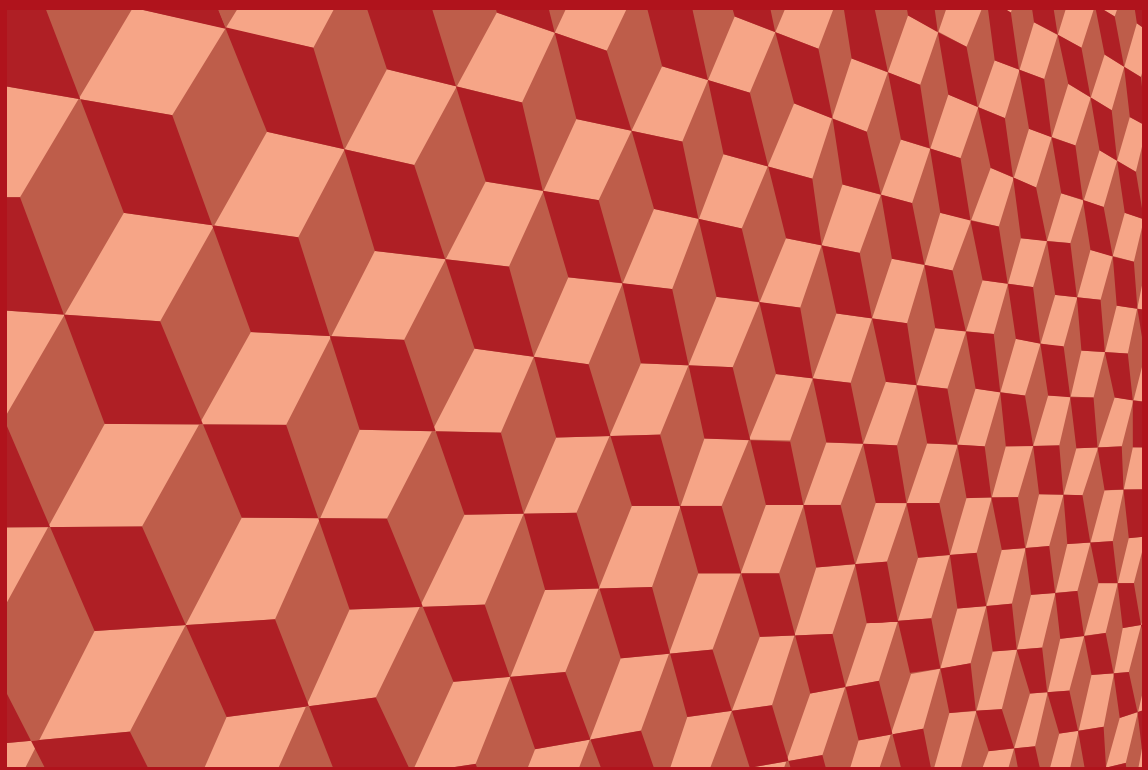
İSTATİSTİK ARAŞTIRMA DERGİSİ
Journal of Statistical Research

Cilt-Volume: 09 Sayı-Number: 01
Temmuz-July 2012

ISSN 1303-6319



TÜRKİYE İSTATİSTİK KURUMU
Turkish Statistical Institute

İSTATİSTİK ARAŞTIRMA DERGİSİ

Journal of Statistical Research

Cilt-Volume: 09 Sayı-Number: 01
Temmuz-July 2012

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: + (312) 410 03 67 - 284 45 00/171
Faks-Fax**:** + (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
3899

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: 0312 410 01 64 * Fax: 0312 418 50 82
Şubat 2013 February 2013
MTB: 2013-086 - 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 dokuz 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'e, 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

Assoc. Prof. Özlem İLK

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

İÇİNDEKİLER CONTENTS

Sayfa
Page

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 Başbakanlık 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

• 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

Prof. Dr	Hasan BAL	Gazi Üniversitesi
Prof. Dr.	Hülya ÇINGI	Hacettepe Üniversitesi
Prof. Dr.	Hüseyin TATLIDİL	Hacettepe Üniversitesi
Prof. Dr	M.Akif BAKIR	Gazi Üniversitesi
Prof. Dr.	Olcay ARSLAN	Ankara Üniversitesi
Prof. Dr.	Sadullah SAKALLIOĞLU	Çukurova Üniversitesi
Prof. Dr	Şanslı ŞENOL	Ege Üniversitesi
Doç. Dr.	Birdal ŞENOĞLU	Ankara Üniversitesi
Doç. Dr	Coşkun KUŞ	Namık Kemal Üniversitesi
Doç. Dr	Fatih TANK	Ankara Üniversitesi
Doç. Dr	Funda YURDAKUL	Gazi Üniversitesi
Doç. Dr	Mehmet YAZICI	Çankaya Üniversitesi
Doç. Dr	Mehmet YILMAZ	Ankara Üniversitesi
Doç. Dr.	Muhammed BEKÇİ	Ege Üniversitesi
Doç. Dr.	Revan ÖZKALE	Çukurova Üniversitesi
Doç. Dr.	Seda ŞENGÜL	Çukurova Üniversitesi
Yrd. Doç. Dr.	Burçak Başbuğ ERKAN	Orta Doğu Teknik Üniversitesi
Yrd. Doç. Dr.	Birol TOPÇU	Namık Kemal Üniversitesi
Yrd. Doç. Dr.	Canan HAMURKAROĞLU	Hacettepe Üniversitesi
Yrd. Doç. Dr.	Celal AYDIN	Gazi Üniversitesi
Yrd. Doç. Dr	Emel KIZILOK KARA	Kırıkkale Üniversitesi
Yrd. Doç. Dr.	Necla GÜNDÜZ TEKİN	Gazi Üniversitesi
Dr.	Nejla ÖZKAYA TURHAN	Ankara Üniversitesi
Daire Başkanı	Hasibe DEDEŞ	Türkiye İstatistik Kurumu
Uzman	Korhan BABADAĞ	Türkiye İstatistik Kurumu

Volume: 09 Number: 01 Category: 01 Page: 1-10 ISSN No: 1303-6319 Cilt: 09 Sayı: 01 Kategori: 01 Sayfa: 1-10

GECİKMESİ DAĞITILMIŞ MODELLERDE YANLI TAHMİN EDİCİLER

Selahattin KAÇIRANLAR*

ÖZET

Gecikmesi sonlu dağıtılmış modeller, aynı değişkenin gecikmeli ve gecikmesiz değerlerine sahip olduğundan sık sık yüksek ilişkili değişkenlere sahip olurlar. Bu modellere En Küçük Kareler (EKK) Yöntemi uygulandığında bazı sorunlarla karşılaşılır. Bu sorunları çözmek için de Almon ve Koyck Modeli gibi modeller önerilmiştir. Bu çalışmada, regresyon analizinde çoklu iç ilişki problemini çözmek için EKK'ya alternatif olarak tanımlanmış Ridge ve Liu tipi tahmin ediciler gibi yanlı tahmin edicilerin Almon metodu ile kombinasyonları ele alınarak alternatif metotların verilmesi amaçlanmıştır. Ayrıca tanımlanan metodlar Almon (1965) verisi kullanılarak karşılaştırılmıştır.

Anahtar Kelimeler: Almon tahmin edici, Gecikmesi sonlu dağıtılmış model, Ridge tahmin edici, Liu tahmin edici.

1. GİRİŞ

Gecikmesi sonlu dağıtılmış model,

$$y_t = \beta_0 x_t + \beta_1 x_{t-1} + \ldots\ldots + \beta_p x_{t-p} + u_t \qquad t = p+1, \ldots\ldots T, \; = \sum_{i=0}^{p} \beta_i x_{t-i} + u_t, u_t \sim IN\left(0, \sigma_u^2\right) \qquad (1)$$

şeklindedir. β_i katsayıları gecikme ağırlıkları olarak adlandırılır. (1) modeli matris formunda

$$y = X\beta + u \qquad (2)$$

şeklinde yazılabilir. Burada

$$y = \begin{bmatrix} y_{p+1} \\ y_{p+2} \\ \vdots \\ y_T \end{bmatrix}, \; \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}, \; X = \begin{bmatrix} x_{p+1} & x_p & \ldots x_1 \\ x_{p+2} & x_{p+1} & \ldots x_2 \\ \vdots & \vdots & \ddots \vdots \\ x_T & x_{T-1} & \ldots x_{T-p} \end{bmatrix}, \; u = \begin{bmatrix} u_{p+1} \\ u_{p+2} \\ \vdots \\ u_T \end{bmatrix}$$

formundadır.

(1) modelinin direk olarak bilinen EKK yöntemiyle tahmin edilmesi durumunda aşağıdaki problemlerle karşılaşılır:

a) Bağımsız değişkenler arasında çoklu iç ilişki problemi olabilir. Çünkü, aynı değişkenin p gecikmeleri modelde yer almaktadır.

*Prof. Dr., Çukurova Üniversitesi, Fen Edebiyat Fakültesi, İstatistik Böltimü, e-posta: skacir@cu.edu.tr

b) Gecikme uzunluğu p 'nin bilinmemesidir. p bilinse bile bu sayı büyük ve örneklem miktarı küçükse parametreleri tahmin edemeyebiliriz.Gecikmesi dağıtılmış modellerdeki bu tür problemleri çözmek amacıyla Koyck ve Almon modeli gibi çeşitli yöntemler önerilmiştir. Bu modellerin hemen hemen hepsi (1)'deki β 'ların davranışı hakkında bazı önbilgilerin tanımlanmasını gerektirmektedir. Genel olarak bu önbilgi stokastik ve stokastik olmayan düzeltilmiş önbilgi şeklinde sınıflandırılmaktadır (Vinod ve Ullah, 1981; Gujarati,1999).

Irving Fisher (1937) ilk olarak stokastik olmayan düzeltilmiş ön bilgiyi

$$\begin{aligned} \beta_i &= (p+1-i)\alpha \qquad 0 \le i \le p \\ &= 0 \qquad\qquad\quad i > p \end{aligned} \tag{3}$$

formunda vermiştir. Burada α bilinmeyen herhangi bir parametredir. (1)'de (3)'ün yerleştirilmesiyle,

$$\begin{aligned} y_t &= \left[\sum_{i=0}^{p}(p+1-i)x_{t-i}\right]\alpha + u_t \\ &= z_t\alpha + u_t \end{aligned} \tag{4}$$

elde edilir. Böylece α 'nın EKK tahmini (4) nolu modelden bulunabilir ve (3)'ün kullanılmasıyla β_i 'lerin tahmini bulunabilir.

β_i üzerinde lineer stokastik olmayan önbilginin genelleştirilmesi,

$$\beta_i = \alpha_0 + \alpha_1 i + \alpha_2 i^2 + \ldots + \alpha_r i^r \qquad p \ge r \ge 0 \tag{5}$$

şeklinde r-inci dereceden bir polinom olarak yazılabilir. β_i gecikme ağırlıklarının bu yapısı Almon (1965) tarafından verilmiştir. Bu yüzden Almon gecikme polinomu olarak bilinmektedir. Tekrar (5)'in (1)'de yerleştirilmesiyle α 'ların tahminlerini ve (5)'in kullanılmasıyla β_i 'lerin tahminlerini bulabiliriz. (5) matris formunda,

$$\beta = A\alpha \tag{6}$$

şeklinde yazılabilir. Burada $A:(p+1)\times(r+1)$ tipinde bir matris ve $\alpha:(r+1)\times 1$ tipinde bir vektör olmak üzere;

$$A = \begin{bmatrix} 1 & 0 & 0 & \ldots 0 \\ 1 & 1 & 1 & \ldots 1 \\ \vdots & \vdots & \vdots & \ddots \vdots \\ 1 & p & p^2 & \cdots p^r \end{bmatrix}, \qquad \alpha = \begin{bmatrix} \alpha_0 \\ \vdots \\ \alpha_r \end{bmatrix}$$

formundadır. X ve A matrislerinin ranklarının $(p+1)<(T-p)$ ve $(r+1)<(p+1)$ olduğu varsayılmaktadır. Eğer $r<p$ ise A ' nın rankı $(r+1)$ olur.

(6) ile verilen β üzerinde stokastik olmayan önbilgi altında, (2)'deki β 'yı Almon metodu ile tahmin edebiliriz. Bunu yapmanın iki yolu vardır. (6), (2)'de yerine yazılırsa,

$$\begin{aligned} y &= XA\alpha + u \\ &= Z\alpha + u \end{aligned} \tag{7}$$

elde edilir. α 'nın EKK tahmin edicisi,

$$\hat{\alpha} = (Z'Z)^{-1} Z'y \qquad (8)$$
$$= (A'X'XA)^{-1} A'X'y$$

dir. β 'nın Almon tahmin edicisi bu durumda,

$$\hat{\beta}_A = A\hat{\alpha} \qquad (9)$$

olur. Eğer $\beta = A\alpha$ doğru ise,

$E(\hat{\beta}_A) = \beta$ ve $Var(\hat{\beta}_A) = A\,Var(\hat{\alpha})A' = \sigma_u^2 A(A'X'XA)^{-1}A' = \sigma_u^2 A(A'SA)^{-1}A'$

dır ve burada $S = X'X$ 'dır. Yani $\hat{\beta}_A$ en iyi lineer yansız tahmin edicidir (BLUE).

$\hat{\beta}_A$'yı elde etmenin diğer bir yolu ise Kısıtlı En Küçük Kareler (RLS) yöntemi uygulamaktadır. A 'nın kolonları lineer bağımsız olduğundan,

$$M = I - A(A'A)^{-1}A' \qquad (10)$$

şeklinde idempotent bir matris tanımlanabilir. Burada I, $(p+1)\times(p+1)$ tipinde bir birim matristir. Bu durumda $\beta = A\alpha$ olması

$$M\beta = 0 \qquad (11)$$

olmasını gerektirir. (2) ile (11)'in birlikte çözülmesiyle RLS tahmin edicisi

$$\hat{\beta}_R = b - S^{-1}M'\left[MS^{-1}M'\right]^{-} Mb \qquad (12)$$

şeklinde elde edilir. Burada b, (2) modelinden elde edilen $b = S^{-1}X'y$ şeklinde EKK tahmin edicisidir ve "$^-$" genelleştirilmiş tersi gösterir. M matrisi, Terasvirta (1976)'nın kullanılmasıyla,

$$M = I - A(A'A)^{-1}A' = R'(RR')^{-1}R = M^2 \qquad (13)$$

şeklinde yazılabilir. Burada $R:(p-r)\times(p+1)$ tipinde tam satır ranklı aşağıdaki gibi bilinen bir matristir.

$$R = \begin{bmatrix} (-1)^0\binom{r+1}{0} & (-1)^1\binom{r+1}{1} & \cdots & (-1)^{r+1}\binom{r+1}{r+1} & 0 & \cdots & 0 \\ 0 & (-1)^0\binom{r+1}{0} & \cdots & (-1)^r\binom{r+1}{r} & (-1)^{r+1}\binom{r+1}{r+1} & \cdots & 0 \\ \vdots & \vdots & & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \cdots & \cdots & & (-1)^{r+1}\binom{r+1}{r+1} \end{bmatrix}$$

$M\beta = 0$ olması aynı zamanda,

$$R\beta = 0 = RA\alpha \qquad (14)$$

olmasını gerektirir. (2) ile (14) birlikte çözülürse RLS tahmin edicisi

$$\hat{\beta}_R = b - S^{-1}R'\left[RS^{-1}R'\right]^{-1}Rb \tag{15}$$

şeklinde tahmin edilir. Bu ifade, (9) ve (12) ile özdeştir. $R\beta = 0$ doğru değilse $\hat{\beta}_R$'de yanlı olacaktır. Ayrıca Vinod ve Ullah (1981) tarafından $\hat{\beta}_R$ ve b tahmin edicilerinin hata kareleri ortalaması (MSE) matrisleri karşılaştırılarak

$$MSE\left(\hat{\beta}_R\right) \leq MSE(b) \Leftrightarrow \beta'\left(RS^{-1}R'\right)^{-1}\beta \leq \sigma_u^2$$

elde edilmiştir. Burada MSE hata kareler ortalaması matrisini göstermektedir.

Almon tahmin edicinin bazı dezavantajlarından dolayı karşılaşılan problemleri gidermek için Hoerl ve Kennard (1970)'ın ridge tahmin edici yaklaşımı alternatif olarak ele alınmıştır (Maddala, 1974; Vinod ve Ullah, 1981; Yeo ve Trivedi, 1989; Chanda ve Maddala, 1984).

(7) modelinden α'nın Ridge tahmin edicisi

$$\begin{aligned}\hat{\alpha}_k &= \left(Z'Z + kI\right)^{-1}Z'y \\ &= \left(A'SA + kI\right)^{-1}A'X'y\end{aligned}$$

ve

$$\hat{\beta}_k = A\hat{\alpha}_k \tag{16}$$

şeklinde elde edilir. Burada I, $(r+1)\times(r+1)$ tipinde bir birim matristir. Ayrıca,

$$MSE\left(\hat{\alpha}_k\right) \leq MSE\left(\hat{\alpha}\right) \Leftrightarrow \alpha'\left[(A'SA)'+2k^{-1}I\right]^{-1}\alpha \leq \sigma_u^2$$

olduğu Vinod ve Ullah (1981)'de gösterilmiştir. Burada,

$$\begin{aligned}V\left(\hat{\alpha}_k\right) &= \sigma_u{}^2\left(Z'Z + kI\right)^{-1}Z'Z\left(Z'Z + kI\right)^{-1} = \sigma_u{}^2 G_k Z'ZG_k \\ Bias\left(\hat{\alpha}_k\right) &= -k\left(Z'Z + kI\right)^{-1}\alpha = -kG_k\alpha\end{aligned}$$

olmak üzere

$$MSE\left(\hat{\alpha}_k\right) = G_k\left(\sigma_u{}^2 Z'Z + k^2\alpha\alpha'\right)G_k \;,\; MSE\left(\hat{\alpha}\right) = \sigma_u{}^2\left(Z'Z\right)^{-1}$$

şeklindedir.

Ancak, ridge tahmin edicisi ve Lindly ve Smith (1972) tarafından verilen genişletilmiş şekli gecikmesi dağıtılmış modeller için çok umut verici değildir (Maddala, 1974). Bu nedenle alternatif tahmin yöntemlerine ihtiyaç duyulmaktadır.

2. ALTERNATİF METODLAR

2.1. Almon İle Kısıtlı Ridge'in Kombinasyonu

Gross (2003)'de (2) modeli ile $R\beta = r$ kısıtlamasını birleştirerek

$$\hat{\beta}_r(k) = \hat{\beta}(k,\beta_0) - S(k)^{-1}R'[R\,S(k)^{-1}R']^{-1}\left(R\hat{\beta}(k,\beta_0) - r\right) \qquad k \geq 0 \qquad (17)$$

şeklinde yeni bir kısıtlı ridge tahmin edici tanımlamıştır.

Burada $\hat{\beta}(k,\beta_0) = S(k)^{-1}\left(X'y + kR'\left(RR'\right)^{-1}r\right)$, $S(k) = X'X + kI = S + kI$ ve $\beta_0 = R'\left(RR'\right)^{-1}r$' dir. Özkale ve Kaçıranlar (2007) bu tahmin edicinin alternatif olarak

$$\hat{\beta}_r(k) = \hat{\beta}(k) - S(k)^{-1}R'[R\,S(k)^{-1}R']^{-1}\left(R\hat{\beta}(k) - r\right) \qquad k \geq 0 \qquad (18)$$

şeklinde yazılabileceğini göstermişlerdir. Burada $\hat{\beta}(k) = S(k)^{-1}X'y$ olup bilinen ridge tahmin edicidir. (18), (12) ve (15) de verilen RLS ile benzer yapıdadır.

Şimdi Almon ile Kısıtlı Ridge' in kombinasyonu yardımıyla aşağıdaki tahmin ediciyi tanımlayabiliriz. (2), (11) ve (18)' in birleştirilmesiyle

$$\hat{\beta}_R(k) = \hat{\beta}_k - S_k{}^{-1}M'[M\,S_k{}^{-1}M']^{-}M\hat{\beta}_k \qquad k \geq 0 \qquad (19)$$

elde edilir. Burada $\hat{\beta}_k = A\hat{\alpha}_k$'dır. Benzer şekilde (2), (14) ve (18)'in göz önüne alınmasıyla Kısıtlı Ridge tahmin edicisi

$$\hat{\beta}_R(k) = \hat{\beta}_k - S_k{}^{-1}R'[R\,S_k{}^{-1}R']^{-1}R\hat{\beta}_k \qquad k \geq 0 \qquad (20)$$

şeklinde de verilebilir. Burada $S_k^{-1} = \left(A'SA + kI\right)^{-1}$'dir. (19) ve (20) de elde edilen tahmin edicilerin (16) ile özdeş olduğu görülür.

2.2. Almon ile Liu Tahmin Edicinin Kombinasyonu

Çoklu iç ilişki problemini gidermek için daha önce ele aldığımız ridge tahmin edici pratikte yaygın kullanıma sahiptir fakat k'yı seçmek için bazı popüler metotları kullanırken karmaşık denklemlerle yüz yüze kalırız. Bu problemi gidermek amacıyla Liu (1993), Ridge ve Stein tipi tahmin edicileri birleştirerek (2) formundaki bir model için

$$\begin{aligned}\hat{\beta}_d &= \left(X'X + I\right)^{-1}\left(X'y + db\right)\\ &= \left(X'X + I\right)^{-1}\left(X'X + dI\right)b, \qquad 0 < d < 1\end{aligned} \qquad (21)$$

şeklinde bir tahmin edici tanımlamıştır. Bu tahmin edici Akdeniz ve Kaçıranlar (1995)'de Liu tahmin edici olarak adlandırılmıştır. $\hat{\beta}_d$'nin Ridge tahmin edici üzerine avantajı d'nin bir lineer fonksiyonu olması ve bu nedenle d'nin seçiminin daha kolay olmasıdır.

Liu (1993)'de $mse(\hat{\beta}_d) \leq mse(b)$ olacak şekilde her zaman bir $0 < d < 1$ olduğu gösterilmiştir. Burada mse hata kareler ortalaması skalerini göstermektedir.

