - HAFTA) * BF17
BF21 = max(0, HAFTA - 38) * BF17
BF23 = max(0, HAFTA - 37) * BF18
BF25 = (ALKOL in (1)) * BF15
BF27 = max(0, HAFTA - 36) * BF25
BF29 = max(0, HAFTA - 38) * BF25
BF31 = max(0, HAFTA - 35) * BF25
BF33 = max(0, HAFTA - 39) * BF25
BF35 = max(0, HAFTA - 40) * BF17
BF36 = max(0, 40 - HAFTA) * BF17
BF39 = (CANLI in (0)) * BF36
BF40 = (CANLI in (1)) * BF36
BF41 = (ALKOL in (1)) * BF40
BF43 = (SIGARA in (0)) * BF41
BF45 = max(0, HAFTA - 32) * BF7
BF49 = max(0, HAFTA - 32) * BF17

Bu temel fonksiyonlar ve katsayılarıyla elde edilen model ise aşağıdaki gibi olacaktır.

$$Y = 1.00629 + 0.60322 * BF5 - 0.376346 * BF9 + 0.42706 * BF11 - 0.0479715 * BF13 - 0.0382146 * BF20 - 0.574262 * BF21 - 0.611076 * BF23 - 0.928152 * BF27 + 0.440698 * BF29 + 0.608121 * BF31 - 0.148641 * BF33 + 0.0452413 * BF35 + 0.0220369 * BF39 + 0.0430005 * BF41 - 0.0313029 * BF43 - 0.023289 * BF45 - 0.0896456 * BF49;$$

Doğru sınıflama oranı ise modelin uygunluğunu ya da verilerin bağımlı değişkenini açıklama yüzdesini gösteren bir değerdir. Modele ait DSO değeri %95.418 olarak hesaplanmıştır. İlgili değişkenler bağımlı değişkenin yaklaşık %95'ini açıklamıştır (Tablo11).

Doğru sınıflama çizelgelerinden modelin duyarlılığı (sensitivity) ve özgüllüğü (specificity) hakkında da yorum yapılabilir.

Tablo 11. MARS için sınıflandırma çizelgesi

Gözlenen		Tahmin edilen		
		DA		toplam
		0.00	1.00	
Adım 1	DA 0.00	850	8	858
	1.00	37	87	124
				95.418

Duyarlılık referans düzeyin doğru tahmin edilme derecesidir. MARS için referans düzey, doğum ağırlığının 2500 gram ve üstünü ifade eden ve 0 olarak kodlanmış düzeydir. Bu durumda duyarlılık=850/858=0.9906 değerine sahiptir.

Özgüllük ise 1 olarak kodlanan doğum ağırlığının 2500 gramdan küçük olduğu değerleri ifade eden düzeyin doğru tahmin edilmesinin derecesidir. MARS yöntemiyle yapılan analiz için özgüllük, 87/124=0.7016 olarak hesaplanmıştır.

4. SONUÇ VE YORUMLAR

Regresyon analizinde amaç bağımlı değişkeni bağımsız değişkenlerin bir fonksiyonu şeklinde yazmaktır. Parametrik regresyon analizinde önceden bilinen modele ait fonksiyonun parametrelerini tahmin etmek amaçlanır. Parametrik olmayan yöntemde ise parametre yerine doğrudan fonksiyonlar tahmin edilir. Model ise bu fonksiyonların bir kombinasyonudur.

Gerçekleştirilen araştırmada verilerin MARS yöntemine dayalı ikili lojistik regresyon ile yapılan analizi sonucunda doğru sınıflandırma oranı %95.418 olarak bulunmuştur. Bu oran oldukça yüksek bir orandır. Oluşturulan modelde, gebelik haftası en fazla etkiye sahip değişken olarak ortaya çıkmıştır (Tablo 9). Aynı şekilde Tablo 10

incelendiği zaman hafta değişkeninin tek başına bir temel fonksiyon oluşturduğu gibi diğer birçok temel fonksiyonun oluşumunda da yer aldığı görülmektedir. Hafta değişkeni sürekli değişken olduğu için birçok düğüm noktası ortaya çıkmıştır. Analiz sonucunda bebeğin doğum ağırlığını, 32-40 arasındaki haftaların etkilediği temel fonksiyonlardan anlaşılmaktadır. Modelin temel fonksiyonlarının katsayıları incelendiğinde en büyük katsayılardan birisinin yine bu değişkene ait olduğu görülmektedir.

Bebeğin doğum ağırlığını etkileyen bir diğer önemli değişken ise, anne adayının daha önce canlı doğum yapıp yapmamasıdır. Daha önce canlı doğum yapmamış bir bireyin aynı zamanda alkol de kullanıyor olmasının bebeğin düşük doğum ağırlıklı olmasına sebep olduğu, 25 numaralı temel fonksiyondan anlaşılmaktadır (Tablo 10).

Yaş değişkeninin de bebeğin doğum ağırlığında etkili olduğu anlaşılmaktadır. 35 yaşına kadar olan yaşlara ait 17 numaralı temel fonksiyonun düşük doğum ağırlığına etki etmediği, 35'in üzerindeki yaşlara ait 18 numaralı temel fonksiyonda ise, yaş değişkeninin temel fonksiyon katsayısı kadar etkili olduğu anlaşılmaktadır (Tablo10).