(7) numaralı model için Liu tahmin edicisi

$$\begin{aligned}\hat{\alpha}_d &= (Z'Z+I)^{-1}(Z'y + d\hat{\alpha}) \\ &= (A'SA+I)^{-1}(A'X'y + d\hat{\alpha})\end{aligned} \tag{22}$$

şeklindedir. Bu tahmin edici

$$\begin{aligned}\hat{\alpha}_d &= (Z'Z+I)^{-1}(Z'Z + dI)\hat{\alpha} \\ &= (A'SA+I)^{-1}(A'SA + dI)\hat{\alpha} \\ &= G_d\hat{\alpha}\end{aligned} \tag{23}$$

şeklinde de verilebilir. β nın tahmin edicisi de $\hat{\beta}_d = A\hat{\alpha}_d$ olur. (23)'den $\hat{\alpha}_d$ tahmin edicisi için

$$\begin{aligned}MSE(\hat{\alpha}_d) &= V(\hat{\alpha}_d) + [Bias(\hat{\alpha}_d)][Bias(\hat{\alpha}_d)]' \\ &= \sigma_u^2 G_d (A'SA)^{-1} G_d' + \left((1-d)^2 (A'SA+I)^{-1}\alpha\alpha'(A'SA+I)^{-1}\right)\end{aligned}$$

elde edilir. $MSE(\hat{\alpha}) = \sigma_u^2 (A'SA)^{-1}$ olduğunu biliyoruz.

Böylece,

$$MSE(\hat{\alpha}_d) - MSE(\hat{\alpha}) = \sigma_u^2 \left[(A'SA)^{-1} - G_d (A'SA)^{-1} G_d'\right] - (1-d)^2 (A'SA+I)^{-1}\alpha\alpha'(A'SA+I)^{-1}$$

dir.

Farebrother (1976), Sakallıoğlu ve ark. (1996)'nın kullanılmasıyla $MSE(\hat{\alpha}) - MSE(\hat{\alpha}_d)$ farkının pd olması için gerek ve yeter koşul aşağıdaki teorem ile verilebilir.

Teorem 2.1 $MSE(\hat{\alpha}) - MSE(\hat{\alpha}_d)$ farkının pd olması için gerek ve yeter koşul

$$\alpha'\left[I + \frac{1+d}{2}\Lambda^{-1}\right]^{-1}\alpha < \frac{2\sigma_u^2}{1-d}$$

olmasıdır.

İspat. $A'SA$ simetrik bir matris olduğundan $T'A'SAT = \Lambda$ olacak şekilde ortogonal bir T matrisi vardır. Burada, $T'T = TT' = I$, $\Lambda = diag(\lambda_1, ..., \lambda_{r+1})$ $A'SA$ nın öz değerlerinden oluşan bir matris ve T, $A'SA$ nın öz vektörlerinden oluşan bir matristir. Bu ayrışım altında,

$$\begin{aligned}MSE(\hat{\alpha}) - MSE(\hat{\alpha}_d) &= \sigma_u^2\left[\Lambda^{-1} - (\Lambda+I)^{-1}(\Lambda+dI)\Lambda^{-1}(\Lambda+dI)(\Lambda+I)^{-1}\right] - (1-d)^2(\Lambda+I)^{-1}\alpha\alpha'(\Lambda+I)^{-1} \\ &= (\Lambda+I)^{-1} W (\Lambda+I)^{-1}\end{aligned}$$

olur. Burada, $W=(1-d)\left\{\left[2\sigma_u^2 I+(1+d)\sigma_u^2\Lambda^{-1}\right]-(1-d)\alpha\alpha'\right\}$, $c=\dfrac{2\sigma_u^2}{1-d}$ ve $B=I+\dfrac{1+d}{2}\Lambda^{-1}$ olmak üzere, $W=(1-d)^2\left[cB-\alpha\alpha'\right]$ şeklinde yazılabilir. Farebrother (1976) nın kullanılmasıyla ispat tamamlanır.

$(1+d)\sigma_u^2\Lambda^{-1}$ pd bir matris olduğundan $MSE(\hat{\alpha}\ )-MSE(\hat{\alpha}_d)$ farkı için yeterli koşul aşağıdaki gibi verilebilir**.**

Teorem 2.2 $d>1-\dfrac{2\sigma_u^2}{\alpha'\alpha}$ ise $MSE(\hat{\alpha}\ )-MSE(\hat{\alpha}_d)$ pd dir.

İspat. $MSE(\hat{\alpha}\ )-MSE(\hat{\alpha}_d)$ farkının pd olması için $W>0$ olmalıdır. $(1+d)\sigma_u^2\Lambda^{-1}$ pd bir matris olduğundan, $2\sigma_u^2 I-(1-d)\alpha\alpha'>0$ olması yeterlidir. Farebrother (1976)'nın kullanılmasıyla ispat tamamlanır.

3. UYGULAMA

Almon (1965) den alınan, 1953-1967 yıllarına ait üçer aylık veriler kullanılarak, bağımsız değişkenin kaynaklar ve bağımlı değişkenin sermaye harcamaları olduğu verinin göz önüne alınmasıyla aşağıdaki sonuçlar elde edilmiştir: Öncelikle "Schwartz Bilgi Kriteri (SC)"nin kullanılmasıyla en küçük SC değeri 12.75, $p=8$ için elde edilmiştir. (5) deki β_i üzerindeki önbilginin beşinci dereceden bir polinom ($r=5$) olması varsayımıyla başlanarak, katsayıların anlamlılık testleri yapılarak optimal polinomun derecesi 2 olarak ($r=2$) elde edilmiştir. Buna göre β parametrelerinin Almon tahminleri Tablo 3.2 de verilmiştir. $\hat{\alpha}$ nın elde edildiği (8) deki Z matrisi için koşul sayısına bakıldığında 63.5 elde edilmiştir. Bu da Z nin kolonları arasındaki yüksek bağımlılığı işaret etmektedir. Bu nedenle ridge tahmin edici yaklaşımı alternatif olarak ele alınmıştır ve (16) yardımıyla farklı k değerleri için ridge tahmin ediciler elde edilmiştir.

Tablo 3.1 Almon verisi için bazı tahminler

Gecikme	Almon(k=0)	k=0.001	k=0.002	k=0.003	k=0.2
0	0.096	0.115	0.118	0.120	0.056
1	0.123	0.127	0.128	0.127	0.065
2	0.140	0.134	0.132	0.130	0.074
3	0.146	0.134	0.131	0.129	0.086
4	0.142	0.129	0.125	0.123	0.098
5	0.127	0.117	0.114	0.113	0.113
6	0.102	0.099	0.098	0.098	0.128
7	0.067	0.074	0.077	0.080	0.146
8	0.021	0.044	0.052	0.056	0.164
Toplam	**0.963**	**0.972**	**0.974**	**0.975**	**0.930**
Koşul sayısı	**63.5**	**41.227**	**32.817**	**28.078**	**3.952**
mse	22.8594	4.267	1.8583	1.1037	0.0233

Burada toplam, bağımsız değişkenin bağımlı değişken üzerindeki uzun dönem etkisini göstermektedir. k =0.2 ridge izi yardımıyla bulunan k değeridir. k =0.2 için bulunan

katsayıların artan bir trende sahip olduğu görülür. $k = 0.003$ için koşul sayısının makul seviyeye indiği ve katsayıların beklentilerle daha uyumlu olduğu görülür. Ayrıca, Almon ile Kısıtlı Ridge'in kombinasyonu yardımıyla tanımladığımız (19) ve (20) deki yeni tahmin edicilerin (16) ile aynı sonucu verdiği görülmüştür.

Almon ile Liu' nun kombinasyonu yardımıyla (22)'de tanımladığımız tahmin edicinin kullanılmasıyla aşağıdaki sonuçlar elde edilmiştir.

Tablo 3.2 Almon verisi için Liu tahmin edici ile bazı tahminler

Gecikme	d=0.1	d=0.3	d=0.4	d=0.5	d=0.6	d=0.7	d=0.8	d=0.9	Almon (d=1)
0	0.039	0.052	0.058	0.064	0.071	0.077	0.083	0.090	0.096
1	0.048	0.065	0.073	0.082	0.090	0.098	0.106	0.115	0.123
2	0.059	0.077	0.086	0.095	0.104	0.113	0.122	0.131	0.140
3	0.070	0.087	0.095	0.104	0.112	0.121	0.129	0.137	0.146
4	0.082	0.095	0.102	0.109	0.115	0.122	0.128	0.135	0.142
5	0.095	0.102	0.106	0.109	0.113	0.116	0.120	0.123	0.127
6	0.109	0.107	0.107	0.106	0.105	0.104	0.104	0.103	0.102
7	0.124	0.111	0.105	0.098	0.092	0.086	0.080	0.073	0.067
8	0.139	0.113	0.100	0.087	0.074	0.061	0.048	0.034	0.021
Toplam	**0.765**	**0.809**	**0.831**	**0.853**	**0.875**	**0.897**	**0.919**	**0.941**	**0.963**
mse	**2.306**	**3.325**	**4.595**	**6.372**	**8.656**	**11.446**	**14.744**	**18.548**	**22.859**

Koşul sayısının en küçük olduğu $k = 0.2$ için elde edilen katsayıların artan bir trende sahip olması sonucu, tablodaki $d \leq 0.3$ değerleri için elde edilmektedir. Teorem 2.2 de verdiğimiz koşula göre bulunan d negatif olduğundan, seçilen her $0 < d < 1$ için teoremde belirttiğimiz gibi Liu tahmin edicinin *mse* skalerinin, Almon tahmin edicinin *mse* skalerinden küçük olduğu görülmüştür. Ayrıca, $0.6 \leq d \leq 0.8$ için katsayı tahminlerinin beklentilerle daha uyumlu olduğu görülmektedir.

4. SONUÇ VE ÖNERİLER

Gecikmesi sonlu dağıtılmış modellerde çoklu iç ilişki olması durumunda Almon ve Ridge tahmin ediciye alternatif olarak, Almon ile Liu' nun kombinasyonu yardımıyla yeni bir tahmin edici tanımlanmıştır. Çoklu iç ilişki problemini gidermesi ve ridge'deki k nın seçimi problemlerinin bu tahmin edici için bulunmaması, yani d'nin seçiminin daha basit olması yeni tahmin edici için bir avantaj sağlamaktadır.

Bu çalışma Çukurova Üniversitesi Bilimsel Araştırma Projeleri birimi tarafından desteklenmiştir (FEF 2009BAP25). Ayrıca, çalışmanın nümerik örnek kısmındaki yardımlarından dolayı Öğr. Gör. Dr. Hüseyin Güler'e teşekkür ederim.

5. KAYNAKLAR

Akdeniz, F. and Kaçıranlar, S., 1995. On the Almost Unbiased Generalized Liu Estimator and Unbiased Estimation of the Bias and mse. Communications in Statistics-Theory and Methods, 24 (7), 1789–1797.

Almon, S., 1965. The Distributed Lag between Capital Appropriations and Expenditures, Econometrica, Vol 30, 96-178.

Chanda, A. K., and Maddala, G. S., 1984. Ridge Estimators for Distributed Lag Models. Communications in Statistics - Theory and Methods, Vol 13, Issue 2, 217-225.

Farebrother, R. W., 1976. Further Results on the Mean Square Error of Ridge Regression, Journal of the Royal Statistical Society, Series B (Methodological) 38(3): 248–250.

Fisher, I., 1937. Income in Theory and İncome Taxation Practice, Econometrica, 5, 1-55.

Groß, J., 2003. Restricted Ridge Estimation, Stat. Prob. Lett. 65, 57-64.

Gujarati, D. N., 1999. Temel Ekonometri. (Çev.Ü. Şenesen, G. G. Şenesen). Literatür Yayıncılık.

Hoerl, A. E., and Kennard, R. W., 1970. Ridge Regression: Biased Estimation for Non - Orthogonal Problems. Technometrics 12, 69-82.

Lindly, D. V., and Smith, A. F. M., 1972. Bayes Estimates for the Linear Model, Journal of the Royal Statistical Society, B Series, 1-41.

Liu, K., 1993. A New Class of Biased Estimate in Linear Regression, Commun. in Statistics - Theory and Methods , 22(2), 393-402.

Maddala, G. S., 1974. Ridge Estimator for Distributed Lag Models, NBER Working Paper Series, no:69.

Özkale, M. R., and Kaçıranlar, S., 2007. The Restricted and Unrestricted Two Parameter Estimators, Communications in Statistics Theory and Methods, Vol 36(15), 2707-2725.

Sakallıoğlu, S., Kaçıranlar, S., and Akdeniz, F., 1996. A Note on Combining Ridge and Least Squares Estimator , Journal of Institute of Mathematics and Computer Science (Math. Series), Vol. 9, No. 2, 193-198.

Teravista, T., 1976. A Note on Bias in Almon Distributed Lag Estimators. Econometrica, 1317-1322.

Vinod, H. D., and Ullah, A., 1981. Recent Advances in Regression Methods, Marcel Dekker, New York.

Yeo, S. J., and Trivedi, P. K., 1989. On Using Ridge - Type Estimators for a Distributed Lag Model, Oxford Bulletin of Economics and Statistics, Vol 51, Issue 1, 85-90.

BIASED ESTIMATORS FOR DISTRIBUTED LAG MODELS

ABSTRACT

The finite distributed lag models often include highly correlated variables since they have lagged and unlagged values of the same variable. Some problems are faced when the ordinary least squares (OLS) method is applied to these models. Models such as Koyck and Almon models, have been suggested to tackle these problems. In this study, providing alternative methods are aimed by introducing the combinations of Almon method with biased estimators such as Ridge and the Liu type estimators, which are alternatives to OLS defined for solving multicollinearity problem. Morover, these defined methods are compared by using Almon(1965) data.

Keywords: Almon estimator, Finite distributed lag model, Ridge estimator, Liu estimator.

Volume: 09 Number: 01 Category: 01 Page: 11-20 ISSN No: 1303-6319 Cilt: 09 Sayı: 01 Kategori: 01 Sayfa: 11-20

GAMMA, WEİBULL VE LOG-NORMAL DAĞILIMLARININ DOĞRU SEÇİM OLASILIKLARINA GÖRE AYRIŞTIRILMASI

Hayrinisa DEMİRCİ BİÇER* **Cemal ATAKAN****

ÖZET

Gamma, Weibull ve Log-Normal dağılımları, çarpık verilerin analizi için sık kullanılan dağılımlardır. Bu çalışmada, verilen bir veri setinin, Gamma, Weibull ya da Log-Normal dağılımlardan hangisi ile modelleneceği problemi üzerinde durulmuştur. Verilen bir veri setinin dağılımının Gamma, Weibull ya da Log-Normal dağılımından hangisinden geldiğine, veri setinin dağılımının sırasıyla ilgili dağılımlara göre oluşturulan yokluk hipotezi altında Monte-Carlo simülasyonları ve asimptotik sonuçlardan hesaplanan doğru seçim olasılıkları ile karar verilmiştir.

Anahtar Kelimeler: Doğru seçim olasılığı, Gamma dağılımı, Log-Normal dağılım, Olabilirlik oranı, Weibull dağılımı.

1. GİRİŞ

Gamma, Weibull ve Log-Normal dağılımları, özellikle, mühendislik, sağlık ve fen alanlarında çarpık verilerin analizi için sık kullanılan dağılımlardır ve oldukça yaygın bir uygulama alanlarına sahiptirler.

Literatürde, verilen bir veri setinin iki olasılık dağılımından hangisi ile modelleneceği ile ilgili birçok çalışma mevcuttur. Atkinson (1969, 1970), Cox (1961, 1962), Dyer (1973) kitleden alınan örneklemin, ele aldıkları iki olasılık dağılımından hangisi ile modelleneceği problemini ele almışlardır. Dumonceaux ve Antle (1973) Log-Normal ve Weibull dağılımlarını ayırt etmek için en çok olabilirlik oranına ilişkin kritik değerler elde etmişlerdir. Bain ve Englehardt (1980) simülasyon sonuçlarına göre Weibull ve Gamma dağlımlarına ilişkin doğru seçim olasılıklarını elde etmişlerdir. Wiens (1999) bir veri setinin dağılımının Log-Normal veya Gamma dağılımı olması durumlarını incelemiştir.

Veri setinin, ortak uygulama alanlarına sahip olan bu dağılımlardan hangisi ile modelleneceği oldukça önemlidir. Bu çalışmada, örneklemin dağılımının Gamma, Weibull ya da Log-Normal dağılımlarının hangisi ile modelleneceği problemi üzerinde durulmuştur. Elde edilen doğru seçim olasılığına göre, veri setinin dağılımının Gamma, Weibull ya da Log-Normal dağılımlarından hangisine uyduğuna karar verilmektedir. Doğru seçim olasılığı, en çok olabilirlik oranının dağılımına göre elde edilebilir. Ancak, bu oranın dağılımına ilişkin analitik ifadeler henüz elde edilemediği için, doğru seçim olasılığı olabilirlik oranının logaritmasının asimptotik dağılımına göre elde edilebilir. Kitleden alınan örneklemin dağılımının sırasıyla ilgili dağılımlara göre oluşturulan yokluk hipotezi altında elde edilen asimptotik dağılımlarına ve Monte-Carlo simülasyonlarında elde edilen sonuçlara göre doğru seçim olasılıkları hesaplanmıştır.

*Arş. Gör. Dr., Kırıkkale Üniversitesi, Fen-Edebiyat Fakültesi, İstatistik Bölümü, Kırıkkale, e-posta:hdbicer@hotmail.com

**Doç. Dr., Ankara Üniversitesi, Fen Fakültesi, İstatistik Bölümü, 06100 Ankara, e-posta: atakan@science.ankara.edu.tr

2. EN ÇOK OLABİLİRLİK ORANLARI

Verilen bir veri setinin dağılımının, Gamma, Weibull ya da Log-Normal dağılımlarından, hangisine uyduğuna karar verebilmek için bu dağılımlar ile ilgili hipotezler altında en çok olabilirlik fonksiyonlarının oranına bağlı bir karar kriteri elde edilebilir.

α ve β parametreli Gamma dağılımına sahip X rasgele değişkeninin olasılık yoğunluk fonksiyonu,

$$f_{GA}\left(x;\alpha,\beta\right)=\frac{1}{\beta^{\alpha}\Gamma\left(\alpha\right)}x^{\alpha-1}e^{-x/\beta}\ ,x>0,\alpha>0,\beta>0 \tag{1}$$

olmak üzere, α ve β parametrelerinin en çok olabilirlik tahmin edicileri arasındaki ilişki

$$\hat{\beta}=\frac{\bar{X}}{\hat{\alpha}} \tag{2}$$

olarak elde edilir.

η ve σ parametreli Log-Normal dağılıma sahip X rasgele değişkenin olasılık yoğunluk fonksiyonu,

$$f_{LN}\left(x;\sigma,\eta\right)=\frac{1}{\sqrt{2\pi}x\sigma}e^{-\frac{\left(\ln x-\ln\eta\right)^{2}}{2\sigma^{2}}}\ ,x>0,\ \sigma>0,\ \eta>0 \tag{3}$$

dır. η ve σ parametrelerinin en çok olabilirlik tahmin edicileri,

$$\hat{\eta}=\left(\prod_{i=1}^{n}X_{i}\right)^{\frac{1}{n}}\quad\text{ve}\quad\hat{\sigma}^{2}=\frac{1}{n}\sum_{i=1}^{n}\left(\ln X_{i}-\ln\hat{\eta}\right)^{2} \tag{4}$$

biçiminde elde edilir.

θ ve λ parametreli Weibull dağılımına sahip X rasgele değişkeninin olasılık yoğunluk fonksiyonu,

$$f_{WE}\left(x;\theta,\lambda\right)=\theta\lambda^{\theta}x^{\theta-1}e^{-\left(x\lambda\right)^{\theta}}\ ,x>0,\ \theta>0,\ \lambda>0 \tag{5}$$

olmak üzere, θ ve λ parametrelerinin en çok olabilirlik tahmin edicileri arasındaki ilişki

$$\hat{\lambda}=\left(n\Big/\sum_{i=1}^{n}X_{i}^{\hat{\theta}}\right)^{1/\hat{\theta}} \tag{6}$$

olarak elde edilir.

En çok olabilirlik oranının logaritması ile kitleden alınan örneklemin dağılımının Gamma ya da Log-Normal dağıldığına karar verilebilir. Yokluk ve alternatif hipotezler sırasıyla,

H_0 : Örneklem Gamma dağılımından alınmıştır.
H_1 : Örneklem Log-Normal dağılımından alınmıştır.

olmak üzere, $\hat{\alpha}$, $\hat{\beta}$, $\hat{\sigma}$ ve $\hat{\eta}$ en çok olabilirlik tahmin edicilerine bağlı olarak en çok olabilirlik oranının logaritması,

$$T_1 = \ln\left[\frac{L_{GA}\left(\hat{\alpha},\hat{\beta}\right)}{L_{LN}\left(\hat{\sigma},\hat{\eta}\right)}\right] = n\left[\ln\left(\Gamma\left(\hat{\alpha}\right)\right) - \ln\left(\hat{\sigma}\right) - \hat{\alpha}\ln\left(\frac{\tilde{X}}{\hat{\beta}}\right) + \hat{\alpha} - \frac{1}{2}\left(1+\ln\left(2\pi\right)\right)\right] \tag{7}$$

olarak elde edilir. Burada $\tilde{X} = \left(\prod_{i=1}^{n} X_i\right)^{1/n}$ örneklemin geometrik ortalamasıdır. Eğer $T_1 > 0$ ise veri setinin dağılımının Gamma dağılımından, aksi halde Log-Normal dağılımından olduğuna karar verilir. H_0'ın doğru olduğu koşulu altında (7) eşitliği ile verilen T_1'in dağılımı β'dan bağımsızdır ve sadece α parametresine bağlıdır. Benzer olarak, H_1'in doğru olduğu koşulu altında ise, T_1'in dağılımı sadece σ'ya bağlıdır.

Benzer biçimde, veri setinin dağılımının Log-Normal ya da Weibull dağıldığına karar verilebilir. Bu durumda, yokluk ve alternatif hipotezler sırasıyla,

H_0 : Örneklem Log-Normal dağılımından alınmıştır.
H_1 : Örneklem Weibull dağılımından alınmıştır.

olmak üzere, $\hat{\sigma}$, $\hat{\eta}$, $\hat{\theta}$ ve $\hat{\lambda}$ en çok olabilirlik tahmin edicilerine bağlı olarak en çok olabilirlik oranının logaritması,

$$T_2 = \ln\left[\frac{L_{LN}\left(\hat{\sigma},\hat{\eta}\right)}{L_{WE}\left(\hat{\theta},\hat{\lambda}\right)}\right] = n\left[\frac{1}{2} - \ln\left(\hat{\sigma}\hat{\theta}\left(\hat{\lambda}\hat{\eta}\right)^{\hat{\theta}}\sqrt{2\pi}\right)\right] \tag{8}$$

olarak elde edilir. Eğer $T_2 > 0$ ise örneklemin Log-Normal dağılımdan, aksi halde Weibull dağılımından geldiğine karar verilir. Eğer veri Log-Normal dağılımından geliyor ise, (8) eşitliği ile verilen T_2'nin dağılımı (σ,η)'dan bağımsızdır. Benzer olarak, eğer veri Weibull dağılımından geliyor ise, T_2'nin dağılımı (θ,λ)'dan bağımsızdır.