Araştırmada kullanılan, ilk on yedi haftada ateşli hastalık geçirme ve kahve kullanma durumları çok düşük miktarda etkiye sahip oldukları için için oluşturulan modelde yer almamışlardır. Bu çalışmanın bir devamı niteliğinde, çalışmada kullanılan değişkenlere alternatif yeni değişkenlerin varlığı ve bebek doğum ağırlığına etkileri araştırılabilir.

5. KAYNAKLAR

Andersen A. M. N., Vastrup P., Wohlfahrt J., Amdersen P. K., Olsen J., Melbye M., 2002. Fever in pregnancy and risk of fetal death: a cohort study. Lancet, 360: 1552-1556.

Craven, P., Wahba, G., 1979. Smoothing Noisy Data with Spline Functions. Numer. Math, 31: 377-403.

Dieterle, F. J., 2003. Multianalyte Quantifications by Means of Integration of Artificial Neural Networks, Genetic Algorithms and Chemometrics for Time-Resolved Analytical Data. Ph.D. thesis, Universität Tübingen, Tübingen, 183 s.

Friedman, J. H., 1991. Multivariate Adaptive Regression Splines (with discussion). Ann. of Statistics, 19: 1-141.

Hastie, T. J., Tibshirani, R. J., 1990. Generalized Additive Models. Chapman & Hall/CRC, New York, 335s.

Kan, B., Yazıcı, B., 2010. Comparison of the Results of Factorial Experiments, Fractional Factorial Experiments, Regression Trees and MARS for Fuel Consumption Data. WSEAS TRANS. on MAT., 2:110-119.

Kayri, M., 2010. The Analysis of Internet Addiction Scale Using Multivariate Adaptive Regression Splines. Iranian J Publ Health, 39: 51-63.

Kim, J. H., 2000. MARS Modeling for Ordinal Categorical Response Data: A Case Study. Soongsil University.

Kolyshkina, I., Brookes, R., 2002. Data Mining Approaches to Modelling Insurance Risk. Electronic Version, Pricewaterhouse Coopers, New York, 20 s.

Kramer, M. S., 1987. Determinants of Low Birth Weight: Methodological Assessment and Meta-analysis. Bull World Health Organ. 1987;65(5):663-737.

Kriner, M., 2007. Survival Analysis with Multivariate Adaptive Regression Splines. Dr. Rer. Nat. Universitat Munchen, 101 s.

Kuhnert, P. M., Do, K. A., Mc, R., 2000. Combining Non-parametric Models with Logistic Regression: an Application to Motor Vehicle Injury Data. Computational Statistics & Data Analysis, 34: 371-386.

Lee, T. S., Chiu, C. C., Chou, Y. C., Lu, C. J., 2004. Mining The Customer Credit Using Classification And Regression Tree and Multivariate Adaptive Regression Splines. Computational Statistics & Data Analysis, Volume 50, Issue 4, 24 February 2006, Pages 1113–1130.

Mina, C., 2009. Profiling Poverty with Multivariate Adaptive Regression Splines. PIDS Discussion Paper Series No. 2009-29, 55 s.

Mina, C., 2010. Employment Choices of Persons with Disability in Metro Manila. PIDS Discussion Paper Series No. 2010-29, 35 s.

Nash, M. S., Bradford, D. F., 2001. Parametric and Nonparametric(MARS; Multivariate AdditiveRegression Splines) Logistic Regressions for Prediction of A Dichotomous Response VariableWith an Example forPresence/Absence of an Amphibian. United States Environmental Protection Agency (EPA), USA, 40s.

Put, R., Massart D. L., Heyden V., 2003. An Application of Multi-variate Adaptive Regression Splines (MARS) in QSRR, IEJMD. BioChemPress.com.

Quiros, E., Felicisimo, A. M., Cuartero, A., 2009. Testing Multivariate Adaptive Regression Splines (MARS) as a Method of Land Cover Classification of TERRA-ASTER Satellite Images. Sensors, 9:9011-9028.

Salford System, 2001. MARS, User Guide. Cal. Stat. SoftWare, Inc., San Diego, California.

Samui, P., Kothari, D.P., 2011. Application of Multivariate Adaptive Regression Splines to Evaporation Losses in Reservoirs, Earthscience, 4:15-20.

Stokes, H. H., Lattyak, W. J., 2005. Multivariate Adaptive Regression Spline (MARS) Modeling Using the B34S ProSeries Econometric System and SCA WorkBench. Scientific Computing Associates Corp., 80 s.

Topak, M. S., 2011. An Empirical Study to Model Corporate Failures In Turkey: A Model Proposal Using Multivariate Adaptive Regression Splines (MARS). Namık Kemal Üniversitesi Sos. Bilim. Metinleri, 15 s.

Tunay, K. B., 2001. Türkiye'de Paranın Gelir Dolaşım Hızlarının MARS Yöntemiyle Tahmini. ODTÜ Geliştirme Dergisi, 28: 431-454.

Tunay, K. B., 2010. Bankacılık Krizleri ve Erken Uyarı Sistemleri: Türk Bankacılık Sektörü İçin Bir Model Önerisi. BDDK Bankacılık ve Finansal Piyasalar Dergisi, 4:9-46.

Tunay, K. B., 2011. Türkiye'de Durgunlukların MARS Yöntemiyle Tahmin ve Kestirimi. Marmara Üniversitesi İİBF Dergisi, XXX:71-91.

Yerlikaya, F., 2008. A New Contribution to Nonlinear Robust Regression and Classification with MARS and Its Applications to Data Mining for Quality Control in Manufacturing. M.Sc. Thesis, Middle East Technical University, Ankara, 102s.

MODELLING THE LOW BIRTH WEIGHT OF NEW BORN BABIES WITH BINARY LOGISTIC REGRESSION BASED ON MARS METHOD

ABSTRACT

Babies with low birth weight have some health problems in later years. Therefore, it is important to estimate before the birth whether a new born baby will have a low birth weight or not. In order to obtain this estimation, logistic regression model is a suitable choice. Logistic regression analysis is a modelling technique which is used when the dependent variable is categorical. It is also easily interpreted. When the dependent variable has only two categories, the logistic regression is called binary logistic regression. Logistic regression has parametric and nonparametric solutions. MARS method is a nonparametric method which can be used as an alternative to the parametric solutions in the analysis of logistic regression. The nonparametric models require fewer assumptions compared to the parametric ones and they are also more flexible. In the application, a binary logistic regression model has been fitted to estimate whether a new born baby will have a low birth weight or not. The model has been estimated based on the MARS method. In the analysis, data belonging to 982 subjects have been investigated by applying the MARS software. In the conclusion part, the findings are interpreted.

Keywords: Low birth weight, Logistic regression, Maximum likelihood, MARS.

Volume: 10 Number: 02 Category: 02 Page: 73-82 ISSN No: 1303-6319 Cilt: 10 Sayı: 02 Kategori: 02 Sayfa: 73-82

AN APPLICATION OF COPLOT ON TURKEY DEMOGRAPHIC AND HEALTH SURVEY 2008 DATA

Yasemin KAYHAN* **Süleyman GÜNAY****

ABSTRACT

CoPlot method, an extension of multidimensional scaling, gives opportunity to investigate the relations between the observations, and between the variables on the same map. CoPlot consists of two graphs drawn on each other. First graph represents the distribution of n multivariate observations in a two dimensional space. The second graph consists of p arrows each representing a variable. By means of CoPlot, researchers can make more detailed comments about the multivariate data set with a single map. Since CoPlot is easy to understand, it has been used in various disciplines such as socioeconomic, economics and medicine but not in the demographic studies. In this study, CoPlot is briefly explained and a simple application of the method on a part of "Turkey Demographic and Health Survey, 2008" data set is presented. Superiority of CoPlot about visually interpreting the multivariate data set is emphasized by using an exceptional data set.

Keywords: Kruskal raw stress, Multidimensional scaling, Scree plot.

1. INTRODUCTION

In multivariate data analysis literature, there are various graphical representation techniques used to visualize a multivariate dataset. The main objective of these methods is to reduce multidimensional data into a lower dimension and then discover the hidden structure by means of a graphical representation of the data. One of the multivariate data analysis techniques which has been used with this purpose is multidimensional scaling (MDS). However, like many multivariate analysis methods, MDS investigates the relations between the observations, and between the variables by producing two different graphs.

MDS analysis requires only the proximities, the dissimilarities or similarities between the pairs of observations (Hastie et al., 2008). MDS tries to find the appropriate low dimensional graphical representation of the observations, so that the distances between the observations match the proximities as close as possible. In MDS analysis, similarities or dissimilarities between the observations are transformed into distances by using some specific distance functions, such as Euclidean distance, City-Block distance. Once the proximities are determined, MDS representation can be produced by using different possible optimization algorithms. The obtained graph shows that the higher the dissimilarity (similarity) measures, the larger (smaller) the corresponding distances (Borg, 2005). Although MDS can be performed either on the observations or the variables, it cannot produce a map to analyze the variables and observations, simultaneously.

CoPlot analysis, an extension of MDS, enables the simultaneous investigation of the relations between the observations, and variables with a single map. This map consists of two graphs drawn on top of each other. The first graph is obtained from MDS, and

*Yrd. Doç. Dr., Hacettepe University, Department of Statistics, Ankara, e-mail: ykayhan@hacettepe.edu.tr
**Prof. Dr., Hacettepe University, Department of Statistics, Ankara, e-mail: sgunay@hacettepe.edu.tr

represents the distribution of the *p*-dimensional observations over two dimensional space. On the second graph, the relations between variables are shown by vectors, and the location of each vector is determined by mapped observations. With this main feature, CoPlot gives researcher an opportunity to make deeper and richer interpretations of the multivariate data (Lipshitz and Raveh, 1998).

Several disciplines such as econometrics (Huang and Liao, 2012), medicine (Bravata et al., 2008), computer science (Talby et al., 1999), management science (Weber et al., 1996) require the analysis of complex multivariate data often coming from large data sets describing numerous variables for many subjects, and the multivariate nature of these data make it difficult to assess the associations of the predictors and the outcomes of interest. So, CoPlot can be seen as a valuable exploratory analysis tool in such analyses.

The objectives of this paper are simply trying to introduce the CoPlot and presenting an application of this method on a demographic data. Although CoPlot have been used previously in several disciplines, it has not been used on demographic data sets. So, this study can be considered as useful in the sense that there are not many applications of CoPlot available in the literature. Analyzing Turkey Demographic and Health Survey 2008 data set with only CoPlot map does not give enough evidence to make any conclusion in general. To get deeper understanding on the issues investigated, more works on detailed statistical analysis and inferences would be needed.

In the following section, general description and methodology of CoPlot are given. In section 3, an application of CoPlot on a part of Turkey Demographic and Health Survey 2008 data set is presented. By using an original data set, interpretive superiority of CoPlot over MDS is displayed. Some concluding remarks are given in the final part.