Örneklemin dağılımının Gamma ya da Weibull dağıldığına karar verilebilmesi için, yokluk ve alternatif hipotezler sırasıyla,

H_0 : Örneklem Gamma dağılımından alınmıştır.
H_1 : Örneklem Weibull dağılımından alınmıştır.

olmak üzere, $\hat{\alpha}$, $\hat{\beta}$, $\hat{\theta}$ ve $\hat{\lambda}$ en çok olabilirlik tahmin edicilerine bağlı olarak en çok olabilirlik oranının logaritması

$$T_3 = \ln\left[\frac{L_{GA}\left(\hat{\alpha},\hat{\beta}\right)}{L_{WE}\left(\hat{\theta},\hat{\lambda}\right)}\right] = n\left[\hat{\alpha}\ln\left(\hat{\alpha}\tilde{X}/\bar{X}\right) - \hat{\theta}\ln\left(\hat{\lambda}\tilde{X}\right) - \ln\left(\hat{\theta}\Gamma\left(\hat{\alpha}\right)\right) - \hat{\alpha} + 1\right] \quad (9)$$

olarak elde edilir. Burada $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ örneklemin aritmetik ortalamasıdır.

Eğer $T_3 > 0$ ise örneklemin dağılımının Gamma dağılımı, aksi halde Weibull dağılımından olduğuna karar verilir.

3. EN ÇOK OLABİLİRLİK ORANLARININ ASİMPTOTİK DAĞILIMLARI

Bu bölümde ilgili veri setinin iki dağılımdan hangisinden geldiğine karar vermek için asimptotik sonuçlara bağlı karar kuralı elde edilmeye çalışılacaktır. Bu amaçla, karşılaştırmaya ilişkin hipotezler

A) A1. $H_0 : X \sim f_{GA}\left(x;\alpha,\beta\right)$ *vs* $H_1 : X \sim f_{LN}\left(x;\sigma,\eta\right)$
   A2. $H_0 : X \sim f_{LN}\left(x;\sigma,\eta\right)$ *vs* $H_1 : X \sim f_{GA}\left(x;\alpha,\beta\right)$
B) B1. $H_0 : X \sim f_{LN}\left(x;\sigma,\eta\right)$ *vs* $H_1 : X \sim f_{WE}\left(x;\theta,\lambda\right)$
   B2. $H_0 : X \sim f_{WE}\left(x;\theta,\lambda\right)$ *vs* $H_1 : X \sim f_{LN}\left(x;\sigma,\eta\right)$
C) C1. $H_0 : X \sim f_{GA}\left(x;\alpha,\beta\right)$ *vs* $H_1 : X \sim f_{WE}\left(x;\theta,\lambda\right)$
   C2. $H_0 : X \sim f_{WE}\left(x;\theta,\lambda\right)$ *vs* $H_1 : X \sim f_{GA}\left(x;\alpha,\beta\right)$

olarak göz önüne alınsın. H_0 yokluk hipotezleri altında en çok olabilirlik oranının logaritmasının asimptotik dağılımlarının elde edilmesinde kullanılacak olan Lemma ve Teorem, A1 durumu göz önüne alınarak, verinin Gamma dağılımına sahip olduğu varsayımı altında verilmiştir. Diğer durumlar için de benzer sonuçlar aynı şekilde elde edilir.

Lemma : $X_1, X_2, \ldots, X_n$ rasgele örneklemi Gamma dağıldığında $n \to \infty$ için aşağıdaki asimptotik sonuçlar elde edilir. Burada a.s. ile hemen hemen her yerde yakınsama ifade edilmektedir.

a. $E_{GA}\left(\ln f_{GA}\left(X;\alpha,\beta\right)\right) = \max_{\bar{\alpha},\bar{\beta}} E_{GA}\left(\ln f_{GA}\left(X;\bar{\alpha},\bar{\beta}\right)\right)$
   olduğunda $\hat{\alpha} \to \alpha$ a.s. ve $\hat{\beta} \to \beta$ a.s. dır.

b. $E_{GA}\left(\ln f_{LN}\left(X;\tilde{\sigma},\tilde{\eta}\right)\right)=\max_{\sigma,\eta} E_{GA}\left(\ln f_{LN}\left(X;\sigma,\eta\right)\right)$
olduğunda $\hat{\sigma}\to\tilde{\sigma}$ a.s. ve $\hat{\eta}\to\tilde{\eta}$ a.s. dır.

c. $T_1^*=\ln\left[\dfrac{L_{GA}\left(\alpha,\beta\right)}{L_{LN}\left(\tilde{\sigma},\tilde{\eta}\right)}\right]$ olarak tanımlansın. $\dfrac{1}{\sqrt{n}}\left[T_1-E_{GA}\left(T_1\right)\right]$ ve $\dfrac{1}{\sqrt{n}}\left[T_1^*-E_{GA}\left(T_1^*\right)\right]$ aynı asimptotik dağılıma sahiptirler.

TEOREM : Eğer veri Gamma dağılımına sahip ise, T_1, $E_{GA}\left(T_1\right)$ ortalama ve $Var_{GA}\left(T_1\right)$ varyansı ile asimptotik normal dağılır.

İspat: Merkezi Limit Teoremi kullanılarak $\dfrac{1}{\sqrt{n}}\left[T_1-E_{GA}\left(T_1\right)\right]$'nin asimptotik normal dağıldığı gösterilebilir. Böylece, Merkezi Limit Teoremi ve Lemma c. den ispat tamamlanmış olur (White, 1982).

Lemma b'den, $\tilde{\sigma}$, $\tilde{\eta}$, $E_{GA}\left(T_1\right)$ ve $Var_{GA}\left(T_1\right)$ ifadelerini elde etmek için

$$\begin{aligned} g\left(\sigma,\eta\right) &= E_{GA}\left(\ln\left(f_{LN}\left(X;\sigma,\eta\right)\right)\right) \\ &= E_{GA}\left[-\frac{1}{2}\ln\left(2\pi\right)-\ln\sigma-\psi\left(\alpha\right)-\ln X-\frac{1}{2\sigma^2}\left(\ln X-\ln\eta\right)^2\right] \\ &= -\frac{1}{2}\ln\left(2\pi\right)-\ln\left(\sigma\beta\right)-\psi\left(\alpha\right)-\frac{1}{2\sigma^2}\left[\psi'\left(\alpha\right)+\left(\psi\left(\alpha\right)+\ln\left(\beta/\eta\right)\right)^2\right] \end{aligned} \qquad (10)$$

fonksiyonunu tanımlayalım. $g\left(\sigma,\eta\right)$'nın sırasıyla σ ve η'ya göre türevlerinin sıfıra eşitlenmesi ile

$$\tilde{\sigma}=\left(\psi'\left(\alpha\right)\right)^{1/2} \text{ ve } \tilde{\eta}=\beta e^{\psi(\alpha)} \qquad (11)$$

olarak elde edilir. Burada $\psi\left(\alpha\right)=\dfrac{d}{d\alpha}\ln\left(\Gamma\left(\alpha\right)\right)$ ve $\psi'\left(\alpha\right)=\dfrac{d}{d\alpha}\psi\left(\alpha\right)$ dır [4].

T_1 'nin asimptotik beklenen değeri ve varyansı;

$$\begin{aligned} \frac{E_{GA}\left(T_1\right)}{n} \approx AE_{GA} &= E_{GA}\left[\ln f_{GA}\left(X;\alpha,1\right)-\ln f_{LN}\left(X;\tilde{\sigma},\tilde{\eta}\right)\right] \\ &= \frac{1}{2}\ln\left(2\pi\right)+\frac{1}{2}+\ln\tilde{\sigma}-\ln\left(\Gamma\left(\alpha\right)\right)+\alpha\left(\psi\left(\alpha\right)-1\right) \end{aligned}$$

ve

$$\frac{Var_{GA}(T_1)}{n} \approx AVar_{GA} = Var_{GA}\left[\ln f_{GA}(X;\alpha,1) - \ln f_{LN}(X;\tilde{\sigma},\tilde{\eta})\right]$$
$$= -\frac{1}{\tilde{\sigma}^2}\Big[\alpha(\alpha+1)\big(\psi'(\alpha+2)+\psi(\alpha+2)\big)^2$$
$$-\psi(\alpha)\big(1-\psi(\alpha)(\alpha+2)\big)-\alpha\psi''(\alpha)$$
$$-\frac{1}{4\tilde{\sigma}^2}\big(\psi'''(\alpha)-3\psi(\alpha)\psi''(\alpha)\big)\Big]$$
$$+\alpha+\alpha^2\tilde{\sigma}^2-\psi(\alpha)(\alpha-2)-\frac{3}{2}$$

dir. Bu durumda, T_1, $E_{GA}(T_1)$ oratalama ve $Var_{GA}(T_1)$ varyanslı asimptotik normal dağıldığından, doğru seçim olasılığı; yani A1 de verilen H_0 hipotezinin red edilememe olasılığı

$$P(T_1>0) \approx \Phi\left(-\frac{E_{GA}(T_1)}{\sqrt{Var_{GA}(T_1)}}\right) = \Phi\left(-\frac{n\times AE_{GA}}{\sqrt{n\times AVar_{GA}}}\right) \qquad (12)$$

dir. Burada Φ standart normal rasgele değişkenin kümülatif dağılım fonksiyonudur.

4. SİMÜLASYON ÇALIŞMASI

Bu bölümde, farklı örneklem büyüklükleri için Bölüm 3 de elde edilen asimptotik sonuçların simülasyon çalışması ile Monte-Carlo simülasyon çalışmalarına bağlı olarak doğru seçim olasılıkları hesaplanmıştır.

A: H_0 hipotezi Gamma olduğu ve alternatif hipotezi Log-Normal olduğu durum ele alınmıştır. $L_{GA}(\alpha,2)$'den örneklem büyüklüğü $n=20$, 40, 60, 80 ve 100 birimlik rasgele örneklemler 10000 tekrarlı üretilerek (7) eşitliği ile verilen T_1'in sıfırdan büyük olup olmadığı kontrol edilmiştir. Doğru seçim olasılığının bir tahmin değerini elde etmek için, T_1'nin sıfırdan büyük olduğu durumların sayısı hesaplanmıştır. Asimptotik sonuçlar kullanılarak (12) eşitliği ile verilen doğru seçim olasılıkları hesaplanarak, sonuçlar Tablo 1'de verilmiştir. H_0 hipotezinin Log-Normal olduğu durum için de $L_{LN}(\sigma,2)$'den örneklem büyüklüğü $n=20$, 40, 60, 80 ve 100 birimlik rasgele örneklemler üretilerek, $T_1<0$ olup olmadığı kontrol edilmiştir. Sonuçlar Tablo 2'de verilmiştir. Her bir tabloda, ilk eleman Monte-Carlo simülasyonlarının sonucunu ve parantez içindeki değerler asimptotik sonuçlardan elde edilen doğru seçim olasılıklarını göstermektedir.

Tablo 1. Veri Gamma dağıldığında Monte-Carlo simülasyonları ve asimptotik sonuçlara bağlı doğru seçim olasılıkları

$\alpha \downarrow n \rightarrow$	20	40	60	80	100
	MC AS	MC AS	MC AS	MC AS	MC AS
2	0.7245(0.7216)	0.8023(0.7970)	0.8512(0.8456)	0.8797(0.8800)	0.9058(0.9055)
4	0.6549(0.6675)	0.7394(0.7299)	0.7844(0.7734)	0.8098(0.8068)	0.8354(0.8336)
6	0.6303(0.6419)	0.7083(0.6965)	0.7419(0.7356)	0.7652(0.7665)	0.7953(0.7919)
8	0.6399(0.6285)	0.6851(0.6786)	0.7251(0.7150)	0.7478(0.7440)	0.7673(0.7683)
10	0.6249(0.6154)	0.6718(0.6609)	0.7059(0.6943)	0.7245(0.7213)	0.7442(0.7441)

Tablo 2. Veri Log-Normal dağıldığında Monte-Carlo simülasyonları ve asimptotik sonuçlara bağlı doğru seçim olasılıkları

$\sigma \downarrow n \rightarrow$	20	40	60	80	100
	MC AS	MC AS	MC AS	MC AS	MC AS
0.5	0.6645(0.6827)	0.7493(0.7414)	0.7912(0.7848)	0.8197(0.8242)	0.8408(0.9583)
0.7	0.7149(0.7207)	0.7994(0.7959)	0.8494(0.8444)	0.8947(0.8789)	0.9054(0.9045)
0.9	0.7603(0.7509)	0.8453(0.8309)	0.8819(0.8796)	0.9152(0.9122)	0.9357(0.9350)
1.1	0.7699(0.7723)	0.8651(0.8545)	0.9051(0.9020)	0.9417(0.9323)	0.9510(0.9525)
1.3	0.7849(0.7876)	0.8798(0.8705)	0.9199(0.9166)	0.9495(0.9448)	0.9632(0.9628)

B: H_0 hipotezi Log-Normal olduğu ve alternatif hipotezi Weibull olduğu durum ele alınmıştır. $L_{LN}(2,2)$'den örneklem büyüklüğü $n = 20,\ 40,\ 60,\ 80$ ve 100 birimlik rasgele örneklemler üretilerek (8) eşitliği ile verilen $T_2 > 0$ olup olmadığı kontrol edilmiştir. Sonuçlar Tablo 3'de verilmiştir. H_0 hipotezi Weibull ve alternatif hipotez Log-Normal olduğu durum ele alınarak $L_{WE}(2,2)$'den n birimlik rasgele örneklemler üretilerek, $T_2 < 0$ olup olmadığı kontrol edilmiştir. Sonuçlar Tablo 4'de verilmiştir.

Tablo 3. Veri Log-Normal dağıldığında Monte-Carlo simülasyonları ve asimptotik sonuçlara bağlı doğru seçim olasılıkları

n	20	40	60	80	100
MC	0.7838	0.8668	0.9162	0.9421	0.9595
AS	0.7808	0.8632	0.9110	0.9394	0.9588

Tablo 4. Veri Weibull dağıldığında Monte-Carlo simülasyonları ve asimptotik sonuçlara bağlı doğru seçim olasılıkları

n	20	40	60	80	100
MC	0.7782	0.8660	0.9090	0.9389	0.9581
AS	0.7766	0.8590	0.9062	0.9360	0.9556

C: H_0 hipotezi Gamma olduğu ve alternatif hipotezi Weibull olduğu durum ele alınmıştır. Örneklem büyüklüğü $n = 20,\ 40,\ 60,\ 80$ ve 100 ve şekil parametresi $\alpha = 2,\ 4,\ 6,\ 8,\ 10$ ve 12 olarak seçilmiştir. $L_{GA}(\alpha,2)$'den n birimlik rasgele örneklem üretildi ve $T_3 > 0$ olup olmadığı kontrol edilmiştir. Sonuçlar Tablo 5'de verilmiştir. Benzer olarak, H_0 hipotezi Weibull ve alternatif hipotez Gamma olduğu durum ele

alınarak, örneklem büyüklüğü $n = 20,\ 40,\ 60,\ 80$ ve100 ve $\theta = 2,\ 4,\ 6,\ 8$ ve 10 olduğu durumlar için $T_3 < 0$ olup olmadığı kontrol edildi. Sonuçlar Tablo 6'da verilmiştir.

Tablo 5. Veri Gamma dağıldığında Monte-Carlo simülasyonları ve asimptotik sonuçlara bağlı doğru seçim olasılıkları

$\alpha \downarrow n \rightarrow$	20	40	60	80	100
	MC AS	MC AS	MC AS	MC AS	MC AS
2	0.8914(0.8907)	0.9014(0.9088)	0.9203(0.9103)	0.9367(0.9329)	0.9601(0.9592)
4	0.8370(0.8384)	0.8420(0.8479)	0.8627(0.8595)	0.8024(0.8797)	0.8923(0.8924)
6	0.7185(0.7068)	0.7476(0.7465)	0.7568(0.7530)	0.7739(0.7681)	0.7839(0.7833)
8	0.5794(0.5812)	0.5750(0.5627)	0.5938(0.5912)	0.6163(0.6088)	0.5962(0.5965)
10	0.5468(0.5340)	0.5369(0.5217)	0.5456(0.5438)	0.5390(0.5343)	0.5320(0.5342)
12	0.5278(0.5115)	0.5073(0.5078)	0.5171(0.5102)	0.5115(0.5097)	0.5095(0.5096)

Tablo 6. Veri Weibull dağıldığında Monte-Carlo simülasyonları ve asimptotik sonuçlara bağlı doğru seçim olasılıkları

$\theta \downarrow\ n \rightarrow$	20	40	60	80	100
	MC AS	MC AS	MC AS	MC AS	MC AS
2	0.5495(0.5271)	0.5513(0.5386)	0.5578(0.5470)	0.5588(0.5544)	0.5615(0.5609)
4	0.5449(0.5224)	0.5501(0.5344)	0.5494(0.5435)	0.5509(0.5492)	0.5541(0.5530)
6	0.5273(0.5143)	0.5309(0.5218)	0.5369(0.5258)	0.5316(0.5300)	0.5339(0.5321)
8	0.5228(0.5103)	0.5198(0.5143)	0.5243(0.5176)	0.5204(0.5193)	0.5237(0.5229)
10	0.5184(0.5073)	0.5201(0.5099)	0.5248(0.5131)	0.5153(0.5148)	0.5165(0.5156)

5. TARTIŞMA VE SONUÇ

Bu çalışmada, verilen bir veri setinin, ortak uygulama alanlarına sahip olan Gamma, Weibull ya da Log-Normal dağılımlarından hangisine uyduğuna karar vermek için ilgili dağılımlara göre oluşturulan yokluk hipotezi altında Monte-Carlo simülasyonlarından ve asimptotik sonuçlardan doğru seçim olasılıkları elde edilmiştir.

Elde edilen tablolardan, örneklem büyüklüğü arttıkça doğru seçim olasılıklarının arttığı ve Monte-Carlo simülasyonu ile elde edilen doğru seçim olasılığı ile asimptotik sonuçlardan elde edilen doğru seçim olasılığı arasındaki farkın küçüldüğü görülmektedir.

Asimptotik sonuçlara göre elde edilen doğru seçim olasılıklarına bakıldığında büyük örneklemlerde olduğu gibi küçük örneklemler için de asimptotik sonuçların Monte-Carlo kadar iyi sonuçlar verdiği görülmektedir.

6. KAYNAKLAR

Atkinson, A., 1969. A Test of Discriminating Between Modes, Biometrika, 56, 337–341.

Atkinson, A., 1970. A Method for Discriminating Between Models (with discussions), Journal of the Royal Statistical society. Ser. B, 32, 323–353.

Bain, L. J., Englehardt, M., 1980. Probability of Correct Selection of Weibull Versus Gamma Based on Likelihood Ratio. Communications in Statistics. Series A, 9, 375–381.

Bernardo, J. M., 1976. Algorithm As 103: Psi (Digamma) Function, J. Roy. Statist. Soc. Ser. C, 25, 315-317.

Cox, D. R., 1961. Tests of Separate Families of Hypotheses, Proceedings of the Fourth Berkeley Symposium in Mathematical Statistics and Probability, Berkeley, University of California Press, 105–123.

Cox, D. R., 1962. Further Results on Tests of Separate Families of Hypotheses, Journal of the Royal Statistical Society, Ser. B, 24, 406–424.

Dumonceaux, R., Antle, C.E., 1973. Discriminating Between the Log-normal and Weibull Distribution, Technometrics, 15, 923–926.

Dyer, A. R., 1973. Discrimination Procedure for Separate Families of Hypotheses, Journal of the American Statistical Association, 68, 970–974.

Johnson, N. L., Kotz, S., Balakrishnan, N., 1994. Continuous Univariate Distributions, Volume 1, John Wiley & Sons, Inc. USA.

Wiens, B.L., 1999. When Log-Normal and Gamma Models Give Different Results: A Case Study, American Statistician, 53, 89–93.

White, H., 1982. Requlority Conditions for Cox's Test of Non-nested Hypotheses, J. Econometrics, 19, 301–318.

DISCRIMINATION OF THE GAMMA, WEIBULL AND LOG-NORMAL DISTRIBUTIONS ACCORDING TO PROBABILITY OF CORRECT SELECTION

ABSTRACT

Gamma, Weibull and Log-Normal distributions are often used distributions for modelling skewed data. In this study, for a given data set the problem of selecting Gamma, Weibull or Log-Normal distributions is focused on for the modelling. Distribution of a given set of data, Gamma, Weibull or Log-Normal distribution has been decided by the probability of correct selection that is calculated by constructing Monte-Carlo simulations and asymptotic results under the null hypothesis.

Keywords: Probability of correct selection, Gamma distribution, Log-Normal distribution, Likelihood ratio tests, Weibull distribution.

Volume: 09 Number: 01 Category: 03 Page: 21-28 ISSN No: 1303-6319 Cilt: 09 Sayı: 01 Kategori: 03 Sayfa: 21-28

ERP YAZILIMI KULLANAN VE KULLANMAYAN İŞLETMELER ARASINDAKİ YÖNETİMSEL VE TEKNOLOJİK FARKLILIKLARIN TANIMLAYICI VERİ MADENCİLİĞİ YÖNTEMLERİYLE İNCELENMESİ

Betül YAVAŞOĞLU* **Ahmet Selman BOZKIR****

ÖZET

ERP (Enterprise Resource Planning–Kurumsal Kaynak Planlama) yazılımları, şirket gereksinimlere göre tasarlanmış, esnek ve üretim/hizmet gibi konularda yüksek verim hedefleyen gelişmiş yazılım çözümleridir. Bu çalışmada tanımlayıcı veri madenciliği yöntemlerinden kümeleme ve birliktelik kuralları analizi ile kurumların ERP kullanmaları halinde kazançlarının ne olacağı, geçiş öncesinde ne gibi altyapısal unsurlara sahip olmaları gerektiği ortaya çıkarılmaya çalışılmıştır. Kümeleme çalışmasında ERP sistemlere sahip olan ve olmayan kurumlar arasındaki doğal farklılıklar ortaya çıkarılmıştır. Sonuçlara bakıldığında güncel kapasite kullanım oranlarının ve bazı departmanlara sahip olma yüzdesinin ERP sistemlere sahip olma ile yüksek derecede pozitif korelasyon gösterdiği tespit edilmiştir. Öte yandan, birliktelik kuralları analizi ile ERP sistemlere sahip olan ve olmayan işletmelerde sıklıkla gözlemlenen kurallar çıkarılmıştır.