2. GENERAL DESCRIPTION OF COPLOT

The final product of CoPlot is a simple picture of the multidimensional dataset. With this plot, one can observe: the similarity between the observations, the correlations among the variables and the mutual relationships between the observations and the variables. The main advantage of CoPlot over a MDS is that the clusters of observations which are highly characterized by a particular variable is mapped together and located in the same direction as that of the variable's vector (Bravata et al., 2008). CoPlot analysis is performed in four steps. MDS analysis is executed in the first three steps, and in the last step the variables' vectors are placed on top of the obtained MDS graph.

Let's assume that X is an *nxp* data matrix. The rows of this matrix are assumed as the observations and will be denoted as *n* points on two dimensional space. The columns of the data matrix correspond to the variables which are exhibited by *p* vectors on CoPlot map.

Step 1: Standardization of the Data Matrix

The dissimilarities between the observations are transformed into distances by using City-Block distance in CoPlot. That is why the variables measured on different scales affect the distance values evaluated by City-Block in such a way that the largest variance variable will dominate the distance measure (Borg, 2005). So the data matrix X is transformed into matrix Z as follows,

$$Z_{ij} = \frac{\left(x_{ij} - \overline{X}_j\right)}{S_j}, \qquad i = 1,2,\ldots,n \; ; \; j = 1,2,\ldots,p \tag{1}$$

where $\overline{X}_j$ and S_j represent mean and standard deviation of the *j*-th column of X, respectively.

Step 2: Creating distance matrix

At this step, dissimilarities between the observations are transformed into distances by using a proper distance function. The distance between each pair of observations is denoted in a symmetric $D_{nxn}\left(d_{ij}\right)$ matrix. To generate D_{nxn}, 2-combination of *n*, $C(n,2)$, different d_{ij} distances are calculated as follows,

$$d_{ij} = \sum_{k=1}^{p} \left| Z_{ik} - Z_{jk} \right| > 0 \tag{2}$$

Step 3: Mapping Distances

The representation of *n* observations on two-dimensional space is generated during this step. If the rank-order of the proximities implies that the distances must have the same rank-order as proximities, we speak of non-metric MDS. In non-metric MDS, it is accepted that rank-order of the distances between the observations are informative. In this case, disparities $(\hat{d}_{ij})$ which are obtained from Euclidean distances of MDS coordinate matrix *Y* are assigned to the proximities in such a way that these values display the same rank-order as the data (Kruskal, 1964a; Borg, 2005).

The aim of non-metric MDS is to obtain a coordinate matrix *Y* such that the distance between rows *i* and *j* of *Y*, $d_{ij}(Y)$, matches disparities as closely as possible. Namely, MDS tries to minimize the following cost function,

$$\sigma^2\left(\hat{d}, Y\right) = \sum_{i=2}^{n} \sum_{j}^{i-1} w_{ij} \left(\hat{d}_{ij} - d_{ij}(Y)\right)^2 \tag{3}$$

This function is known as raw-Stress introduced by Kruskal (1964a), and the value of the function measures the quality of MDS representation. The w_{ij} is a user defined weight that should be non-negative and relates to missing information. The minimization of this function has no closed form solution, so it must be solved by iterative algorithms. Within these iterative algorithms, one of them is known as SMACOF (De Leeuw, 1977). By using this iterative majorization algorithm, optimal *Y* can be obtained.

Step 4: Adding Vectors

At this step, vectors corresponding to the variables are drawn onto the map obtained from Step 3. For each variable, CoPlot produces a vector coming up from the center of mass of the points denoted on the MDS map.

To find the direction of the vector *j*, initial angle between the *j*-th vector and the *x* axis is assumed to be zero. Then projections of all points on to the vector are calculated. The goal is to find the angle which maximizes the correlation between *n* projected scores and the *z*-score values for variable *j*. The angle which makes the correlation maximum decides the direction of the vector for variable *j*. This procedure is separately performed for all variables in the data set (Lipshitz and Raveh, 1998).

This vector representation has some advantages for the interpretation of the data. By considering the correlation values of the vectors, it may be decided that which variables should be kept in graphical representation or discarded. The vectors for highly correlated variables are located in the same direction. The vectors for highly negatively correlated variables are located along the same axis but in opposite directions. Two vectors which are orthogonal to each other imply that corresponding variables are not correlated. Obviously, the main advantage of this representation is to allow the simultaneous consideration of both variables and observations.

3. APPLICATION

For a better explanation of the procedure, some applications of the use of CoPlot are given in this section. The used data set, Turkey Demographic and Health Survey - 2008, is taken from Hacettepe University Institute of Population Studies (TDHS-2008, with the permission number 2012/8). 2,473 women who have terminated their pregnancy within the duration of their marriage are selected. Respondents (observations) are separated with respect to 12 regions. Table 1 represents the number of respondents within corresponding regions.

Table 1. Number of respondents from each region

Regions	Number of Respondents
Istanbul	187
West Marmara	134
East Marmara	199
Aegean	199
Central Anatolia	191
Mediterranean	333
West Black Sea	232
East Black Sea	124
West Anatolia	160
North East Anatolia	188
Central East Anatolia	204
South East Anatolia	322

Ten variables such as; respondent's current age (1), number of household member (2), wealth index which is an indicator of the level of wealth (3), total children ever born (4), age of respondent at 1st birth (5), age of respondent at first marriage (6), years since first marriage (7), partner's educational attainment (8), partner's age (9), number of living children (10) are selected from the survey data set. Observations are classified on the MDS and CoPlot maps with respect to the respondent's educational attainment as

follows; not educated women – square shaped points, highest educated women – cross shaped points, and the rest of the women – circle shaped points. Obtained Figure 1 and Figure 2 for the first region (Istanbul) represent distinction between MDS and CoPlot, and display what is gained in interpretability. Figure 1 shows the map produced by MDS of the ten variables describing 187 respondents using City-Block distance to calculate the dissimilarities between the observations. Figure 1(a) shows the embedding of 187 observations in a two dimensional space, and high educated women and not educated women generate two different clusters. In Figure 1(b), each variable is shown as a point and the points are arranged in such a way that their distances correspond to the correlations. That is, two points are close to each other (such as Wealth Index and Number of Household Members), if their corresponding correlation is high. However, one cannot decide the direction of this correlation. Conversely, Wealth Index and Number of Household Members are found as highly negatively correlated variables, see Figure 2. With MDS analysis, it is possible to obtain a graph that displays the relations between either the observations or the relations between variables, but seeing the observations and variables at the same graph simultaneously is not possible.

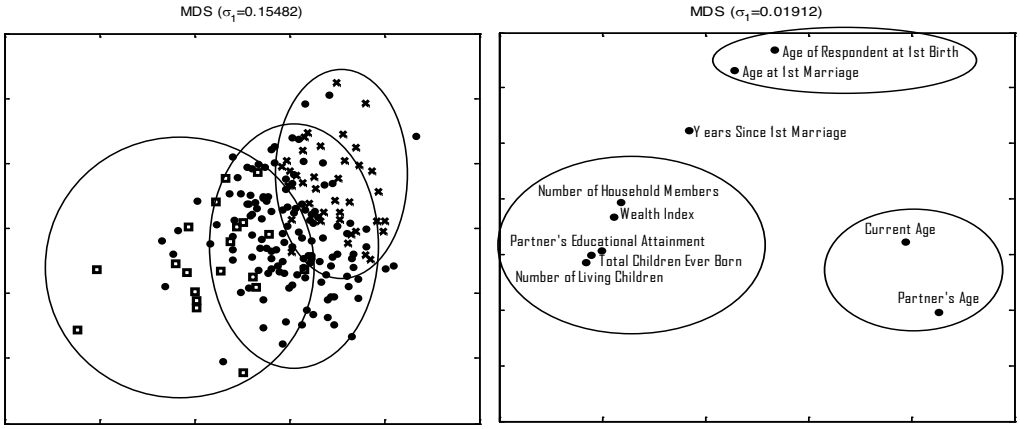


Figure 1. MDS representations of observations (a) and variables (b) for Istanbul

The map of Coplot is a simple picture of the multivariate data. From Figure 2, the following results can be concluded: Ages of the women who live in Istanbul are generally the same with the ages of their husband (1 and 9). Number of living children is highly correlated with the number of total children ever born (4 and 10). So it can be thought that child mortality is not high in this region. Age of 1-st marriage and age of 1-st birth are close to each other (5 and 6). So it can be said that women got pregnant when they got married. From the vectors 2 and 8, it is said that well educated husband do not prefer to live with big families. As the distribution of the observations and the directions of the vector are considered together, it can be said that, well educated women prefer to become mother at later age, and well educated women are richer. Not educated women have more children, and they live with crowded families. It is obvious that CoPlot map of the data set gives much more information.

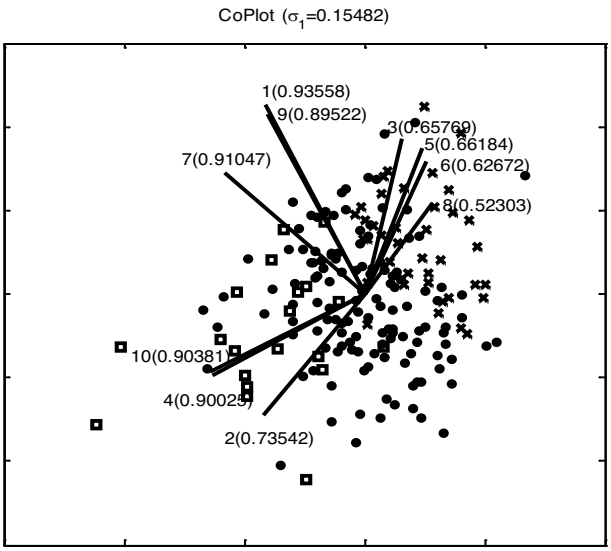


Figure 2. CoPlot representation of Istanbul

The following findings with CoPlot about rest of the regions are indicated by Figure 3 and Figure 4: For Aegean region, obtained map, therefore the comment, is nearly same as Istanbul. For Central Anatolia, variables 1, 7 and 9 are correlated. It may be concluded that not educated women got married at an early age with peers. For Central East Anatolia, not educated women rate is high relative to Istanbul, Aegean and Central Anatolia. For East Marmara, high educated women's number of living children and number of total children ever born are low (4 and 10 are in the opposite direction of cross-shaped cluster). Similar to Central East Anatolia, in Mediterranean region, the number of living children and the number of total children ever born are high for not educated women (4 and 10 are in the same direction of square-shaped cluster). For East Blacksea, high educated women get married and be mother at a late age. Similar comments can be given for the rest of the regions. We do not claim that a single map can be sufficient for these comments. These results need to be discussed sociologically and/or psychologically in details and need to be extended with more comparative statistical analysis. However, CoPlot is useful to give a researcher insight into understanding the possible relations in the data and for further analysis.