Anahtar Kelimeler: Birliktelik kuralları analizi, ERP, Kümeleme, Veri madenciliği.

1. GİRİŞ

Enterprise Resource Planning (ERP) yazılımları, tanım olarak işletme içindeki materyal, bilgi ve finansal kaynak akışlarının ortak bir veritabanı üzerinden yürütülmesini sağlayan, iş süreçleri ve bilgi teknolojilerinin biraraya getirilmesiyle üretilmiş entegre yazılım sistemleridir (Su ve Yang, 2009). Bununla birlikte ERP yazılımları, işletme gereksinimlerine cevap veren, esnek, parametrik ve üretim/hizmet gibi konularda yüksek verim hedefleyen çözümlerdir. ERP'nin çıkışı, 60'lı yılların sonlarında üretim ve dağıtım şirketlerinin stoklarını daha sağlıklı yönetebilmek ve malzeme tedariklerini planlayabilmek için kullandıkları MRP (Material Requirements Planning – Malzeme İhtiyaç Planlaması) yazılımlarının doğuşu ile başlamaktadır. İlerleyen yıllarla birlikte ek modüllerin de (üretim ve kapasite planlaması, finans ve pazarlama) eklenmesiyle MRP II'ye geçilmiş, 90'lı yıllara gelindiğinde ise muhasebe, lojistik ve üretimle ilgili yeni özelliklerin dahil edilmesiyle birlikte bu yazılım teknolojisi evrimleşerek ERP adını almıştır (Yavasoglu B., 2011 ve Jacobs ve Weston, 2007).

Ancak, özellikle 2000 yılından sonra çok hızlı büyüme sergileyen internet ve uygulamaları, önemi daha çok artan müşteri ilişkileri yönetimi ve modern dünyanın kahini sayılan veri madenciliğinin iş süreçlerine girmiş hali olan iş zekasının ERP'ye dahil olması ile ERP gelişerek ERP II ortaya çıkmıştır.

*İstatistikçi, Türkiye İstatistik Kurumu İstanbul Bölge Müdürlüğü, İstanbul, e-posta: byavasoglu@gmail.com

**Arş. Gör., Hacettepe Üniversitesi Bilgisayar Mühendisliği Bölümü, Ankara, e-posta: selman@cs.hacettepe.edu.tr

Görüldüğü üzere ERP sürekli evrimleşmiş ve gelişmiştir ve her seferinde işletmelerdeki sahası ve rolü daha da artmıştır. Ancak bu gelişimle birlikte işletmelerin ERP'ye terfi etme kararı almaları eskiye nazaran zorlaşmıştır. Cebeci (2009)'ye göre ERP yazılımlarının yüksek maliyeti, olası entegrasyon sorunları ve personel adaptasyonu gibi konular işletmeler açısından risk teşkil etmektedir. Aynı zamanda, Saatcioglu (2007), ERP yazılımlarını hem karmaşık olarak nitelendirmiş hem de bu sistemlerin para, zaman ve deneyim konularında ciddi yatırım gerektirdiğinin altını çizmiştir. Buna karşın Wu (2010), ERP'ye geçişin tamamlanmasıyla elde edilecek faydalar göz önüne alındığında birçok işletmenin ERP'ye sıcak baktığını ve bu riski göze alarak ERP'ye terfi ettiğini belirtmiştir. Bu noktada, işletmeler açısından ERP'ye terfi öncesinde ERP'nin kurum için ne derece gerektiği, kuruma kazandıracağı faydalar ve işletmenin bu dönüşüm için gerekli altyapısal unsurlara sahip olup olmadığı öncelikle sorgulanması gereken noktalar olarak ortaya çıkmaktadır. Literatür incelendiğinde, ERP'ye geçiş halinde elde edilecek faydalar ve adaptasyon konuları çeşitli çalışmalarda sıklıkla işlenmektedir. Bir örnek olarak Buonanno v.d. (2009), küçük-orta ölçekli işletmeler ve büyük işletmelerin ERP'ye adaptasyon sürecinde yaşadıkları farklılıkları ortaya koymuştur. Ancak konuyla ilgili literatür taraması yapıldığında günümüzün önemli veri analiz enstürmanlarından biri olan veri madenciliği yöntemlerinin yer aldığı bir çalışmaya rastlanılmamıştır. Bu çalışmada, kurumların ERP sistemine geçmeleri halinde kazanımlarının ne olacağı ve geçiş öncesinde hangi altyapısal unsurlara sahip olmaları gerektiği, günümüzde önemi daha da artmakta olan veri madenciliği yöntemlerinden kümeleme ve birliktelik analizi yardımıyla ortaya çıkarılmaya çalışılmıştır. Çalışma kapsamınca, önceden yapılan bir anket çalışmasından elde edilen veriler ışığında, ERP yazılımına sahip olan/olmayan yurtiçi işletmelerin profilleri ve ERP hakkındaki görüşleri elde edilmiş, sonrasında kümeleme analizi ile ERP'ye terfi etmiş/etmemiş işletmeler arasındaki farklılıklar belirlenmeye çalışılmıştır.. İkinci safhada birliktelik kuralları (association rules) yöntemi kullanılarak, kurumların ERP'ye sahip olma/olmama değişkeninin diğer değişkenlerle birlikte gözlemlenen birliktelik örüntüleri ortaya çıkarılmıştır.

2. VERİ MADENCİLİĞİ

Veri madenciliği tanım olarak büyük miktarda veri içersindeki anlamlı örüntü ve bilginin istatistik, yapay zeka ve makine öğrenmesi gibi yöntembilimler ışığında akıllı algoritmalar yardımıyla keşfedilmesi sürecidir (Bozkir vd., 2008). Veri madenciliği yöntemleri temelde iki gruba ayrılmaktadır. İlk yöntem grubu olan kestirimsel yöntemlerde amaç, eldeki mevcut verinin yardımıyla hiç görülmemiş örneklerin tahmin edilmesi iken tanımlayıcı yöntemlerde amaç veri içersindeki anlamlı ilişkilerin, korelasyonların ve örüntülerin ortaya çıkarılmasıdır. Karar/regresyon ağaçları, destek vektör makineleri ve yapay sinir ağları gibi yöntemler kestirimsel yöntemlerde yer alırken, çalışmada kullanılan kümeleme ve birliktelik kuralları analizi, tanımlayıcı veri madenciliği grubuna üye yöntemlerdir.

Kümeleme, en basit tarifiyle nesneleri özelliklerine göre verilen bir sayıda gruba ayırma işlemidir (Tang ve MacLennan, 2005). Kümeleme yöntemlerinin temeli uzaklık ilkesine dayalıdır. Örneklemdeki nesnelerin her bir niteliği bir boyut olarak ele alınır, nesne bu *n* boyutlu uzayda bir nokta olarak konumlandırılır. İkinci aşamada uzaklık tabanlı (ör: öklid) kümeleme algoritması bu noktalardan verilen küme sayısı kadar noktayı başlangıç için rasgele seçer ve bunları ilk küme merkezi olarak atar. Devamında diğer her seçilecek noktayı bu noktalara olan uzaklıklarına bakarak bir kümeye atar ve o

kümenin merkezini tekrardan yeni gelen örneği gözeterek hesaplar ve işlem tüm noktalar bir kümeye atanana kadar sürer. Kümeleme yöntemleri genel olarak katı kümeleme ve esnek kümeleme olarak ikiye ayrılmaktadır. Katı kümeleme yöntemlerinde (ör: K-means) her bir nokta ancak ve ancak bir kümeye ait iken esnek kümeleme yöntemlerinde (ör: EM, bulanık kümeleme) her örnek belli bir üyelik derecesi ile birçok kümeye ait olabilir (Tang ve MacLennan, 2005). Çalışma kapsamınca esnek kümeleme yapabilen EM (Expectation Maximization – Beklenti Ençoklaştırımı) tabanlı Microsoft Clustering (Microsoft Kümeleme) algoritması kullanılmıştır. EM algoritması prensip olarak her bir boyut (değişken) için bir ortalamaya ve standart sapmaya sahip çan eğrileri çıkarır. Bir örnek bu eğrilerin altında kaldığı noktanın yerine göre birçok kümeye farklı üyelik dereceleri ile atanmaktadır (Tang ve MacLennan, 2005). En yüksek üyelik derecesi kazanılan küme, örneğin nihai atandığı küme olarak belirlenmektedir. Bu şekilde bir yaklaşımla uzaklık ölçütüne bağlı katı kümeleme yerine olasılıksal bir çatı altında esnek kümeleme yapılabilmekte ve örneklerin üyelikleri gözlemlenebilmektedir.

Diğer bir tanımlayıcı veri madenciliği yöntemi olan birliktelik kuralları analizi tanım olarak, veri içersinde sıklıkla beraber varolan örnekleri tespit etme işlemidir. Bu konuyla ilgili çalışmalar 80'li yıllarda başlamıştır ancak şu an için literatürde en yaygın kullanılan Apriori algoritması Agrawal ve arkadaşları tarafından ortaya konulmuştur (Agrawal vd., 1983). Apriori algoritması, gerçekte ağaç tabanlı çalışan ve belli bir eşik değerinin altında yer alan birliktelikleri budayıp, geçmeyi başarabilen birlikteliklerle yola devam eden iki aşamalı bir algoritmadır. İlk aşamada sıklıkla görülen öğe kümeleri (itemsets) tespit edilirken ikinci aşamada elde kalan birlikteliklerden kurallar (rules) çıkarılır. A, B $\Rightarrow$ C şeklinde oluşan kurallarda A ve B kuralın önceli (antecedent), C ise izleyeni (consequent) şeklinde adlandırılmaktadır ve A ile B varken C'de olmaktadır şeklinde okunabilmektedir. Algoritmanın çalışması için minimum destek (her kuralda gerekli minimal örnek sayısı) ve güven değeri (A ve B olurken C'nin de olması olasılığının minimal yüzdesi) gibi parametreler belirtilmelidir.

3. VERİ

Çalışmada kullanılan veri kümesi Türkiye içersinde yer alan 129 adet büyük ve orta ölçekteki işletmelere uygulanan bir ankete dayanmaktadır. Anket içersinde 30 adet soru bulunmaktadır. Sorulardan üçü hariç diğerleri kategorik türdedir. Anketin güvenilirliğinin bir ölçütü olan *Cronbach Alpha* değeri %94'tür.

Veri kümesi içerisinde 47 işletme ERP yazılımına sahip iken, 82 işletme böyle bir teknolojiye sahip değildir. Tablo 1'de görüleceği üzere işletmenin sektörü, faaliyet alanı, çalışan sayısı, gelir durumu, çeşitli birimlere ve sertifikasyon belgelerine sahip olma durumu, 2006, 2007 ve 2008 yıllarındaki kapasite kullanım oranı ile son olarak ERP yazılımına sahip olma durumu soruları ankette yer almıştır. Kapasite kullanım oranlarının yer aldığı sorulara %15 lik bir kesim yanıt vermemiştir.

Tablo 1. Çalışmada kullanılan verilerin nitelikleri

Değişken Adı	Veri Türü	Alabileceği Değerler
Faaliyet Sektör	Kesikli	İmalat, Hizmet, Enerji, İnşaat, Diğer
Faaliyet Alan	Kesikli	Bilgisayar / Elektronik, Gıda, İlaç, Kağıt Urunleri, Kauçuk/Plastik, Kimya, Mermer, Metal, Mobilya, Otomotiv,Tekstil, Diğer
Çalışan Sayısı	Kesikli	10'dan az, 10 – 49, 50 – 249, 250'den fazla
Gelir Durumu	Kesikli	10.000 TL'den az, 10.000 – 50.000 TL, 50.000 – 100.000 TL, 100.000 – 250.000 TL, 250.000 – 500.000 TL, 500.000 – 1.000.000 TL, 1.000.000 TL'den fazla
Bilgisayar Entegrasyonu	Kesikli	Kesinlikle Katılıyorum, Katılıyorum, Kararsızım, Katılmıyorum, Kesinlikle Katılmıyorum
Geleneksel Tezgah, CNC Tezgah, Özel Amaç Tezgah, Otomatik Üretim Hattı, Robot, Servis İşletmesi, ISO9000-1994, ISO9000-2000, ISO14000, CE, OHSAS18001, Üretim Birimi , İnsan Kaynakları Birimi, Finans Birimi, Satınalma Birimi, Arge Birimi, Kalite Birimi, Birim Yok	0/1	Var - Yok
2006 Kapasite Kullanımı, 2007 Kapasite Kullanımı, 2008 Kapasite Kullanımı	Sürekli	0 – 100 arası sayısal değer
Erp Kurumda Var mı?	0/1	Yok

1. YÖNTEM VE BULGULAR

Çalışma kapsamınca kümeleme ve birliktelik kuralları çalışması yapılmıştır. Kullanılan yöntemlere ait algoritmalar Microsoft Analysis Services for SQL Server 2008 içerisinde yer alan Microsoft Clustering (EM tabanlı) ve Apriori algoritmalarıdır. Çalışmanın üzerinde yer aldığı platform olarak Microsoft Analysis Services'in seçilmesinin temel nedenleri, diğer bazı veri madenciliği paketlerinin (Orange, Weka vb.) aksine kümeleme yöntemlerinde küme sayısı verilmeksizin doğal kümeleme yapılmasına olanak tanıması (Tang ve MacLennan, 2005), doğrudan ilişkisel veritabanı üzerinden çalışabilmesi ve kaliteli görselleştirme hizmetleri sunabilmesidir. Öte yandan kümeleme yönteminin kullanım amaçları ve genel beklentiler şu şekilde listelenebilir:

1. ERP sistemine sahip olan ve olmayan işletmelerin kümeleme çalışması yerine doğrudan iki grup olarak ele alınıp, değişken dağılımlarının farklarını ortaya koymak yerine doğal kümeleme yapılmış ve işletmelerin gerçekten de sahip oldukları niteliklere göre ERP'ye sahip olan veya olmayan kümelerde yer alıp almadıkları tespit edilmiştir.
2. Küme sayısı verilmeksizin verinin gerçekten de iki doğal kümeye ayrılıp ayrılmadığı gözlemlenmiştir (ERP'ye sahip olanlar ve olmayanlar).

Kümeleme analizi için katı kümeleme yapan K-means algoritması yerine esnek kümeleme yapan EM algoritması tercih edilmiştir. Doğal kümeleme yapabilmek için CLUSTER_COUNT parametresi 0'a ayarlanarak 129 işletmenin herhangi bir manipülasyon olmaksızın doğal biçimde kümelenmesi sağlanmıştır. Algoritma iteratif olarak 5 defa çalıştırılmış ve her seferinde 2 adet küme elde edilmiştir.

Şekil 1. Kümeleme sonuçlarından bir kesit

Gerçekte Microsoft Analysis Services sonuçları tek sutunda listelemektedir. Ancak bu yayının sayfa sayısının kısıtlı olması nedeniyle şekil üzerinde birleştirme yapılarak iki sutun yan yana gelecek şekilde daha çok bilginin sunulması amaçlanmış ve Şekil 1 elde edilmiştir. Şekil 1'de görüleceği üzere 129 işletme, ilki 80 diğeri 49 işletme içeren iki kümeye ayrılmıştır. Birinci kümede ERP sahibi olan ve olmayan işletmeler sırası ile %54 ve %46 olarak belirlenmiştir. İkinci kümede ise ERP sahibi olmayan işletmeler yoğunluktadır ve %87 seviyesindedir. Bundan yola çıkılarak ikinci kümenin ERP sahibi olmayan işletmeleri temsil edebileceği, ilk kümenin ise homojen bir dağılıma sahip olduğu varsayımı ile yorumlamalar yapılmış ve Tablo 2'de sunulmuştur.

Tablo 2'de sunulan değişkenler dışındaki değişkenlerde (robot teknolojisi, servis işletmesi, ISO9000-1994 ve diğer birimlere sahiplik) kümeler bazında gözlemlenebilir bir fark yakalanamamıştır.

Tablo 2. Kümeleme sonrası elde edilen bulgular

ERP sahibi işletmeler (Küme 1)	ERP sahibi olmayan işletmeler (Küme 2)
• 2006, 2007 ve 2008 kapasite kullanım oranları küme 2'ye nazaran daha yüksek çıkmıştır. • Çalışan sayısı %50'ye yakın oranda 50-250 kişi arasındadır. • Gelir, %70 oranında 1.000.000 TL ve üzeridir. • Faaliyet sektörü küme 2'dedaha eşit dağılım gösterirken bu kümedeki işletmeler büyük oranda imalat sektöründe yer almaktadır. • %100 finans birimine sahiptir. • Arge birimine %45 oranında sahiptir. • Kalite birimine %90 oranında sahiptir. • Pazarlama ve üretim birimlerine sahip olma oranı %90'ın üstündedidir. • Otomatik üretim hattı %50 oranında vardır. • ISO9000-2000 belgesi % 66 oranında vardır. • Satınalma birimi %93 oranında vardır. • CNC tezgah oranı küme 2'yegöre yüzde 20 yüksektir. • CE, OHSAS 18001 ve ISO 14000 belgelerine sahiplik çok fazla olmamak kaydıyla küme 2'yegöre daha yüksektir.	• Çalışan sayısı, belirgin biçimde küme 1'den düşüktür. • Bu kümedeki işletmelerin ancak %19'u İK (İnsan Kaynakları) birimine sahiptir. • Finans birimine sahip olma oranı %50 düzeyindedir. • Gelir, %50 oranında 250.000 TL ve altındadır. • Arge birimi ancak %5 oranında mevcuttur. • Kalite birimi ancak %8 oranında mevcuttur. • Pazarlama birimine sahip olma oranı %55 düzeyindedir. • Üretim birimine sahip olma oranı %60 düzeyindedir. • Otomatik üretim hattına sahip olma ancak %12 düzeyindedir. • ISO9000-2000 belgesi % 18 oranında vardır. • Satınalma birimine sahiplik %23 düzeyindedir. • Geleneksel tezgah oranı ilk kümeye oranla %25 daha fazladır (%45). • Otomatik üretim hattına sahip olma yüzdesi ilk kümeye göre %20 daha düşüktür (%16)

İkinci aşamada, birliktelik kuralları analizi ile işletmelerde sıklıkla birlikte gözlemlenen durumlar (ör: belli bir birime sahip olma, gelir durumu vb.) tespit edilmeye çalışılmıştır. Bu amaçla Apriori algoritması kullanılmıştır. Ancak çıkacak olası yüzbinlerce kuraldan önemli ve sık oluşanları bulabilmek amacıyla minimum destek değeri %20, minimum güven değeri %50 olarak belirlenmiştir. Kurallar önem (importance) değerine göre sıralanmıştır. Önem değeri (literatürde lift olarak da bilinir) birliktelik kuralları analizinde güven değeri kadar önemli diğer bir parametredir ve tanım olarak kuralın öncülü ve sonculu arasındaki pozitif ya da negatif ilişkiyi ortaya koymaktadır. Microsoft Analysis Services paketi içersinde yer alan Apriori algoritmasından çıkan kurallarda önem değerinin 0'dan büyük olması kuralın öncülünün olması halinde sonculunun ne oranda gerçekleşeceğini, 0 olması kuralın herhangi bir değer taşımadığını, 0'dan düşük olması öncül ve soncul arasında negatif bir ilişki olduğunu göstermektedir (Tang ve MacLennan, 2005). Elde edilen kurallardan bir kısmı Tablo 3'de sunulmuştur.

Tablo 3. Birliktelik kuralları bulguları

Güven	Önem	Kural
0,625	0,5539	KALITEBIRIM = True, IKBIRIM = True ⇒ ERPVAR = True
1,000	0,4870	CALISANSAYI = 250'den fazla, ARGEBIRIM = True ⇒ ERPVAR= True
0,521	0,4532	IKBIRIM = True, SERVISISLETMESI = False -> ERPVAR = True
0,833	0,4376	CALISANSAYI = 250'den fazla, IKBIRIM = True -> ERPVAR = True
0,563	0,2657	CALISANSAYI = 50 – 249, KALITEBIRIM = True -> ERPVAR = True

Örnek olarak ikinci kural ele alınarak yorumlanmak istenirse çalışan sayısının 250'den fazla olduğu ve arge birimine sahip işletmelerin tamamı ERP kullanmaktadır. Sonuncu kuralın okunuşu ise "Çalışan sayısının 50-250 arasında olduğu ve kalite birimine sahip işletmelerde %56 güven değeri ile ERP yazılımı bulunmaktadır" şeklinde olacaktır.

5. SONUÇ VE TARTIŞMA

Bu çalışmada tanımlayıcı veri madenciliği yöntemlerinden kümeleme ve birliktelik kuralları kullanılarak, kurulumu maliyetli ve riskler taşıyan ancak sonrasında uzun dönemde önemli verimlilikler sağlayan ERP yazılımlarının işletmelerdeki etkileri, bu teknolojiye sahip olan ve olmayan işletmeler arasındaki farklılıklar ve sık gözlemlenen örüntüler ortaya çıkarılmaya çalışılmıştır. Kümeleme çalışması ile işletmelerin sahip oldukları özelliklere dayanılarak iki grup elde edilmiştir ve sonuçlar incelendiğinde şu an ERP sistemine sahip olmayan ancak mevcut nitelikleri bakımından bu sisteme sahip işletmelere yakın birçok işletme gözlemlenmiştir. Dolayısıyla bu işletmelerin ERP sistemine geçebilmek için gerekli altyapısal koşulları sağladıkları ve geçiş için aday olabilecekleri öngörülmektedir. Öte yandan bir kısım işletmenin ERP teknolojisine belirli bir olgunluk düzeyine erişmeden geçtiği ortaya çıkmaktadır. Ayrıca kapasite kullanım oranlarının yükselmesiyle ERP sahibi olma arasında korelasyon saptanmıştır.

Sonuç olarak gerek kümeleme gerekse de birliktelik kuralı analizi ile işletmelerin ERP yazılımına terfi etmeden önce karşılamaları gereken altyapısal (ör: birimler ve teknolojiler) unsurlar ortaya konmuştur. Ayrıca kullanılan yaklaşım ve yöntemler bu alanda yapılan çalışmalarda bu güne değin kullanılmamış olması nedeniyle alana katkı özelliği taşımaktadır.