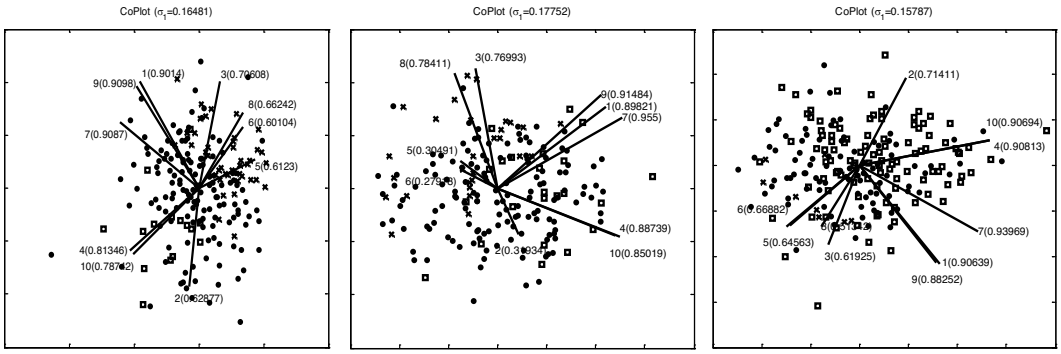


Figure 3. CoPlot maps of Aegean, Central Anatolia and Central East Anatolia

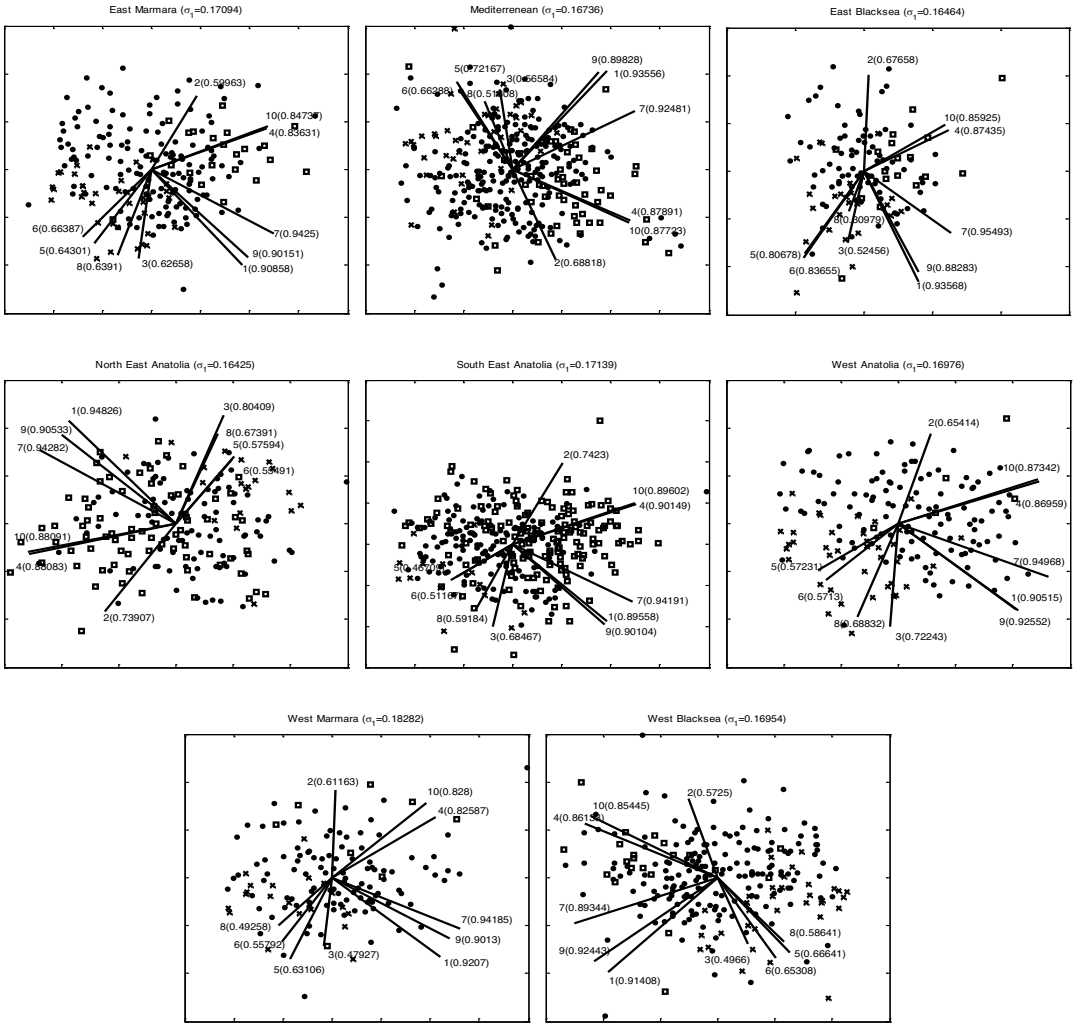


Figure 4. CoPlot maps of 8 regions

Kruskal-stress, σ_1, describes how well the map represents the observations in lower dimension. Generally MDS representations having Kruskal stress value 0.10 are accepted as fair and over 0.20 as poor (Kruskal, 1964b). To improve the goodness of fit, investigating the scree plot may be helpful. For example, from Figure 4, West Marmara has the highest σ_1 value. Figure 5 is the scree plot of this region for 30, 90 and 134 observations. For 30 observations, since σ_1 is almost 0.15, two dimensional MDS representation of the region can be accepted as fair. For 134 observations, to reach a fair representation at least 3 dimensions are needed. It is obvious that, σ_1 decreases with decreasing number of observations.