6. KAYNAKLAR

Su Y. F. & Yang C., 2009. A structural equation model for analyzing the impact of ERP on SCM, Expert Systems with Applications, 37(1), 256-469.

Yavasoglu B., 2011. Kurumsal Kaynak Planlama Sisteminin Etkinliği ve Uygulama, Yüksek Lisans Tezi, Gazi Üniversitesi Bilişim Enstitüsü Yönetim Bilişim Sistemleri.

Cebeci, U., 2009. Fuzzy AHP-based decision support system for selecting ERP systems in textile industry by using balanced scorecard, Expert Systems with Applications, 36(5), 8900–8909.

Saatcioglu, O. Y., 2007. What determines user satisfaction in ERP projects: Benefits, barriers or risks?, Proceedings of European and Mediterranean Conference on Information Systems 2007 (EMCIS2007) S. 24–26.

Wu W. W., 2010. Segmenting and mining the ERP users' perceived benefits using the rough set approach, Expert Systems with Applications, doi:10.1016/j.eswa.

Jacobs F. R., Weston F. C., 2007. Enterprise resource planning (ERP)—A brief history, Journal of Operations Management, 25, 357-363.

Buonanno G., Faverio P., Pigni F., Ravarini A., 2009. Factors affecting ERP system adoption: A comparative analysis between SMEs and large companies, Journal of Enterprise Information, 18(4), 384-426.

Bozkir A. S., Gök B., Sezer E., 2008. İnternetin eğitimsel amaçlar için kullanımını etkileyen faktörlerin veri madenciliği ile tespiti, Bilimde Modern Yöntemler Sempozyumu, Eskişehir.

Tang Z. H. and MacLennan J., 2005. Data Mining with SQL Server 2005, Wiley Publishing.

Agrawal R., Imielinski T., Swami A., 1993. Mining association rules between sets of items in large databases. In Proceedings of the ACM SIGMOD international conference on management of data, 207–216.

INVESTIGATION OF THE DIFFERENCES BETWEEN COMPANIES WITH AND WITHOUT USAGE OF ERP SOFTWARE IN TERMS OF MANAGEMENT AND TECHNOLOGY VIA DESCRIPTIVE DATA MINING METHODS

ABSTRACT

ERP (Enterprise Resource Planning) software are the advanced solutions which tailored for whole enterprise needs and target high efficiency in production & service issues. In this study, the benefits of ERP upgrading in enterprises and the required infrastructural issues prior to this upgrading are revealed by employing the clustering and association rules, which are among the descriptive data mining methods. In clustering study, the natural differences are revealed between the companies which have ERP system and which do not. When the results are examined.İt is established that recent capacity usage ratios and owning some departments have high correlations with having ERP system. On the other hand, frequently observed patterns are extracted in companies with and without ERP by association rules analysis.

Keywords: Associaton rules analysis, ERP, Clustering, Data mining.

Volume: 09 Number: 01 Category: 01 Page: 29-36 ISSN No: 1303-6319 Cilt: 09 Sayı: 01 Kategori: 01 Sayfa: 29-36

İKİ DAĞILIŞ PARAMETRESİNİN EŞİTLİĞİ İÇİN DOĞRUSAL SIRA İSTATİSTİĞİNE DAYALI BİR TEST

Irmak ACARLAR* **Bülent ALTUNKAYNAK****

ÖZET

Bu çalışmada sıralı alternatifler için bir test istatistiği önerilmiştir. Bu test istatistiğinin dağılım özellikleri ve güç karşılaştırmaları yapılmıştır. Güç karşılaştırmalarında Klotz, Kamat ve Siegel-Tukey testleri kullanılmıştır. Özellikle $n=n_1+n_2$ tek sayı ve küçük iken önerilen test istatistiğinin, karşılaştırılan diğer üç testten daha güçlü olduğu kanıtlanmıştır.

Anahtar Kelimeler: Doğrusal sıra istatistikleri, Güç karşılaştırmaları, Kamat Testi, Siegel-Tukey Testi.

1. GİRİŞ

σ_1^2 ve σ_2^2 dağılış parametrelerinin bilinmediği iki yığından n_1 ve n_2 hacimli bağımsız örnekler $X_1,\ldots,X_{n_1}$ ve $Y_1,\ldots,Y_{n_2}$ olsun. Bu parametrelerin yansız tahmin edicileri S_1^2 ve S_2^2 olmak üzere, yığınların dağılımlarının normalliği altında "dağılış parametrelerinin eşitliği" hipotezinin testi için $S_1^2 / S_2^2 \sim F_{n_1-1,n_2-1}$ istatistiği kullanılır. Ancak, bu varsayımın sağlanmaması durumunda F istatistiği robust değildir (Gibbons ve Chakraborti, 2003). Bu nedenle, normallik varsayımı sağlanmıyorken iki dağılış parametresinin eşitliği hipotezinin testi için Doğrusal Sıra Sayıları İstatistiği formuna dayanan testler önerilmiştir. Bu forma dayanan testler için temel varsayım örneklerin geldikleri yığınların medyanlarının aynı olduğudur.

Doğrusal Sıra Sayıları İstatistiği formuna dayalı testlerden en iyi bilineni Siegel-Tukey testidir. Bu testin dezavantajı gözlem sayısı tek olduğunda bir gözlemi veriden çıkartmasıdır. Özellikle gözlem sayısının az olduğu durumlarda bu durum testin gücünün azalmasına neden olabilmektedir. Bu çalışmada, iki dağılış parametresinin eşitliği için sıra istatistiğine dayalı bir test istatistiği önerilmiştir. Bunun dışında Doğrusal Sıra İstatistiği formuna dayalı olan testlere Kamat testi, Mood Testi, Freund-Ansari-Bradley Testi, David-Barton Testi ve Klotz Normal Skorlar Testi örnek verilebilir (Gibbons ve Chakraborti, 2003). Konum parametrelerinin eşitliği varsayımını gerektirmeyen yöntemlerden bazıları Deshpande ve Kusum (1984) ve Maharajan vd. (2011) tarafından çalışılmıştır.

Çalışmanın ikinci bölümünde literatürde iki dağılış parametresinin eşitliği hipotezinin testi için önerilen parametrik olmayan testlerden Kamat testi, Siegel-Tukey testi ve Klotz normal skorlar testi tanıtılmıştır. Üçüncü bölümde bu parametrik olmayan testlere alternatif bir yöntem verilmiştir. Dördüncü bölümde, önerilen yöntemle ikinci bölümde tanıtılan yöntemler testin gücü bakımından simülasyonla karşılaştırılmıştır. Son olarak beşinci bölümde sonuçlar tartışılmıştır.

*Araş. Gör., Gazi Üniversitesi, Fen Fakültesi, İstatistik Bölümü, e-posta: irmakacarlar@gazi.edu.tr
**Doç. Dr., Gazi Üniversitesi, Fen Fakültesi, İstatistik Bölümü, Ankara, e-posta: bulenta@gazi.edu.tr

2. İKİ DAĞILIŞ PARAMETRESİNİN EŞİTLİĞİ İÇİN PARAMETRİK OLMAYAN BAZI TESTLER

Sıralı istatistiklere dayalı olan testlerden biri Kamat (1956) tarafından önerilmiştir. $n_1 \le n_2$ olmak üzere D_{n_1,n_2} ile gösterilen Kamat test istatistiği bir dağılış ölçüsü olan açıklığa dayalıdır. Kamat test istatistiği,

$$D_{n_1,n_2} = R_{n_1} - R_{n_2} + n_2 \tag{1}$$

ile verilir (Kamat, 1956). Burada R_{n_1} ve R_{n_2} sırasıyla birleştirilmiş örnekte X ve Y gözlemlerine atanan sıra sayılarının açıklıklarıdır. X gözlemlerine atanan sıra sayıları Y gözlemlerine atanan sıra sayılarından daha uç değerler alıyorsa bu test istatistiği Rosenbaum (1965) tarafından önerilen R istatistiği türünden $D_{n_1,n_2} = R + n_2$ biçimine dönüşür. Kamat test istatistiği için bazı kritik değerler (Kamat, 1956)'te mevcuttur.

Rasgele olmayan ağırlık katsayıları a_i ve rasgele değişken $Z_i \quad i = 1,...,n$ de

$$Z_i = \begin{cases} 1, & X \text{ ve } Y \text{ gözlemleri ortak sıralandığında } i.\text{sıradaki } X \text{ gözlemi ise} \\ 0, & \text{aksi halde} \end{cases} \tag{2}$$

olmak üzere, doğrusal sıra sayılarına dayalı istatistik

$$LR = \sum_{i=1}^{n} a_i Z_i \tag{3}$$

biçiminde tanımlanır (Gibbons ve Chakraborti, 2003). Burada $n = n_1 + n_2$ 'dir. a_i katsayılarının farklı seçimleri ile farklı test istatistikleri önerilmiştir. $[\bullet]$ işlemi tam değer fonksiyonunu ifade etmek üzere Siegel and Tukey (1960) tarafından önerilen test istatistiği

$$a_i = \begin{cases} 2i-1 & ,i \text{ tek ise, } 1 \le i \le [n/2] \\ 2i & ,i \text{ çift ise, } 1 < i \le [n/2] \\ 2(n-i)+1 & ,i \text{ tek ise, } [n/2] < i \le n \\ 2(n-i)+2 & ,i \text{ çift ise, } [n/2] < i < n \end{cases} \tag{4}$$

katsayılarına dayanır. Siegel-Tukey testinde n tek iken X ve Y gözlemleri ortak sıralandığında ortadaki gözlem veriden atılmakta ve yine bu ağırlık katsayıları kullanılmaktadır (Siegel ve Tukey, 1960). Bu durumda bir gözlem veriden atıldığı için veride bilgi kaybı oluşmaktadır.

Uygulamada sıkça kullanılan bir başka parametrik olmayan test Klotz normal-skorlar testidir. Klotz (1962) tarafından önerilen bu testteki temel düşünce ağırlık katsayılarının standart normal dağılımın ters dönüşüm fonksiyonuyla sürekli hale dönüştürülmesi ve

bunlardan yararlanarak doğrusal sıra istatistiğinin hesaplanmasıdır. $\Phi(\bullet)$ fonksiyonu standart normal dağılım için dağılım fonksiyonu olmak üzere test istatistiği,

$$K_n = \sum_{i=1}^{n} \left\{ \Phi^{-1}\left(\frac{i}{n+1} \right) \right\}^2 Z_i \tag{5}$$

ile verilir. $n \le 20$ iken Klotz normal skorlar testi için kritik değerler (Klotz, 1962)'te mevcuttur. Siegel-Tukey test istatistiğinde olduğu gibi örnek hacminin büyük değerleri için K_n istatistiği normal dağılıma yakınsar.

3. ÖNERİLEN TEST İSTATİSTİĞİ

Önerilen Test İstatistiği şu şekilde verilebilir. Olasılık yoğunluk fonksiyonları $f(x)$ ve $f(y)$ olan yığınlardan n_1 ve n_2 hacimli bağımsız örnekler $X_1, X_2, ..., X_{n_1}$ ve $Y_1, Y_2, ..., Y_{n_2}$ olsun. Bu iki yığının dağılış parametreleri, sırasıyla σ_1 ve σ_2 olsun. M_X ve M_Y eşit iken Z_i eşitlik (2) de verildiği gibi elde edilir. $[\bullet]$ işlemi tam değer fonksiyonunu ifade etmek üzere bu çalışmada önerilen testte ağırlık katsayıları

$$a_i = \begin{cases} n+2-2i-s+2sr-r & , i \text{ tek ise, } 1 \le i \le [n/2] \\ n+1-2i+s-2sr+r & , i \text{ çift ise, } 1 < i \le [n/2] \\ 2i-n-1+s & , i \text{ tek ise, } [n/2] < i \le n \\ 2i-n-s & , i \text{ çift ise, } [n/2] < i \le n \end{cases} \tag{6}$$

olarak elde edilir. Burada

$$r = \begin{cases} 0, & n \text{ tek ise} \\ 1, & n \text{ çift ise} \end{cases} \quad \text{ve} \quad s = \begin{cases} 0, & [n/2] \text{ tek ise} \\ 1, & [n/2] \text{ çift ise} \end{cases} \tag{7}$$

şeklindedir. Ayrıca n tek iken $a\left(\left[\frac{n}{2} \right] + 1 \right) = 1$ olarak tanımlanır. Bu durumda, $H_0 : \sigma_1 = \sigma_2$ hipotezinin testi için önerilen test istatistiği,

$$BI = \sum_{i=1}^{n} a_i Z_i \tag{8}$$

olarak verilebilir.

Bu çalışmada önerilen test istatistiği gözlemlerin tümünü kullanmakta, dolayısıyla *n* tek iken veriden bir gözlemin atılmasını önlemektedir. BI istatistiğinin dağılımı Siegel-Tukey istatistiğinin dağılımı ile aynıdır. Bu nedenle simülasyon çalışmasında BI testi için kritik değerlerin oluşturulmasında (Siegel ve Tukey, 1960)'deki olasılık dağılımları

kullanılmıştır. Büyük hacimli örnekler için BI istatistiği $(n_1 n_2 + n_1(n_1+1))/2$ ortalama ve $n_1 n_2 (n_1 + n_2 + 1)/12$ varyans ile normal dağılıma sahiptir.

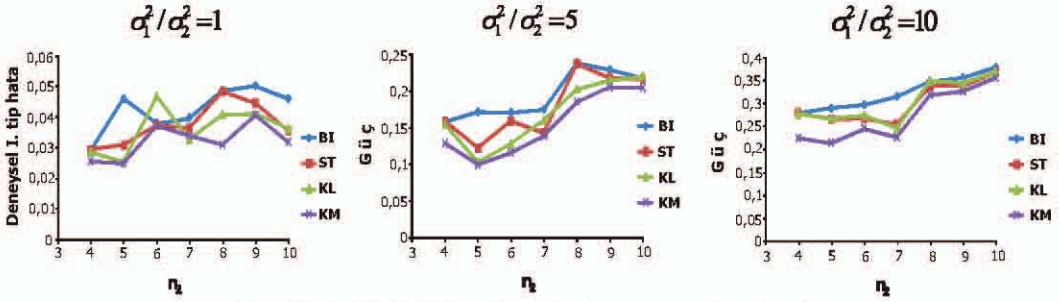
4. SİMÜLASYON ÇALIŞMASI

Bu bölümde deneysel birinci tip hataların ve varyanstaki farklılaşmaya göre güç değerlerinin belirlenmesine yönelik bir simülasyon çalışması yapılmıştır. Simülasyon çalışmasında önerilen test istatistiğinden elde edilen sonuçlar Siegel-Tukey, Kamat ve Klotz testlerinden elde edilen sonuçlar ile karşılaştırılmıştır. Karşılaştırmalarda üç farklı dağılım kullanılmıştır. Bunlar Tekdüze(0,1), Ki-kare(1) ve Üstel(1) dağılımlardır. Parantez içindeki ifadeler dağılıma ait parametreleri göstermektedir. Varyanstaki farklılaşma σ_1^2/σ_2^2 şeklinde ifade edilmiştir. $\sigma_1^2/\sigma_2^2 = 1$ ifadesi iki kitle varyansının eşit olduğu anlamına gelmektedir. Bu durumda yokluk hipotezinin reddedilme olasılığı deneysel I. tip hataya karşılık gelmektedir. $\sigma_1^2/\sigma_2^2 = 5$ ifadesi birinci kitle varyansının ikinci kitle varyansından 5 kat büyük olduğu anlamına gelirken, $\sigma_1^2/\sigma_2^2 = 10$ ifadesi de benzer şekilde 10 kat büyük olduğu anlamına gelmektedir. Dolayısıyla σ_1^2/σ_2^2 ifadesinin 5 ve 10'a eşit olduğu durumlarda yokluk hipotezinin reddedilme olasılığı testin gücüne karşılık gelmektedir. Her bir durumda yokluk hipotezinin reddedilme olasılıkları 50000 iterasyon için elde edilmiştir. Çalışmada, iki kitle için örnek çapları $n_1, n_2 = 4(1)10 \quad (n_1 \le n_2)$ olacak şekilde alınmıştır. Simülasyon çalışmasına ait bazı sonuçlar aşağıda verilmiştir.

Tablo 1. Tekdüze (0,1) dağılımı için simülasyon sonuçları

		σ_1^2/σ_2^2											
		1				5				10			
n_1	n_2	*BI*	*ST*	*KL*	*KM*	*BI*	*ST*	*KL*	*KM*	*BI*	*ST*	*KL*	*KM*
4	4	0.0293	0.0297	0.0288	0.0258	0.1563	0.1565	0.1530	0.1270	0.2785	0.2786	0.2771	0.2241
	5	0.0461	0.0311	0.0256	0.0251	0.1697	0.1210	0.1014	0.0982	0.2894	0.2656	0.2662	0.2138
	6	0.0382	0.0376	0.0470	0.0374	0.1692	0.1578	0.1267	0.1144	0.2971	0.2671	0.2743	0.2448
	7	0.0399	0.0367	0.0329	0.0341	0.1730	0.1414	0.1582	0.1370	0.3149	0.2543	0.2460	0.2263
	8	0.0488	0.0486	0.0411	0.0311	0.2368	0.2356	0.2006	0.1836	0.3471	0.3370	0.3482	0.3172
	9	0.0504	0.0448	0.0412	0.0409	0.2269	0.2161	0.2131	0.2032	0.3561	0.3369	0.3433	0.3252
	10	0.0462	0.0357	0.0363	0.0320	0.2156	0.2151	0.2182	0.2026	0.3778	0.3671	0.3679	0.3549
5	5	0.0307	0.0313	0.0501	0.0304	0.1639	0.1607	0.1550	0.1527	0.1680	0.1683	0.2369	0.1032
	6	0.0306	0.0330	0.0437	0.0188	0.1971	0.1815	0.1715	0.1623	0.2815	0.2971	0.2468	0.1720
	7	0.0480	0.0485	0.0434	0.0118	0.2414	0.2408	0.2060	0.1999	0.3810	0.3741	0.2769	0.1690
	8	0.0449	0.0457	0.0523	0.0433	0.2897	0.2776	0.2289	0.2153	0.4157	0.4479	0.3319	0.3196
	9	0.0414	0.0430	0.0531	0.0326	0.2991	0.2837	0.2648	0.2428	0.3910	0.3925	0.3430	0.2385
	10	0.0388	0.0394	0.0478	0.0232	0.3057	0.3013	0.3003	0.2856	0.4579	0.4388	0.3267	0.2266
6	6	0.0414	0.0411	0.0487	0.0103	0.2460	0.2450	0.2006	0.1726	0.3817	0.3807	0.3682	0.2468
	7	0.0342	0.0462	0.0480	0.0352	0.2223	0.2693	0.2262	0.1705	0.3568	0.4246	0.3657	0.2831
	8	0.0432	0.0431	0.0472	0.0234	0.2885	0.2920	0.2536	0.1997	0.4614	0.4668	0.3551	0.2735
	9	0.0487	0.0422	0.0466	0.0161	0.3191	0.3000	0.2729	0.2209	0.5023	0.4639	0.3848	0.2883
	10	0.0406	0.0419	0.0476	0.0421	0.3285	0.3281	0.3058	0.2854	0.5236	0.5219	0.3916	0.3288
7	7	0.0382	0.0398	0.0525	0.0213	0.2499	0.2520	0.2448	0.2021	0.3919	0.3950	0.3336	0.2491
	8	0.0397	0.0403	0.0473	0.0467	0.3081	0.3182	0.2955	0.2860	0.4847	0.4989	0.3421	0.3373
	9	0.0430	0.0415	0.0510	0.0327	0.3240	0.3234	0.3022	0.2889	0.5081	0.5078	0.3863	0.3843
	10	0.0437	0.0416	0.0512	0.0235	0.3701	0.3699	0.3466	0.2920	0.5747	0.5705	0.4012	0.3973
8	8	0.0496	0.0484	0.0500	0.0290	0.3721	0.3731	0.3540	0.3153	0.5545	0.5614	0.5160	0.5032
	9	0.0451	0.0461	0.0499	0.0367	0.3731	0.3761	0.3670	0.3512	0.5671	0.5649	0.5437	0.5043
	10	0.0420	0.0445	0.0494	0.0398	0.4025	0.4000	0.3898	0.3879	0.6180	0.6149	0.5990	0.5323
9	9	0.0400	0.0419	0.0519	0.0379	0.3709	0.3728	0.3937	0.3818	0.5713	0.5711	0.5481	0.5313
	10	0.0431	0.0425	0.0515	0.0234	0.4280	0.4378	0.4002	0.3613	0.6725	0.6477	0.5801	0.5768
10	10	0.0434	0.0427	0.0482	0.0428	0.4509	0.4505	0.4628	0.4466	0.6674	0.6667	0.6575	0.6205

Tablo 1'deki değerler σ_1^2/σ_2^2 oranının 1 olduğu durum için deneysel I. tip hatayı, bu oranın 5 ve 10 oldukları durumlar için ilgili değerler testin gücünü ifade etmektedir. Bu sonuçlar incelendiğinde özellikle n_1+n_2 değerinin tek sayı olduğu durumlarda önerilen test istatistiğinin karşılaştırılan diğer istatistiklere göre genelde daha iyi sonuçlar verdiği görülmektedir. $n_1=4$ durumu için elde edilen grafiklerden bu durum daha belirgin bir şekilde görülmektedir.

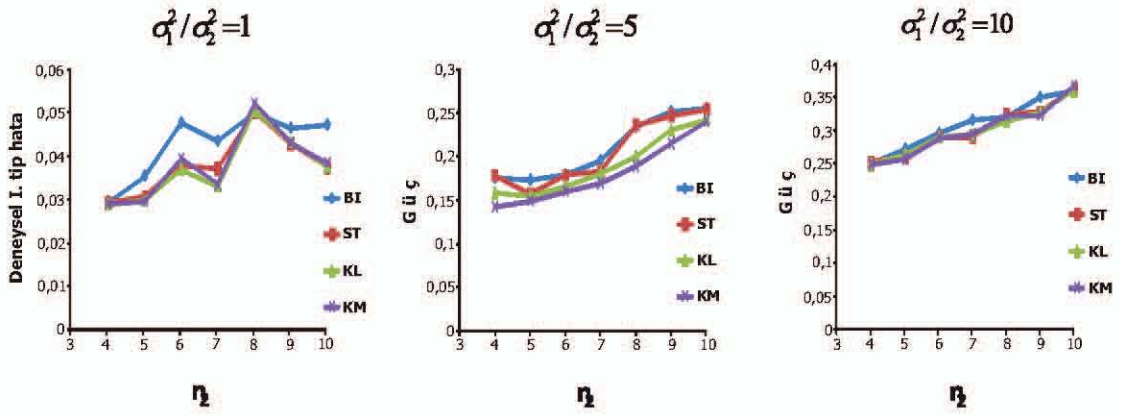


Şekil 1. Tekdüze (0,1) dağılımı için simülasyon sonuçları (n_1=4)

Tablo 2. Ki-kare (1) dağılımı için simülasyon sonuçları

		σ_1^2/σ_2^2											
		1				5				10			
n_1	n_2	*BI*	*ST*	*KL*	*KM*	*BI*	*ST*	*KL*	*KM*	*BI*	*ST*	*KL*	*KM*
4	4	0.0291	0.0291	0.0286	0.0287	0.1746	0.1772	0.1570	0.1409	0.2524	0.2548	0.2509	0.2506
	5	0.0350	0.0304	0.0294	0.0293	0.1715	0.1564	0.1533	0.1478	0.2740	0.2620	0.2661	0.2587
	6	0.0473	0.0375	0.0367	0.0391	0.1783	0.1779	0.1645	0.1590	0.2975	0.2920	0.2921	0.2905
	7	0.0432	0.0369	0.0327	0.0331	0.1945	0.1819	0.1800	0.1682	0.3177	0.2911	0.2961	0.2961
	8	0.0496	0.0499	0.0502	0.0518	0.2332	0.2350	0.1991	0.1879	0.3234	0.3265	0.3159	0.3237
	9	0.0461	0.0424	0.0429	0.0428	0.2510	0.2460	0.2293	0.2139	0.3523	0.3303	0.3298	0.3250
	10	0.0469	0.0370	0.0372	0.0382	0.2547	0.2530	0.2411	0.2388	0.3624	0.3638	0.3614	0.3690
5	5	0.0321	0.0321	0.0475	0.0319	0.1846	0.1893	0.1748	0.1616	0.2680	0.2679	0.2566	0.2544
	6	0.0302	0.0342	0.0452	0.0190	0.2035	0.1904	0.1790	0.1758	0.2849	0.2768	0.2702	0.2774
	7	0.0501	0.0492	0.0424	0.0122	0.2204	0.2202	0.2081	0.2049	0.2978	0.2977	0.2843	0.2885
	8	0.0443	0.0492	0.0513	0.0451	0.2403	0.2231	0.2200	0.2198	0.3247	0.3038	0.3039	0.3057
	9	0.0424	0.0443	0.0524	0.0314	0.2562	0.2522	0.2484	0.2452	0.3278	0.3256	0.3204	0.3106
	10	0.0401	0.0397	0.0505	0.0246	0.2892	0.2723	0.2755	0.2717	0.3419	0.3375	0.3321	0.3358
6	6	0.0415	0.0407	0.0469	0.0100	0.1982	0.1983	0.1846	0.1811	0.3088	0.3064	0.3011	0.2915
	7	0.0374	0.0448	0.0476	0.0355	0.2164	0.2004	0.1940	0.1948	0.3301	0.3221	0.3159	0.3181
	8	0.0423	0.0439	0.0458	0.0224	0.2346	0.2405	0.2181	0.2192	0.3462	0.3462	0.3401	0.3301
	9	0.0482	0.0416	0.0471	0.0162	0.2637	0.2596	0.2442	0.2390	0.3764	0.3663	0.3631	0.3573
	10	0.0442	0.0425	0.0480	0.0407	0.3044	0.3047	0.2939	0.2890	0.4013	0.4030	0.4048	0.3957
7	7	0.0372	0.0384	0.0517	0.0200	0.2332	0.2395	0.2275	0.2272	0.3431	0.3404	0.3398	0.3320
	8	0.0392	0.0393	0.0485	0.0480	0.2583	0.2466	0.2471	0.2424	0.3822	0.3695	0.3717	0.3742
	9	0.0413	0.0422	0.0507	0.0329	0.2914	0.2917	0.2850	0.2784	0.4163	0.4156	0.4129	0.4185
	10	0.0444	0.0421	0.0516	0.0234	0.3433	0.3325	0.3201	0.3201	0.4574	0.4307	0.4480	0.4352
8	8	0.0494	0.0502	0.0498	0.0290	0.3081	0.3084	0.3038	0.3014	0.4255	0.4278	0.4311	0.4250
	9	0.0473	0.0463	0.0502	0.0370	0.3503	0.3313	0.3302	0.3263	0.4596	0.4499	0.4459	0.4421
	10	0.0442	0.0424	0.0509	0.0391	0.3632	0.3622	0.3628	0.3645	0.4887	0.4988	0.4866	0.4888
9	9	0.0391	0.0415	0.0510	0.0366	0.3610	0.3621	0.3696	0.3568	0.4908	0.4847	0.5071	0.4910
	10	0.0450	0.0436	0.0505	0.0243	0.3983	0.3792	0.3719	0.3751	0.5670	0.5443	0.5514	0.5535
10	10	0.0448	0.0457	0.0475	0.0435	0.4092	0.4093	0.4019	0.4004	0.5889	0.5872	0.5896	0.5813

Tablo 2'deki farklı oranlar için deneysel I. tip hata ve testin gücüne ilişkin sonuçlar incelendiğinde özellikle n_1+n_2 değerinin tek sayı olduğu durumlarda önerilen test istatistiğinin karşılaştırılan diğer istatistiklere göre genelde daha iyi sonuçlar verdiği görülmektedir. Örneğin, $\sigma_1^2/\sigma_2^2=10$, $n_1=4$ ve $n_2=9$ durumu için önerilen istatistiğe ait güç değeri 0.35 iken diğer testlerde 0.33 veya altındadır. $n_1=4$ durumu için elde edilen grafikler aşağıda verilmiştir.

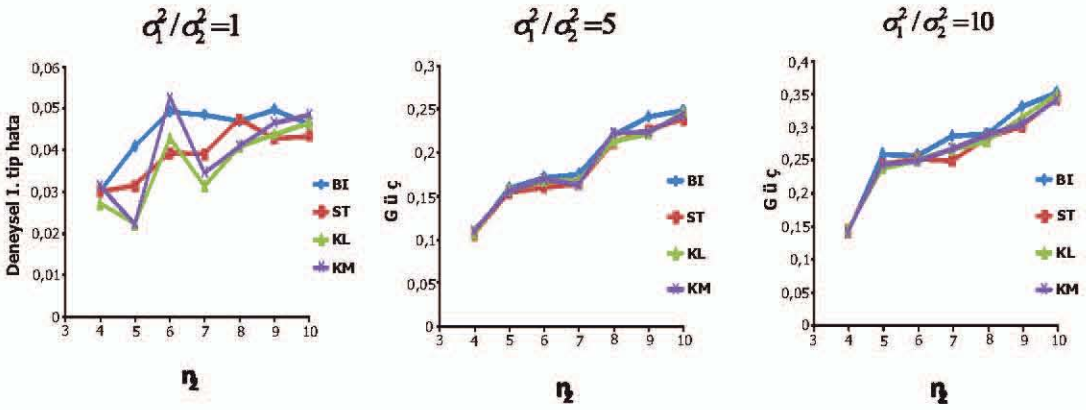


Şekil 2. Ki-kare (1) dağılımı için simülasyon sonuçları (n_1=4)

Tablo 3. Üstel (1) dağılımı için simülasyon sonuçları

		σ_1^2/σ_2^2 = 1				5				10			
n_1	n_2	BI	ST	KL	KM	BI	ST	KL	KM	BI	ST	KL	KM
4	4	0,0301	0,0301	0,0270	0,0313	0,1056	0,1056	0,1070	0,1092	0,1415	0,1445	0,1456	0,1422
	5	0,0407	0,0314	0,0221	0,0222	0,1581	0,1535	0,1565	0,1549	0,2598	0,2452	0,2400	0,2455
	6	0,0492	0,0392	0,0427	0,0525	0,1703	0,1603	0,1675	0,1692	0,2591	0,2541	0,2519	0,2508
	7	0,0485	0,0390	0,0313	0,0345	0,1745	0,1640	0,1672	0,1617	0,2876	0,2513	0,2683	0,2690
	8	0,0470	0,0473	0,0407	0,0409	0,2196	0,2108	0,2129	0,2219	0,2916	0,2843	0,2803	0,2898
	9	0,0497	0,0428	0,0437	0,0465	0,2404	0,2258	0,2211	0,2232	0,3317	0,3030	0,3168	0,3064
	10	0,0463	0,0433	0,0465	0,0485	0,2481	0,2384	0,2460	0,2428	0,3542	0,3442	0,3497	0,3425
5	5	0,0313	0,0317	0,0501	0,0314	0,1629	0,1633	0,1661	0,1663	0,2397	0,2591	0,2427	0,2444
	6	0,0495	0,0349	0,0437	0,0197	0,1855	0,1767	0,1695	0,1666	0,2924	0,2624	0,2746	0,2688
	7	0,0470	0,0473	0,0434	0,0132	0,2149	0,2131	0,2041	0,1901	0,3020	0,3196	0,3251	0,3165
	8	0,0455	0,0479	0,0523	0,0419	0,2586	0,2487	0,2451	0,2259	0,3554	0,3386	0,3355	0,3399
	9	0,0407	0,0420	0,0531	0,0329	0,2592	0,2570	0,2542	0,2595	0,3664	0,3677	0,3786	0,3804
	10	0,0391	0,0406	0,0478	0,0240	0,2865	0,2789	0,2676	0,2691	0,3800	0,3666	0,3754	0,3647
6	6	0,0419	0,0396	0,0487	0,0110	0,2178	0,2194	0,2248	0,2162	0,3044	0,3025	0,3122	0,3090
	7	0,0345	0,0449	0,0480	0,0349	0,2340	0,2145	0,2230	0,2221	0,3614	0,3410	0,3551	0,3434
	8	0,0412	0,0428	0,0472	0,0231	0,2538	0,2550	0,2537	0,2513	0,3770	0,3823	0,3890	0,3813
	9	0,0509	0,0424	0,0466	0,0151	0,2938	0,2880	0,2857	0,2854	0,4018	0,3803	0,4062	0,3960
	10	0,0425	0,0415	0,0476	0,0429	0,3154	0,3252	0,3187	0,3122	0,4298	0,4206	0,4240	0,4255
7	7	0,0380	0,0384	0,0505	0,0209	0,2552	0,2579	0,2698	0,2507	0,3760	0,3708	0,3681	0,3871
	8	0,0404	0,0399	0,0479	0,0493	0,3170	0,2808	0,2997	0,3064	0,4033	0,3890	0,3983	0,4052
	9	0,0419	0,0418	0,0480	0,0312	0,3278	0,3273	0,3136	0,3262	0,4230	0,4259	0,4380	0,4208
	10	0,0437	0,0427	0,0509	0,0239	0,3501	0,3358	0,3468	0,3439	0,4503	0,4350	0,4494	0,4357
8	8	0,0505	0,0504	0,0496	0,0288	0,3255	0,3208	0,3266	0,3112	0,4208	0,4263	0,4354	0,4336
	9	0,0462	0,0460	0,0480	0,0382	0,3693	0,3405	0,3409	0,3359	0,4696	0,4426	0,4581	0,4547
	10	0,0439	0,0441	0,0486	0,0384	0,3895	0,3876	0,3924	0,3874	0,5054	0,5106	0,5235	0,5104
9	9	0,0408	0,0413	0,0511	0,0364	0,3917	0,3937	0,4013	0,3903	0,5040	0,5101	0,5241	0,5346
	10	0,0431	0,0415	0,0479	0,0249	0,4294	0,4192	0,4257	0,4104	0,5446	0,5245	0,5348	0,5368
10	10	0,0430	0,0433	0,0459	0,0443	0,4377	0,4373	0,4410	0,4353	0,5782	0,5799	0,5806	0,5921

Tablo 3'de verilen deneysel I. tip hata ve testin gücüne ilişkin sonuçlarda Tablo 1 ve 2 deki sonuçlara paralellik göstermektedir. Benzer şekilde $n_1 = 4$ durumu için elde edilen grafikler aşağıda verilmiştir.



Şekil 3. Üstel (1) dağılımı için simülasyon sonuçları (n_1=4)

5. SONUÇ

Simülasyon ile elde edilen sonuçlardan ve grafiklerden önerilen test istatistiğinin iyi bilinen Siegel-Tukey, Kamat ve Klotz testleri kadar iyi sonuçlar verdiği görülmüştür. Deneysel I. tip hata ve testin gücü bakımından önerilen test ile Siegel-Tukey testlerinin karşılaştırılmasından elde edilen sonuçlarda özellikle $n=n_1+n_2$ sayısı tek iken önerilen test istatistiğinin Siegel-Tukey test istatistiğine göre daha iyi sonuçlar verdiği görülmektedir. Bu durum, gözlem sayısı tek iken Siegel-Tukey testinin bir gözlemi veriden çıkartmasından kaynaklanmıştır. Gözlem sayısı arttıkça tüm testler için elde edilen güç değerlerinin birbirine yaklaştığı ve arttığı görülmüştür. Güç değerlerindeki artış özellikle standart normal dağılımı kullanan Klotz testinde daha belirgin olarak ortaya çıkmıştır.

6. KAYNAKLAR

Deshpande, J. V., Kusum, K., 1984. A Test for the Nonparametric Two-Sample Scale Problem. Australian Journal of Statistics, 26, 16-24.

Gibbons, J. D., Chakraborti, S., 2003. Nonparametric Statistical Inference 4th edn. Marcel Dekker, New York.

Kamat, A. R., 1956. A Two Sample Distribution Free Test. Biometrika, 43, 377-387.

Klotz, J., 1962. Nonparametric Test for Scale. Ann. Math. Statist, 33, 498-512.

Mahajan, K.K., Gaur, A., Arora, S., 2011. A Nonparametric Test for a Two-Sample Scale Problem Based on Subsample Medians. Statistical Probability Letters, doi: 10.1016. Article in press.

Rosenbaum, S., 1965. On Some Two-Sample Non-parametric Tests. Journal of the American Statistical Association, 60, 1118–1126.

Siegel, S., and J. W. Tukey, 1960. A Nonparametric Sum of Ranks Procedure for Relative Spread in Unpaired Samples. Journal of the American Statistical Association, 55, 429–445.

A TEST BASED ON LINEAR RANK STATISTICS FOR THE EQUALITY OF TWO DISPERSION PARAMETERS

ABSTRACT

This paper suggests a test statistic for ordered alternatives. Distributional properties of this test statistic were examined, and power comparisons were made. Klotz, Kamat and Siegel-Tukey tests were used in power comparisons. Especially when $n=n_1+n_2$ is odd number and small, the suggested statistic was proven to be more powerful in comparison with the other three test statistics.

Keywords: Linear rank statistics, Power comparisons, Kamat Test, Siegel-Tukey Test.

Volume: 09 Number: 01 Category: 01 Page: 37-44 ISSN No: 1303-6319 Cilt: 09 Sayı: 01 Kategori: 01 Sayfa: 37-44

FARLIE-GUMBEL-MORGENSTERN KAPULA AİLESİ, BAZI GENİŞLETMELERİ VE BİR UYGULAMA

Irmak ACARLAR* **Harun KINACI****

ÖZET

Son yirmi yılda iki ya da daha fazla değişken arasındaki bağımlılık yapısının incelenmesinde kapulalar oldukça önem kazanmıştır. Bu yüzden kapulaların matematiksel ve istatistiksel özellikleri üzerine yapılan araştırmaların sayısı gün geçtikçe artmaktadır ve uygulama alanı da aynı doğrultuda genişlemektedir. Bu çalışmada Huang ve Kotz (1999) ve Lai ve Xie (2000) tarafından önerilen FGM ailesinin bazı genişletmeleri Türk Lirası bazında Amerikan Doları ve altın fiyatları üzerinden incelenmiştir.

Anahtar Kelimeler: FGM kapula ailesi, İki değişkenli dağılım, Pozitif bölge bağımlılığı.

1. GİRİŞ

Yirminci yüz yılın başlarından itibaren rasgele değişkenlere ilişkin çok değişkenli dağılım fonksiyonu ile bu değişkenlerin tek değişkenli marjinalleri arasındaki ilişki istatistikçiler tarafından oldukça ilgi duyulan bir konu olmuştur. Bu konu ile ilgili Frechet (1951) ve Dall'Aglio (1959) sabit marjinallere sahip ve bağımlı olan rasgele değişkenlerin iki ve üç değişkenli dağılım fonksiyonları üzerine çalışmalar yapmışlardır. Bağımlı rasgele değişkenlerin ortak dağılım fonksiyonuna ilişkin Sklar (1959) bağımlılık yapısını destekleyecek bir teorem geliştirmiştir ve bağımlılık fonksiyonlarına ilişkin çalışmaların çoğu bu teoreme dayalı olarak yapılmıştır. Kapula (copula) olarak tanımlanan bu fonksiyonlar için genel bir tanım, çok değişkenli dağılım fonksiyonlarını bir değişkenli marjinal dağılım fonksiyonlarına bağlayan fonksiyonlar olarak verilebilir (Çelebioğlu, 2007). Kapula "bağ" anlamına gelen Latince kökenli *copulare* kelimesinden türemiştir.

$I^2 = [0,1]^2$ birim karesinde tanımlı iki değişkenli bir C kapulası marjinalleri standart tek düze olan iki değişkenli birikimli bir dağılım fonksiyonudur ve (Nelsen, 2006) de verilen üç özelliği sağlar.

$\forall x, y \in \Re$ için iki değişkenli dağılım fonksiyonu H ve marjinalleri F ve G ile tanımlansın. C bir kapulayı ifade etmek üzere $\forall x, y \in \Re$ için $H(x,y) = C(F(x), G(y))$ eşitliği sağlanır (Nelsen, 2006; Sklar, 1959). Sklar (1959) tarafından verilen bu eşitlik çok değişkenli dağılım fonksiyonunun kapulalar yardımıyla oluşturulmasında temel teşkil etmektedir.

*Araş. Gör., Gazi Üniversitesi, Fen Fakültesi, İstatistik Bölümü, Ankara, e-posta: irmakacarlar@gazi.edu.tr

**Araş. Gör., Gazi Üniversitesi, Fen Fakültesi, İstatistik Bölümü, Ankara, e-posta: hkinaci@gazi.edu.tr

Ortak dağılım fonksiyonu H ve marjinaller sırasıyla F ve G olmak üzere $\forall x, y \in \Re$ için $H(x,y) - F(x)G(y) \geq 0$ ise X ve Y rasgele değişkenleri *pozitif bölge bağımlılığı*'na (positive quadrant dependence, PQD) sahiptir. Bu durumda X ve Y rasgele değişkenlerinin C kapulası $\forall u, v \in [0,1]$ olmak üzere $C(u,v) \geq uv$ özelliğini sağlar (Nelsen, 2006).

Literatürde çok çalışılan bir kapula olan Farlie-Gumbel-Morgenstern (FGM) kapula ailesi ikinci dereceden bir kapuladır. FGM kapulası ilk kez Eyraud (1938) tarafından önerilmiştir. Daha sonra Morgenstern (1956) bu aileyi ele alarak Cauchy marjinalleri için incelemiş, Gumbel (1960) benzer bir şekilde üstel marjinaller üzerine çalışmış ve Farlie (1960) de bu ailenin genel bir formunu oluşturmuştur (Nelsen, 2006). Bazı kaynaklarda bu kapula ailesi Eyraud-Farlie-Gumbel-Morgenstern olarak da isimlendirilmiştir. İkinci dereceden bir kapula olan FGM kapula ailesi,

$$C_\theta^{FGM}(u,v) = uv + \theta uv(1-u)(1-v) \tag{1}$$

ile verilir ve C_θ^{FGM}, $\theta \in [-1,1]$ olduğu sürece 2-artandır (Nelsen, 2006). u ve v yerine sırasıyla X ve Y rasgele değişkenlerinin marjinalleri olan $F(x)$ ve $G(y)$ alınırsa FGM kapulası ile oluşturulan ortak dağılım fonksiyonu

$$H_\theta(x,y) = F(x)G(y) + \theta F(x)G(y)[1-F(x)][1-G(y)] \quad -1 \leq \theta \leq 1 \tag{2}$$

şeklinde yazılabilir. Birliktelik parametresi olarak da adlandırılan θ için uygun bir değer marjinallere ilişkin korelasyon katsayısından elde edilebilir. Literatürde θ'nın tahminleri için genellikle korelasyon katsayısının sağlam alternatifleri olan ve birliktelik ölçüsü olarak da adlandırılan Spearman'ın Rho'su ve Kendall'ın Tau'sunun kullanılması önerilmektedir. Bu iki birliktelik ölçüsü için formülasyonlar sırasıyla,

$$\rho_S = 12 \int_0^1 \int_0^1 C(u,v)\,du\,dv - 3 \tag{3}$$

ve

$$\tau = 4 \int_0^1 \int_0^1 C(u,v)\,dC(u,v) - 1 \tag{4}$$

eşitlikleri ile verilir (Nelsen, 2006). FGM kapula ailesi için birliktelik ölçülerinden Spearman'ın Rho'su $\rho_S \in [-1/3, 1/3]$ kapalı aralığında değer aldığı için Eşitlik (1) zayıf bir bağımlılığın modellenmesine izin vermektedir. Bundan dolayı daha güçlü bir bağımlılığı modellemek amacıyla FGM kapula ailesinin genişletmeleri üzerine çalışmalar yapılmıştır.

Bu çalışmanın ikinci bölümünde FGM kapula ailesinin iki genişletmesi olan Huang ve Kotz ailesi ve Lai ve Xie ailesi tanıtılmıştır. Üçüncü bölümde FGM kapula ailesi, Huang ve Kotz ailesi ve Lai ve Xie ailesi'ne ilişkin bir uygulamaya yer verilmiştir. Dördüncü bölümde ise bulunan sonuçlar tartışılmıştır.

2. FGM KAPULA AİLESİNİN BAZI GENİŞLETMELERİ

FGM ailesinin genişletmeleri daha yüksek pozitif bağımlılığın modellenmesini sağladığı için oldukça çalışılan bir konudur. Bu aile için pozitif bölge bağımlılığı (1) eşitliğinde sağ tarafta yer alan

$$w(u,v) = \theta uv(1-u)(1-v) \geq 0 \tag{5}$$

olması durumunda söz konusudur. FGM ailesinin genişletmeleri $w(u,v)$ fonksiyonu üzerinde yapılan değişikliklerden hareketle elde edilir. Bu bölümde FGM ailesinin bazı genişletmelerinden Huang ve Kotz (1999) tarafından önerilen iki kapula ve Lai ve Xie (2000) tarafından önerilen kapula tanıtılmıştır.