From Figure 4, it can be seen that, some of these ten variables do not have high correlation coefficient values for every region. For example, at East Black Sea, correlation coefficient values of vectors 3 and 8 are low (0.52456 and 0.30979, respectively). In CoPlot map, low correlation coefficient value, namely less than 0.70, implies that that variable is not interpretive. From Figure 6, when these two vectors are

discarded from the CoPlot analysis, it is seen that correlation coefficient values of the rest of the variables are increasing and the σ_1 is reducing. Therefore, it can be suggested that vectors with low correlation coefficients should be omitted from the analysis.

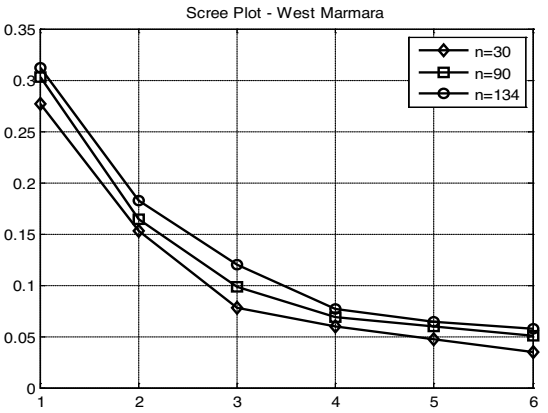


Figure 5. Scree plot for West Marmara

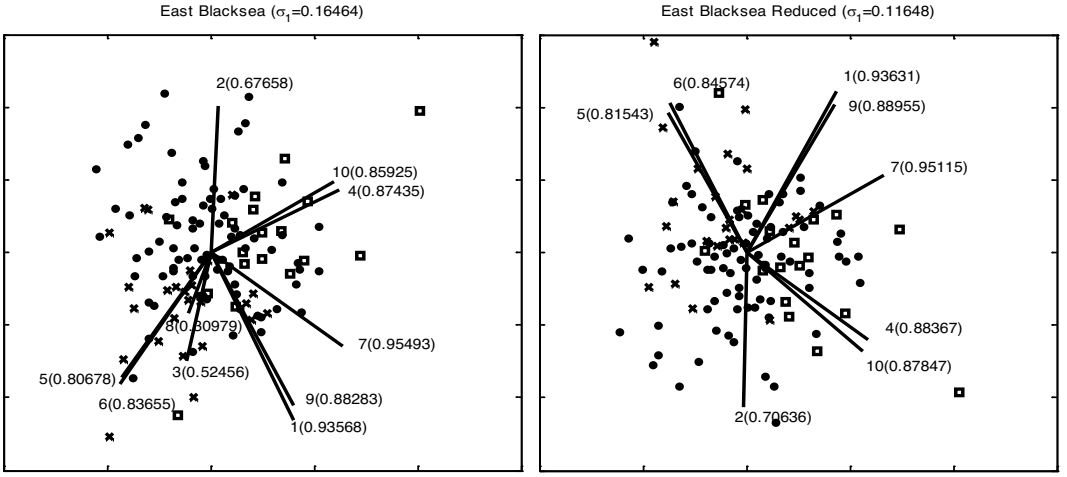


Figure 6. Effect of reducing the low correlated variables on CoPlot for East Black Sea

4. CONCLUSION

CoPlot is a graphical representation of a multivariate data set by projecting onto the two-dimensional space. It provides the possibility to analyze the observations and variables on the same graph. Previously, CoPlot has been applied to data sets from different disciplines (Raveh, 2000; Bravata et al., 2008). In this study, it has been applied to a demographic data for the first time. The purpose of this study is to show how to use CoPlot for a demographic data set. Our analysis of the demographic data suggests that CoPlot may be a useful analyzing tool for exploring the relation between observations and variables in multivariate data. At the application part of this study, it can be seen that with a simple analysis one can roughly get the picture of the regions,

and with these pictures similarities or dissimilarities between the regions would be observed. In CoPlot analysis, there are two goodness of fit measures such as Kruskall raw stress value and Pearson correlation coefficient. These measures enable the user whether to keep or delete specific variables and observations from the map. Possible problems and solutions of the problems such as low goodness of fit or correlation coefficient value are also discussed in the application part. Besides, the superiority of CoPlot on the visual interpretation of multivariate data is emphasized.

5. REFERENCES

Borg, I., Groenen, P. J. F., 2005. Modern Multidimensional Scaling, 2nd edition, Springer.

Bravata, D. M., Shojania, K. G., Olkin, I., Raveh, A., 2008. CoPlot: A Tool for Visualizing Multivariate Data in Medicine, Statistics in Medicine, 27, 2234-2247.

De Leeuw, J., 1977. Application of Convex Analysis to Multidimensional Scaling, In: J. Barra et al. Recent Developments in Statistics, 133-145.

Hastie, T., Tibshirani, R., Friedman, J., 2008. The Elements of Statistical Learning, 2nd edition.

Huang, H., Liao, W., 2012. A CoPlot-based Efficiency Measurement to Commercial Banks, Journal of Software 7:10, 2247-2251.

Kruskal, J. B., 1964a. Nonmetric Multidimensional Scaling: A Numerical Method, Psychometrika, 29:2, 115-129.

Kruskal, J. B., 1964b. Multidimensional Scaling by Optimizing Goodness of Fit to A Nonmetric Hypothesis, Psychometrika, 29:1, 1-27.

Lipshitz, G., Raveh, A., 1998. Socio-economic Differences Among Localities: A New Method of Multivariate Analysis, Regional Studies, 32:8, 747-757.