2.1. FGM Kapula Ailesinin Huang ve Kotz Genişletmeleri

Daha güçlü pozitif bölge bağımlılığını elde etmek amacıyla Huang ve Kotz (1999) FGM kapulasının çekirdek (kernel) genişletmeleri üzerine çalışmışlardır. Önerdikleri iki aileden biri $(1-u)^{\gamma}$ türünden çekirdek genişletmeye dayalıdır. Bu aile,

$$C_{\theta,\gamma}^{HK1}(u,v) = uv\left(1+\theta(1-u)^{\gamma}(1-v)^{\gamma}\right) \qquad \gamma>0, \qquad 0<u,v<1 \tag{6}$$

biçimindedir[Huang ve Kotz, 1999]. Bu genişletme için ortak olasılık fonksiyonu,

$$h_{\theta}(u,v) = 1+\theta(1-u)^{\gamma-1}(1-v)^{\gamma-1}\left[1-(1+\gamma)u\right]\left[1-(1+\gamma)v\right] \tag{7}$$

ile verilir. Burada θ parametresi için değer aralığı $-1 \leq \theta \leq \left[(\gamma+1)/(\gamma-1)\right]^{\gamma-1}$ ile tanımlanır. $C_{\theta,\gamma}^{HK1}$ için tekdüze marjinal dağılımlar arasındaki korelasyon katsayısı,

$$\rho = 12\left(\frac{1}{(\gamma+2)(\gamma+1)}\right)^{2}\theta \tag{8}$$

ile verilmiştir[Huang ve Kotz, 1999].

FGM kapula ailesinin Huang ve Kotz (1999) tarafından önerilen diğer bir genişletmesi $1-u^{\gamma}$ tipi çekirdek genişletmedir. Bu kapula,

$$C_{\theta,\gamma}^{HK2}(u,v) = uv\left[1+\theta\left(1-u^{\gamma}\right)\left(1-v^{\gamma}\right)\right] \qquad \gamma>0,\ 0<u,v<1 \tag{9}$$

ile verilir[Huang ve Kotz, 1999]. Burada da θ parametresi değer aralığı $-\{\max(1,\gamma)\}^{-2} \leq \theta \leq \gamma^{-1}$ ile tanımlanır. Bu kapula için birliktelik ölçüsü ise,

$$\rho = 3\left(\frac{\gamma}{\gamma+2}\right)^2 \theta \qquad (10)$$

ile verilmiştir[Huang ve Kotz, 1999]. Bu iki aile ile ilgili özellikler Huang ve Kotz (1999)' da mevcuttur.

2.2. FGM Kapula Ailesinin Lai ve Xie Genişletmesi

Literatürde yer alan FGM kapula ailesinin genişletmelerinden bir diğeri Lai ve Xie (2000) tarafından önerilmiştir. Bu kapula,

$$C^{LX}_{\theta,p,q}(u,v) = uv + \theta u^p v^p (1-u)^q (1-v)^q \qquad p \geq 1,\ q \geq 1,\ 0 \leq \theta \leq 1 \qquad (11)$$

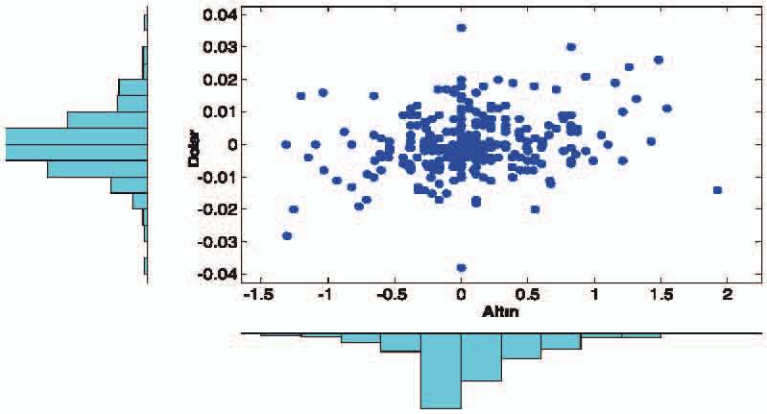
biçimindedir. Lai ve Xie ailesine ilişkin korelasyon katsayısı $\rho = 12\theta[\mathrm{B}(p+1, q+1)]^2$ ile verilir (Lai ve Xie, 2000). Burada $B(a,b)$ beta fonksiyonudur.

3. UYGULAMA

Uygulama kısmında FGM kapulası ile FGM kapulasının genişletmeleri bir veri seti üzerinden ki-kare istatistiği yardımıyla karşılaştırılmıştır. Bu karşılaştırmada ki-kare istatistiğinde yer alan beklenen frekansların hesaplanması için bunlara karşılık gelen ortak olasılıklar, ele alınan kapulalarla belirlenmiştir. Uygulama için 01.01.2010 ve 01.01.2011 tarihleri arasındaki TL bazında altın ve dolar fiyatları verisi ele alınmıştır (www.altinkaynak.com, erişim tarihi:12.01.2011). TL bazında altın fiyatındaki değişim miktarları ve dolar fiyatında değişim miktarları arasındaki bağımlılık yapısı kapula yöntemi ile incelenmiştir.

Kapula tahmin yöntemleri parametrik, yarı parametrik ve parametrik olmayan yöntemler olmak üzere üç başlık altında toplanır. Parametrik ve yarı parametrik yöntemler için ilgili değişkenlerin dağılımlarının biliniyor olması gereklidir. Fakat parametrik olmayan yöntemler dağılıma bağlı olmayan yöntemler olduğundan ilgili kapulanın parametrelerinin tahmininde marjinallere ilişkin bir bilgiye ihtiyaç duyulmaz.

Bu çalışmada ele alınan kapulaların parametrelerinin tahmini için Spearman ρ_S değerine dayalı olan parametrik olmayan yöntem kullanılmıştır (Çelebioğlu, 2007). Bu yöntemin kullanılmasının nedeni hem bu yöntemin mevcut yöntemlere nazaran hesaplama kolaylığı sağlaması hem de dağılıma bağlı olmayan bir yöntem olduğundan marjinal dağılımlara ihtiyaç duyulmamasıdır (Çelebioğlu, 2007). 01.01.2010 ve 01.01.2011 tarihleri arasındaki TL bazında altın ve dolar fiyatlarındaki değişim miktarları arasındaki bağımlılık yapısı incelenirken önce bu iki dizi arasındaki serpme diyagramı oluşturulmuş ve serpme diyagramı Şekil 1'de verilmiştir.



Şekil 1. TL bazında altın ve dolar fiyatlarındaki değişim miktarlarına ilişkin serpme diyagramı

Şekil 1'den de görüldüğü üzere altın ve dolar fiyatları arasındaki ilişki aynı yönlü fakat güçlü değildir. Bu iki değişken arasındaki birliktelik ölçüsü de $\hat{\rho}_s = 0.22$ olarak elde edilmiştir. Değişkenler arasındaki bağımlılık yapısının kapulalarla modellenmesinde doğru kapulanın seçilmesi oldukça önemlidir. $\hat{\rho}_s$'nin bu değeri için belirlenen iki değişken arasındaki bağımlılık yapısının incelenmesinde FGM kapula ailesinden yararlanılabilir. Bu uygulamada ele alınan altın ve dolar değişkenlerine ilişkin birliktelik ölçüsü tahmini $\hat{\rho}_s$, FGM kapula ailesine ilişkin birliktelik ölçüsü olan ρ_s'nin $[-1/3, 1/3]$ aralığında olmasından dolayı FGM kapula ailesi kullanılabilir. X ve Y rasgele değişkenlerine uygulanan monoton artan dönüşümler C kapulasının değerini değiştirmemektedir (Nelsen, 2006). Buradan hareketle genelliği bozmaksızın ele alınan iki diziye sıra sayıları atanmış ve bu sıra sayılarına dayalı olarak sınıf sayısı 4 olacak biçimde sınıflama yapılmıştır. Sınıflama yapılan veri için gözlenen frekanslar Tablo1'deki gibidir.

Tablo 1. TL bazında dolar ve altın fiyatlarındaki değişim miktarlarına ilişkin gözlenen frekanslar

		TL bazında Dolar fiyatlarındaki değişim				
	Değişkenler	$Y_{(74)}$	$Y_{(148)}$	$Y_{(222)}$	$Y_{(295)}$	TOPLAM
TL bazında Altın fiyatlarındaki değişim	$X_{(74)}$	26	17	16	14	73
	$X_{(148)}$	23	24	18	16	81
	$X_{(222)}$	13	18	18	18	67
	$X_{(295)}$	17	12	19	26	74
	TOPLAM	79	71	71	74	295

Tablo 2, Tablo 3 ve Tablo 4'te FGM kapulası, Huang ve Kotz (1999) tarafından önerilen iki kapula ve Lai ve Xie (2000) tarafından önerilen kapulaların farklı durumları için $\hat{\theta}$ tahminleri, beklenen frekanslar ve χ_h^2 değerleri verilmiştir.

Tablo 2. $C_\theta^{FGM}, \gamma = 2$ **için** C_θ^{HK1} **ve** C_θ^{HK2} **,** ***p=q*=2 için** C_θ^{LX} **kapulaları ile elde edilen** $\hat{\theta}$, χ_h^2 **ve beklenen frekanslar**

Kapula	$\hat{\theta}$	Beklenen Frekanslar	χ^2
C_θ^{FGM}	0,6612	27 20 15 11 24 20 19 18 16 15 17 19 13 16 20 25	5,4456
C_θ^{HK1}, $\gamma = 2$ için	2,6449	35 15 9 13 19 20 21 21 10 17 20 19 15 19 21 20	16,5435
C_θ^{HK2}, $\gamma = 2$ için	0,2939	25 19 17 12 25 21 19 16 17 16 16 18 13 14 19 20	4,9514
C_θ^{LX}, p=q=2 için	16,5308	26 22 13 12 27 23 16 16 13 13 20 21 13 14 22 25	7,4185

Tablo 3. $\gamma = 3$ **için** C_θ^{HK1} **ve** C_θ^{HK2} **,** ***p=q*=3 için** C_θ^{LX} **kapulaları ile elde edilen** $\hat{\theta}$, χ_h^2 **ve beklenen frekanslar**

Kapula	$\hat{\theta}$	Beklenen Frekanslar	χ^2
C_θ^{HK1}, $\gamma = 3$ için	7.3470	44 7 6 16 11 24 25 22 8 21 21 18 17 19 19 19	61,1898
C_θ^{HK2}, $\gamma = 3$ için	0.2041	23 20 17 13 25 22 19 15 18 16 16 17 13 13 18 30	5,5381
C_θ^{LX}, p=q=3 için	360.0046	25 23 11 14 29 27 11 14 11 9 25 23 15 12 24 23	25,9843

Tablo4. $\gamma = 1.5$ **için** C_θ^{HK1} **ve** C_θ^{HK2} **,** ***p=q*=1.5 için** C_θ^{LX} **kapulaları ile elde edilen** $\hat{\theta}$, χ_h^2 **ve beklenen frekanslar**

Kapula	$\hat{\theta}$	Beklenen Frekanslar	χ^2
C_θ^{HK1}, $\gamma = 1,5$ için	1.4063	31 18 12 12 22 20 19 20 12 16 19 20 13 18 21 22	8,3600
C_θ^{HK2}, $\gamma = 1,5$ için	0.4000	25 19 16 12 24 21 19 17 16 15 17 19 13 15 20 27	5,0232
C_θ^{LX}, p=q=1,5 için	3.3879	27 21 14 12 26 21 18 27 14 14 18 20 13 15 21 25	5,4249

Genelde en uygun kapulanın belirlenmesi işlemi hesaplanan χ^2 değerlerinden en küçüğüne karşılık gelen kapula ailesi seçilerek sonuçlandırılır (Çelebioğlu, 2007). Bu

bilgi doğrultusunda χ^2 kritik değerinin $\chi^2_{9,0.99} = 21{,}6660$ olduğu durumda uygunluğu belirlenen kapulalar arasında en küçük χ^2_h değerine karşılık gelen kapula $\gamma = 2$ iken Huang ve Kotz (1999) tarafından önerilen ikinci kapula olan C_θ^{HK2}' dir.

4. SONUÇ

Bu çalışmada ilk olarak FGM kapula ailesi ve Huang ve Kotz aileleri ile Lai ve Xie ailesi kısaca tanıtılmıştır. Sonra 01.01.2010 - 01.01.2011 tarihleri arasında altın ve dolar fiyatlarındaki değişimler alınarak bu iki değişken arasındaki bağımlılık yapısı incelenmiştir. Sonra farklı kapulalarla elde edilen beklenen değerlerden hesaplanan χ^2 değeriyle bunlara ilişkin kritik değer karşılaştırılmış ve bu veri seti için uygun kapulalar belirlenmiştir. Sonuç olarak bu veri seti için FGM kapulası, γ' nın 1.5 ve 2 değerleri için Huang ve Kotz' un birinci kapulası, γ'nın 1.5, 2 ve 3 değerleri için Huang ve Kotz' un ikinci kapulasının uygunluğu gözlenmiştir. Bununla birlikte γ'nın 3 değeri için Huang ve Kotz'un birinci kapulası ile *p* ve *q*'nun 1.5 , 2 ve 3 değerleri için Lai ve Xie'nin kapulasına ilişkin $\hat{\theta}$ tahmini, bu parametre ile ilgili sınırların dışında bir tahmindir. γ=2 için Huang ve Kotz'un ikinci kapulası, en küçük χ^2 değerine sahip olduğundan bu kapula beklenen frekansların elde edilmesi bakımından en uygun olduğu sonucuna varılır.

5. KAYNAKLAR

Çelebioğlu, S., 2007. Üretici Fiyat Endeksi ve Tüketici Fiyat Endeksi Arasındaki Bağımlılık Yapısı Üzerine Bir Çalışma. 16. İstatistik Araştırma Sempozyumu Bildiriler Kitabı, 55-66.

Dall'Aglio, G., 1959. Sulla Compatibilit`a Dele Funzioni di Ripartizioni Doppia. Rend. Mat., 5,385–413.

Eyraud, H., 1938. Les Principes de la Mesure Des Correlations. Ann Univ Lyon Series., A1: 30-47.

Farlie D. J. G., 1960. The Performance of Some Correlation Coefficients for a General Bivariate Distribution. Biometrika, 47:307-323.

Frechet, M., 1951. Les Tableaux de Corr´elation Dont Les Marges Sont Donn´ees, Ann. Univ., Lyon, 9 (Sec. A), 53–77.

Gumbel E. J., 1960a. Bivariate Exponential Distributions. J. Amer. Statist. Assoc., 55:698-707.

Huang J. S., Kotz S., 1999. Modifications of the Farlie-Gumbel-Morgenstern Distributions. a Tough Hill to Climb. Metrika, 49:135–145.

Lai C. D., Xie, M., 2000. A New Family of Positive Quadrant Dependent Bivariate Distributions. Stat. Probab. Lett., 46:359–364.

Morgenstern D., 1956. Einfache Beispiele Zweidimensionaler Verteilungen. Mitteilingsblatt fur Mathematische Statistik, 8:234-235.

Nelsen, R. B., 2006. An İntroduction to Copulas 2nd.edn. Springer Series.

Sklar, A., 1959. Fonctions de r'epartition a n dimensions et leurs marges. Publ. Inst. Statist. Univ. Paris, 8, 229–231.

www.altinkaynak.com (erişim tarihi: 12.01.2011).

FARLIE-GUMBEL-MORGENSTERN COPULA FAMILY, SOME EXTENSIONS AND AN APPLICATION

ABSTRACT

In the last twenty years, the copulas gained a considerable importance in investigating the dependence structure between two or more random variables. Therefore the number of researches on the mathematical and statistical properties of copulas increases from year to year, and hence the application area for copulas becomes so widespread in the same direction. In this study, some extensions of FGM family, proposed by Huang and Kotz (1999) and Lai and Xie (2000) are examined on the Turkish Lira basis of changes in American Dollar and gold prices.

Keywords: FGM copula family, Bivariate distributions, Positive quadrant dependence.

Volume: 09 Number: 01 Category: 02 Page: 45-52 ISSN No: 1303-6319 Cilt: 09 Sayı: 01 Kategori: 02 Sayfa: 45-52

REGRESYON ANALİZİNDE GÖZLEMLERİN AYKIRI DEĞER HARİTASI İLE SINIFLANDIRILMASI

Yasemin KAYHAN ATILGAN* **Süleyman GÜNAY****

ÖZET

Uygulamalarda, üzerinde çalışılan çok boyutlu veri kümeleri, genellikle verinin çoğunluğuna uymayan aykırı gözlemler içerir. Regresyon analizinin önemli aşamalarından bir tanesi de artık analizi ile bu aykırı gözlemleri doğru belirlemektir. Ancak, bu amaçla kullanılan klasik istatistiksel yöntemler aykırı değerlerden çok fazla etkilenir. Dolayısıyla klasik tahmin edicilere dayalı, artık analiz teknikleri araştırmacıyı yanlış yönlendirebilir. Bu çalışmada, çok boyutlu veri kümesindeki gözlemleri incelemek için kullanılan ve klasik tahmin ediciler yerine sağlam tahmin edicilere dayalı olarak oluşturulan aykırı değer haritası basitçe açıklanmıştır. Çalışmanın amacı ise, farklı tahmin ediciler kullanılarak oluşturulan regresyon modelleri ve bu modellere bağlı olarak elde edilen haritaları karşılaştırarak, hangi tahmin edicinin daha güvenilir aykırı değer haritası oluşturacağını tartışmaktır.

Anahtar Kelimeler: Aykırı değer, Sağlam regresyon, Sağlam tahmin ediciler, Uç gözlem.

1. GİRİŞ

Regresyon analizinde veriyi modellemeye geçmeden önce uygulanacak istatistiksel analiz yöntemlerinin geçerliliğini garanti altına alan bir takım varsayımlar vardır. Varsayımların sağlanmadığı durumlarda ilk olarak 'artık analizi' ile varsayım bozulumu yaratan gözlem / gözlemlerin belirlenmesi amaçlanır, daha sonra cevap değişkeni ve açıklayıcı değişkenlere uygun dönüşümler uygulanarak ya da modele yüksek dereceden terimler eklenerek sorun çözülmeye çalışılır. Bu nedenle aykırı değerlerin ya da uç gözlemlerin doğru olarak belirlenmesi regresyon analizinin önemli aşamalarından biridir.

İki değişkenli veri kümeleri ile çalışırken gözlemlerin saçılım grafiklerinden yararlanarak aykırı değerler görsel olarak belirlenebilir, ancak çok boyutlu veri kümelerine geçildiğinde benzer grafikler elde edilemediği için, aykırı değerlerin görsel olarak saptanması iki boyutlu durumdaki kadar kolay değildir. Bu nedenle araştırmacılar çok boyutlu veri kümelerinde şüpheli gözlemleri kolay ve doğru bir biçimde saptayacak yöntemler geliştirilmiştir ve bunlardan bir tanesi de 'aykırı değer haritası / outlier map'dır. Klasik tahmin ediciler yerine sağlam tahmin ediciler kullanılarak oluşturulan harita ile veriler düzenli gözlemler, dikey aykırı değerler, iyi uç gözlemler ve kötü uç gözlemler olarak tek bir grafik yardımıyla sınıflandırılmaktadır. Amaç, araştırmacıya veri kümesindeki muhtemel aykırı değerleri görsel olarak değerlendirme imkanı sunmaktır. Bu çalışmada ilk olarak, Mahalanobis uzaklık ile Sağlam uzaklık kavramlarının farklılığına değinilmiştir.

*Dr., Hacettepe Üniversitesi, Fen Fakültesi, İstatistik Bölümü, e-posta: ykayhan@hacettepe.edu.tr

**Prof. Dr., Hacettepe Üniversitesi, Fen Fakültesi, İstatistik Bölümü, e-posta: sgunay@hacettepe.edu.tr

Daha sonra, klasik tahmin edici - Mahalanobis uzaklık ile oluşturulacak aykırı değer haritası ile, sağlam tahmin ediciler - sağlam uzaklık kullanılarak oluşturulacak haritalar karşılaştırılmıştır. Son olarak, "hangi sağlam tahmin edici kullanılırsa elde edilecek aykırı değer haritası daha güvenilir olur" sorusunu araştırmak amacıyla üç farklı sağlam tahmin edici ile aykırı değer haritaları elde edilerek sonuçlar tartışılmıştır.

2. AYKIRI DEĞER ANALİZİ

En genel tanımı ile aykırı değerler, verinin çoğunluğu ile aynı yapıyı göstermeyen ya da çözümlemelerde kullanılan genel varsayımlardan sapmalar gösteren gözlemlerdir (Hubert vd., 2008). Bu gözlemler genel olarak iki grupta incelenir; y cevap değişkeni doğrultusunda gözlenen, veri kümesinde yer alan diğer gözlem değerlerine göre pozitif ya da negatif yönde daha büyük değerli gözlemler 'dikey aykırı değerler / vertical outliers', x açıklayıcı değişken doğrultusunda gözlenen büyük değerli gözlemler ise 'uç gözlemler / leverage points' olarak adlandırılır (Croux, 2007). Bir uç gözlem, x-uzayında verinin çoğunluğunun yer aldığı düzlemden farklı bir doğrultuda yer alıyorsa 'kötü uç gözlem / bad leverage point', verinin çoğunluğunun yer aldığı düzlem ile aynı doğrultuda yer alıyorsa 'iyi uç gözlem / good leverage point' denir (Rousseeuw ve Zomeren, 1990).