Raveh, A., 2000. CoPlot: A Graphic Display Method for Geometrical Representations of MCDM, European Journal of Operational Research, 125, 670-678.

Talby, D., Feitelson, D. G., Raveh, A., 1999. Comparing Logs and Models of Parallel Workloads Using the CoPlot Method, Lecture Notes in Computer Science 1659, 43-66.

Weber, Y., Shenkar, O., Raveh, A., 1996. National and Corporate Cultural Fit in Mergers/Acquisitions: An Exploratory Study, Management Science 42:8, 1215-1227.

TÜRKİYE NÜFUS VE SAĞLIK ARAŞTIRMASI 2008 VERİSİ ÜZERİNDE BİR COPLOT UYGULAMASI

ÖZET

Çok boyutlu ölçeklemenin bir uzantısı olan CoPlot yöntemi, gözlemler arasındaki ve değişkenler arasındaki ilişkileri aynı grafik üzerinde inceleme fırsatı verir. CoPlot, birbiri üzerine çizdirilen iki grafikten oluşur. İlk grafik n sayıda çok değişkenli gözlemin iki boyutlu uzaydaki dağılımını gösterir. İkinci grafik herbiri bir değişkeni temsil eden p sayıda oktan oluşur. CoPlot sayesinde araştırmacılar tek bir grafik ile çok değişkenli veri kümesi hakkında daha detaylı yorumlar yapabilirler. CoPlot kolay anlaşılabilir olduğu için sosyo-ekonomi, ekonomi, tıp gibi çeşitli disiplinlerde kullanılmıştır, ancak demografik çalışmalarda kullanılmamıştır. Bu çalışmada, CoPlot kısaca açıklanacak ve "Türkiye Demografik ve Sağlık Anketi 2008" veri kümesinin bir parçası üzerinde basit bir uygulaması sunulacaktır. CoPlot yönteminin çok değişkenli veri kümesini görsel olarak yorumlama üstünlüğü bu özgün veri kümesi ile vurgulanacaktır.

Anahtar Kelimeler: Kruskal raw stress, Çok boyutlu ölçekleme, Scree plot.

DANIŞMA KURULU ÜYELERİ - ADVISORY BOARD MEMBERS

Ali YAZICI	**Atılım Üniversitesi**
Alper GÜVEL	**Çukurova Üniversitesi**
Asaf Savaş AKAT	**Bilgi Üniversitesi**
Aşır GENÇ	**Selçuk Üniversitesi**
Aydın ÖZTÜRK	**Ege Üniversitesi**
Ayşe GÜNDÜZ HOŞGÖR	**Orta Doğu Teknik Üniversitesi**
Bedriye SARAÇOĞLU	**Gazi Üniversitesi**
Coşkun Can AKTAN	**Dokuz Eylül Üniversitesi**
Deniz GÖKÇE	**Bahçeşehir Üniversitesi**
Ekrem ERDEM	**Erciyes Üniversitesi**
Ercan UYGUR	**Türkiye Ekonomi Kurumu**
Erdem BAŞCI	**T.C. Merkez Bankası**
Erinç YELDAN	**Bilkent Üniversitesi**
Erol TAYMAZ	**Orta Doğu Teknik Üniversitesi**
Eser KARAKAŞ	**Bahçeşehir Üniversitesi**
Fatih ÖZATAY	**TOBB Ekonomi ve Teknoloji Üniversitesi**
Fatin SEZGİN	**Bilkent Üniversitesi**
Fikri AKDENİZ	**Çukurova Üniversitesi**
Fikri ÖZTÜRK	**Ankara Üniversitesi**
Gülay BAŞARIR KIROĞLU	**Mimar Sinan Güzel Sanatlar Üniversitesi**
Güven SAK	**TOBB Ekonomi ve Teknoloji Üniversitesi**
Haluk LEVENT	**Galatasaray Üniversitesi**
Hamza EROL	**Çukurova Üniversitesi**
İlhan TEKELİ	**Orta Doğu Teknik Üniversitesi**
İmdat KARA	**Başkent Üniversitesi**
İnsan TUNALI	**Koç Üniversitesi**
Levent KANDİLLER	**Yaşar Üniversitesi**
Mehmet KAYTAZ	**Işık Üniversitesi**
Meltem DAYIOĞLU TAYFUR	**Orta Doğu Teknik Üniversitesi**
Metin TOPRAK	**İstanbul Üniversitesi**
Mustafa ACAR	**Aksaray Üniversitesi**
Mustafa AYTAÇ	**Uludağ Üniversitesi**
Nihat BOZDAĞ	**Gazi Üniversitesi**
Onur BASKAN	**Ege Üniversitesi**
Orhan GÜVENEN	**Bilkent Üniversitesi**
Ömer Faruk ÇOLAK	**Gazi Üniversitesi**
Ömer L. GEBİZLİOĞLU	**Kadir Has Üniversitesi**
Özkan ÜNVER	**Ufuk Üniversitesi**
Öztaş AYHAN	**Orta Doğu Teknik Üniversitesi**
Savaş ALPAY	**SESRTCIC**
Seyfettin GÜRSOY	**Bahçeşehir Üniversitesi**
Süleyman GÜNAY	**Hacettepe Üniversitesi**
Turan EROL	**Ankara Strateji Enstitüsü**
Ümit OKTAY FIRAT	**Marmara Üniversitesi**
Yasin AKTAY	**Selçuk Üniversitesi**
Yılmaz AKDİ	**Ankara Üniversitesi**