2.1. Sağlam Uzaklık

Analizlerde veri kümesindeki aykırı gözlemleri belirlemek amacıyla sıkça kullanılan bir yöntem 'Mahalanobis Uzaklık / Mahalanobis Distance / MD'dır. Bu uzaklık, veri kümesinde yer alan bir gözlemin, örneklem ortalama vektörüne olan uzaklığının örneklem kovaryans matrisi ile standartlaştırılmış ölçüsüdür. Veri kümesinde ortaya çıkabilecek bir grup aykırı değer, örneklem ortalama vektörünü kendi doğrultusunda çekebilir ve varyans kovaryans matrisini şişirerek varlıklarını gizleyebilir. Dolayısıyla örneklem ortalamasına ve kovaryansına dayalı olarak hesaplanan ve sıkça kullanılan MD yanıltıcı olabilir. Dolayısıyla MD yerine aykırı değerlerden etkilenmeyen ya da daha az etkilenen sağlam yöntemlerin kullanılmasını tercih etmek doğal bir yaklaşımdır. Bu amaçla Campell 1980 yılında, MD de yer alan konum ve ölçeğin tahmini için M tahmin edicilerini kullanmayı önermiştir. Ancak, M tahmin edicisinin kırılma noktası veri kümesindeki değişken sayısına bağlıdır ve değişken sayısı arttıkça sıfıra yakınsamaktadır. Bu da veri kümesindeki değişken sayısı arttığında tahmin edicinin aykırı değerlere karşı olan direncini azaltmaktadır. Bu soruna bir çözüm olarak 1985 yılında Rousseeuw yüksek kırılma noktasına sahip 'En Küçük Hacimli Elipsoid / Minimum Volume Elipsoid / MVE' tahmin edicisinin kullanılmasını önermiştir (Rousseeuw ve Zomeren, 1990). Böylece aykırı değerlerden MD kadar etkilenmeyen 'Sağlam Uzaklık / Robust Distance / RD' kavramı ortaya çıkmıştır. MVE tahmin edicisi veri kümesinde ortaya çıkabilecek aykırı değerlere karşı dirençlidir. Kırılma noktası %50'ye ulaşabilir ancak tahmin edici asimtotik olarak normal dağılıma yakınsamaz, etkinliği düşüktür. Bu nedenle MVE tahmin edicisine alternatif olarak daha yüksek etkiliğe sahip 'En Küçük Kovaryans Determinant / Minimum Covariance Determinant /MCD' tahmin edicisinin önerilmiştir. MCD tahmin edicisine dayalı olarak hesaplanan RD'ler aşağıdaki eşitlik ile hesaplanır,

$$RD_i = \sqrt{(\mathbf{x}_i - \mathbf{T}(\mathbf{X}))\mathbf{C}(\mathbf{X})^{-1}(\mathbf{x}_i - \mathbf{T}(\mathbf{X}))'} \quad (1)$$

MCD tahmin edicisi n gözlemli örneklemde, kovaryans matrisinin determinantı minimum olan ve h gözlemi içeren elipsoidi bulmayı amaçlar. Bu h gözlemin ortalaması veri kümesinin MCD konum tahmini, $\mathbf{T}(\mathbf{X})$, olacaktır. Ölçek tahmini, $\mathbf{C}(\mathbf{X})$, ise yine belirlenen bu h gözlemin kovaryans matrisi ile hesaplanır. Literatürde yer alan FAST-MCD algoritması (Verboven ve Hubert, 2005) ile de, MCD tahminlerinin kolaylıkla elde edilmesi mümkündür (Hubert vd., 2008). MCD konum ve ölçek tahmin edicileri aykırı değerlerden etkilenmedikleri için bu tahmin edicilere dayalı olarak hesaplanan RD de aykırı değerlerden etkilenmez ve veri kümesi ile benzer yapı göstermeyen gözlemler kolayca saptanabilir (Dallal ve Rousseeuw, 1992).

2.2. Bazı Sağlam Tahmin Ediciler

Regresyon analizi iki ya da daha çok değişken arasında bir ilişki olup olmadığının araştırılması ve ilişki varsa bunun matematiksel bir fonksiyon ile tanımlanması olarak açıklanabilir. Amaç, bilinmeyen regresyon katsayılarını tahmin ederek veri kümesine en iyi uyum gösteren regresyon modelini belirlemektir. Bu amaçla en çok kullanılan yöntem klasik 'En Küçük Kareler / Least Squares / LS' regresyondur. Bilinmeyen regresyon parametreleri, 'artık / residual' kareler toplamının minimum yapılması esasına dayalı olarak hesaplanır. LS regresyon, hatalar sıfır ortalama ve sabit varyans ile normal dağılıma sahip olduğu durumda optimal çözümü üretir. Ancak bu varsayımların geçerli olmadığı durumlarda hesaplanan parametre tahminleri yanıltıcı olabilmektedir. Bu nedenle LS regresyona seçenek olacak sağlam tahmin ediciler türetilmiştir. Geliştirilen birçok sağlam tahmin edici ilk bakışta aykırı değerlere karşı dirençli gibi görülebilir. Ancak bu tahmin ediciler y doğrultusunda ortaya çıkacak aykırı değerlere karşı sağlam iken, x doğrultusundaki uç gözlemlerden olumsuz etkilenebilir. Regresyon analizinde ise, veri kümeleri genellikle uç gözlemler içerir. Bu sebeple regresyon modelinden elde edilen artıklar ile aykırı değer analizi yapılırken kullanılan sağlam tahmin edicinin hangi tür aykırı değerlerin varlığında güvenilir sonuçlar ürettiğine dikkat edilmeli ve tahmin edicinin yüksek kırılma noktasına sahip olup olmadığı dikkate alınmalıdır.

Dikey aykırı değerlere karşı sağlamlık özelliğine sahip olmalarına rağmen kötü uç gözlemlerden olumsuz etkilenen tahmin edicilere örnek olarak Huber tarafından 1973 yılında geliştirilen M tahmin edicileri verilebilir. M tahmin edicileri uç gözlemlerden fazla etkilendikleri etkilendikleri için kırılma noktaları $1/n$'dir. Sonraki yıllarda araştırmacılar hem dikey aykırı değerlere hem de kötü uç gözlemlere karşı sağlam, yüksek kırılma noktasına sahip sağlam tahmin ediciler araştırmaya başlamışlardır. Bunlara tipik bir örnek 'En küçük Ortanca Kareler / Least Median of Squares / LMS' tahmin edicisidir. LMS tahmin edicisinin kırılma noktası %50'ye yakınsar, dolayısıyla hem dikey aykırı değerlerin hem de kötü uç gözlemlerin varlığından verinin çoğunluğuna uyan regresyon modelini oluşturur (Rousseeuw ve Leroy, 1987). Ancak asimtotik olarak normal dağılıma yakınsamaz ve etkinliği düşüktür. Bu sebeple istatistiksel çıkarsamalarda tercih edilmese de aykırı değer belirlemede kullanılmaktadır. LMS tahmin edicisinin etkiliğinin düşük olması nedeniyle ona seçenek olarak geliştirilen, kırılma noktası ve etkinliği yüksek bir başka tahmin edici de 'En küçük kırpılmış kareler / Least trimmed squares / LTS' tahmin edicisidir. Son

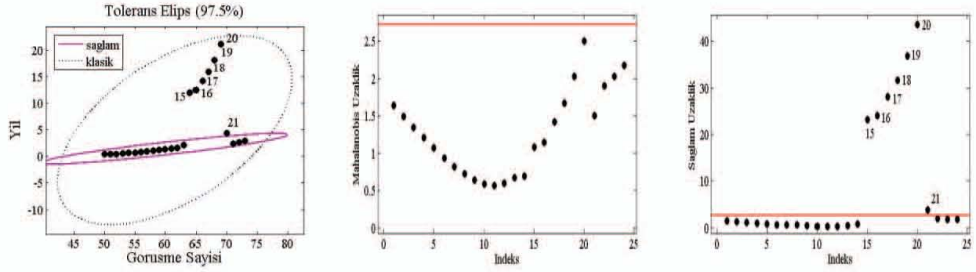
zamanlarda hem hesaplama kolaylığı hemde istatistiksel çıkarsamlarda kullanılabilmesi sebebiyle regresyon analizinde sıklıkla tercih edilmektedir.

2.3. Aykırı Değer Haritası

Veri kümesine regresyon analizi uygulandıktan sonra elde edilen modelden yararlanarak, standartlaştırılmış artık grafiği oluşturulur. Bu grafik sayesinde verinin çoğunluğu ile aynı yapıyı göstermeyen gözlemler yani aykırı değerler belirlenebilir. Ancak bu grafik yardımıyla, belirlenen noktaların dikey aykırı değer mi yoksa uç gözlem mi olduğu kararına varılamaz. Benzer bir problem de RD grafikleri için geçerlidir. RD grafiği ile veri kümesindeki uç gözlemler belirlenebilir. Ancak RD'ler hesaplanırken y_i değerleri dikkate alınmadığından, bu gözlemlerin iyi uç gözlem mi yoksa kötü uç gözlem mi olduğu anlaşılamaz. Bu nedenle hem dikey aykırı değerleri, hem iyi uç gözlemleri, hem de kötü uç gözlemleri tek bir grafik üzerinde gözlemleme olanağı sunan aykırı değer haritası geliştirilmiştir. Sağlam standartlaştırılmış model artıkları, $r_i/\hat{\sigma}$, ve RD'leri kullanarak oluşturulan bu haritada gözlemler, "düzenli (regular) gözlemler - küçük RD ve küçük $r_i/\hat{\sigma}$" , "dikey aykırı değerler - küçük RD ve büyük $r_i/\hat{\sigma}$", "iyi uç gözlemler - büyük RD ve küçük $r_i/\hat{\sigma}$" ve "kötü uç gözlemler - büyük RD ve büyük $r_i/\hat{\sigma}$" olmak üzere dört kategoriye ayrılır (Rousseeuw ve Zomeren, 1990).

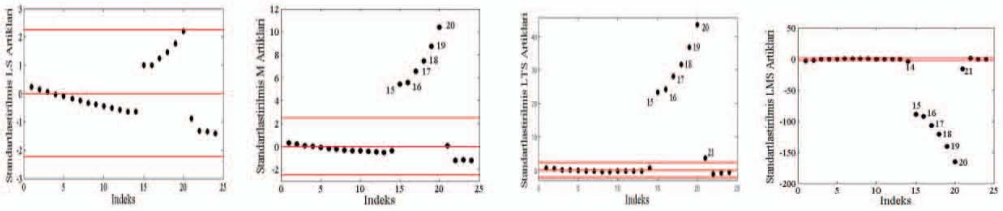
3. UYGULAMA

Bu bölümde, veri kümesinde hem x hem de y doğrultusunda ortaya çıkacak aykırı değerlerin, tahmin edicileri ve dolayısıyla bu tahmin edicilere dayalı olarak elde edilen aykırı değer haritalarını nasıl etkileyeceği ortaya koyulmuş ve yüksek kırılma noktasına sahip tahmin edicilerin kullanılmasının gerekliliği vurgulanmıştır. İlk örnek için, 1950-1973 yılları arasında Belçika'da yapılan uluslararası telefon görüşmelerinin sayısının yer aldığı veri kümesi kullanılmaktadır (Rousseeuw ve Leroy, 1987). Bu veri kümesini kullanmaktaki amaç aykırı değerlerin y cevap değişkeni doğrultusunda gözlenmesi durumunda kullanılan sağlam tahmin edicileri ve bu tahmin edicilere dayalı olarak elde edilen aykırı değer haritalarını karşılaştırmaktır. Şekil 1'de klasik örneklem ortalama ve varyansı ile elde edilen tolerans elips, MCD tahmin edicisi kullanılarak elde edilen tolerans elips, RD ve MD grafikleri verilmiştir. Açıkça görüldüğü gibi veri kümesinde yer alan bir grup aykırı değer, klasik tahminleri etkilemiş ve varlıklarını gizlemiştir. Ancak sağlam tahmin ediciler kullanılarak elde edilen grafikler, bu gözlemlerin aykırı değer olduğuna işaret etmektedir.



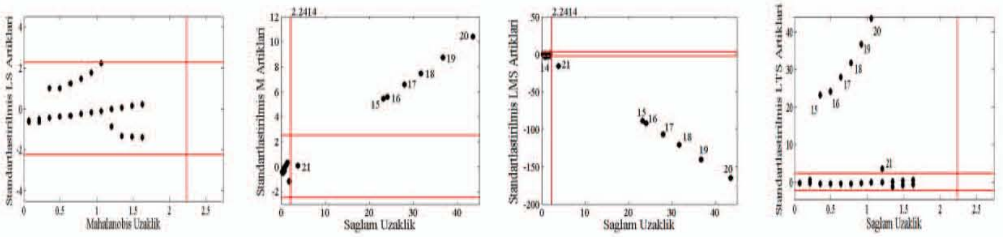
Şekil 1. Sağlam ve klasik tahmin ediciler ile hesaplanan tolerans elipsler, Mahalanobis uzaklık ve sağlam uzaklık.

Uzaklık grafiklerine bakıldığında MD ile veri kümesinde hiçbir aykırı değer tespit edilemezken, RD grafiğinden 15, 16, 17, 18, 19, 20 ve 21 no'lu gözlemlerin verinin çoğunluğu ile aynı yapıyı göstermeyen gözlemler olarak saptamıştır. Daha sonra veri kümesine sırasıyla LS, M, LMS ve LTS regresyon uygulanmıştır. Oluşturulan modellerden standartlaştırılmış artık grafikleri çizdirilmiş ve Şekil 2'de verilmiştir. LS regresyonla veri kümesinde aykırı değer saptanamamış, M tahmin edicisi kullanıldığında, 15, 16, 17, 18, 19 ve 20 no'lu gözlemler, bu gözlemlere ek olarak, LTS tahmin edicisi ile 21, LMS tahmin edicisi ile de 21 ve 14 no'lu gözlemler aykırı değer olarak belirlenmiştir. Daha öncede değinildiği gibi sadece uzaklık grafiklerine ya da sadece artık grafiklerine bakarak belirlenen bu şüpheli gözlemlerin dikey aykırı değer mi, iyi uç gözlem mi yoksa kötü uç gözlem mi olduğu anlaşılamamaktadır.



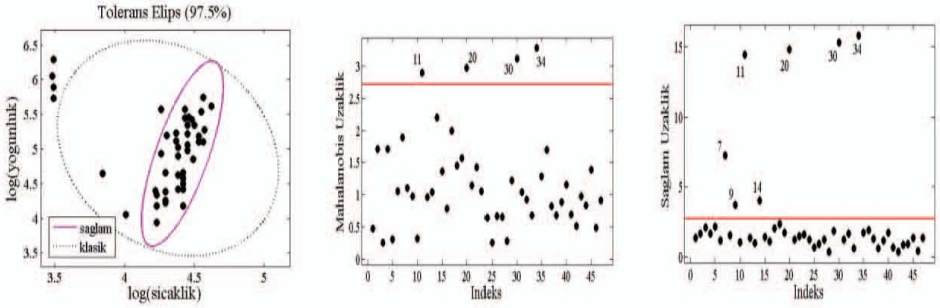
Şekil 2. LS, M, LTS ve LMS regresyon modelleri için standartlaştırılmış artık grafikleri.

Bu sebeple, her bir regresyon modeli için hem hesaplanan uzaklıkların, hem de standartlaştırılmış artıkların tek bir grafik üzerinde birleştirildiği aykırı değer haritaları oluşturulmuş ve Şekil 3'de verilmiştir. Haritalar incelendiğinde, LS regresyon, veri kümesindeki gözlemlerin tamamını düzenli gözlem, M regresyon ise 21 no'lu gözlemi iyi uç gözlem, 15, 16, 17, 18, 19 ve 20 no'lu gözlemleri kötü uç gözlem olarak saptamıştır. LMS regresyon sonucu 14 no'lu gözlem dikey aykırı değer, 15, 16, 17, 18, 19, 20 ve 21 no'lu gözlemler kötü üç gözlem, LTS regresyon ile de 15, 16, 17, 18,19,20 ve 21 no'lu gözlemlerin hepsi dikey aykırı değer olarak tespit edilmiştir.



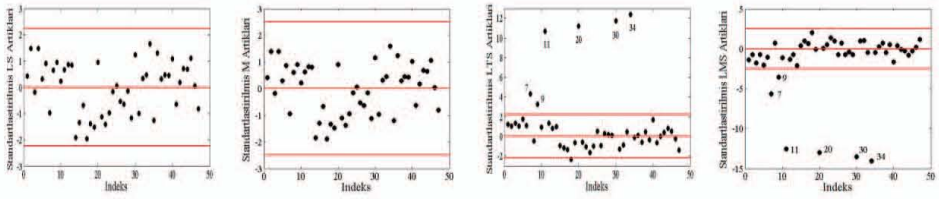
Şekil 3. Aykırı değer haritaları.

İkinci örnekte, x açıklayıcı değişken doğrultusunda gözlenen kötü uç gözlemlerin tahmin ediciler ve dolayısıyla aykırı değer haritaları üzerindeki etkisi incelenmektedir. Bu amaçla, Hertzsprung-Russell'in 47 yıldızdan oluşan CYG OB1 veri kümesi kullanılmıştır (Rousseeuw, Leroy, 1987). Klasik ve sağlam tahmin ediciler ile oluşturulan tolerans elipsler, MD ve RD grafikleri Şekil 4'de verilmektedir. Uzaklık grafikleri incelendiğinde MD ile 11, 20, 30 ve 34 no'lu gözlemler, RD ile bunlara ek olarak 7, 9 ve 14 no'lu gözlemler aykırı gözlem olarak belirlenmiştir.

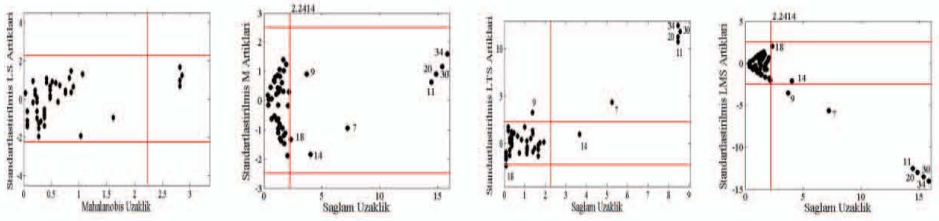


Şekil 4. Sağlam ve klasik tahmin ediciler ile hesaplanan tolerans elipsler, Mahalanobis uzaklık ve sağlam uzaklık.

Şekil 5 ile verilen standartlaştırılmış artık grafikleri değerlendirildiğinde LS ve M tahmin edicileri ile veri kümesindeki aykırı gözlemlerin belirlenemediği gözlenmektedir. LMS ve LST tahmin edicileri ise 7, 9, 11, 20, 30 ve 34 nolu gözlemlerin veri kümesinin çoğunluğu ile aynı yapıyı göstermeyen gözlemler olduğunu tespit etmiştir. Şekil 6 ile verilen aykırı değer haritaları karşılaştırıldığında ise, LS ve M tahmin edicileri veri kümesinde yer alan aykırı değerleri iyi uç gözlem olarak sınıflandırmaktadır. LMS tahmin edicisi ile 7, 9, 11, 20, 30, 34 nolu gözlemler kötü uç gözlem olarak belirlenmiştir. LTS tahmin edicisi ile de 9 ve 18 nolu gözlemler dikey aykırı değer, 7, 11, 20, 30 ve 34 nolu gözlemler de kötü uç gözlem olarak saptanmıştır.



Şekil 5. LS, M, LTS ve LMS regresyon modelleri için standartlaştırılmış artık grafikleri.

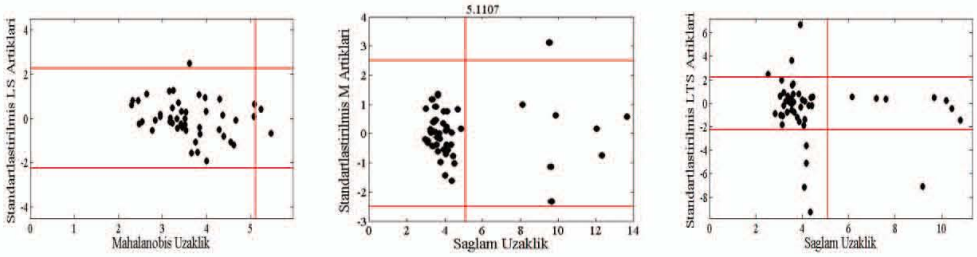


Şekil 6. Aykırı değer haritaları.

Bu iki örnek değerlendirildiğinde klasik bir tahmin edici olan LS'in veri kümesinde hem x hem de y doğrultusunda ortaya çıkabilecek aykırı değerlerden fazlasıyla etkilendiğini ve aykırı değer haritası oluşturulurken kullanılmaması gerektiği açıktır. Benzer biçimde M sağlam tahmin edicisi kullanılarak bu haritaların oluşturulmasının, kötü uç gözlemlerin varlığında LS tahmin edicisi gibi yanıltıcı olabileceği söylenebilir. LMS tahmin edicisinin, veri kümesinden ufak sapmalar gösteren gözlemleri de aykırı değer olarak sınıflandırabileceği görülmektedir. Ayrıca veri kümesinden alt örneklem seçmeye dayalı olarak hesaplanan LMS tahmin edicisi, veri kümesindeki değişken sayısı arttığında hesaplama süresi açısından etkin olmayacaktır. Sonuç olarak, hem x

hem de y doğrultusundaki aykırı değerlerden etkilenmeyen ve iteratif olarak hesaplanması daha kolay olan LTS tahmin edicisi kullanılarak aykırı değer haritası oluşturmak tercih edilebilir.

Son olarak, çok boyutlu veri kümelerine örnek olması amacıyla 1973 yılında Amerika Birleşik Devletleri'nin 47 eyaletindeki suç oranlarına ilişkin bir çalışmadan alınan 47 gözlem ve 14 değişkenli veri kümesine üzerinden LS, M ve LTS regresyona dayalı aykırı değer haritaları elde edilmiş ve Şekil 7'de verilmiştir.



Şekil 7. LS, M ve LTS ile oluşturulan aykırı değer haritaları.

LS tahmin edicisi ve MD kullanılarak elde edilen harita ile, veri kümesinde sadece 1 adet aykırı gözlem belirlenmiş, M tahmin edicisi ve RD kullanıldığında 1 adet kötü uç gözlemin varlığına işaret edilmiş, LTS tahmin edicisi ve RD kullanıldığında ise veri kümesinde 7 adet dikey aykırı değer ve 1 tane de kötü uç gözlem olabileceği tespit edilmiştir.

4. SONUÇ VE TARTIŞMA

İki değişkenli veri kümeleri ile çalışırken, gözlemlerin saçılım grafiğinin ya da standartlaştırılmış artıklar grafiğinin incelenmesi ile görsel olarak veri kümesinin çoğunluğu ile aynı yapıyı göstermeyen gözlemlerin belirlenmesi mümkün iken, çok boyutlu veri kümelerine geçildiğinde benzer grafikler elde edilememektedir. Bu gözlemleri saptamak için kullanılan klasik yöntemler yanıltıcı sonuçlar üretebilmektedir. Bu problemin çözümüne ilişkin önerilen aykırı değer haritaları ile veriyi 4 grupta kategorize etmek ve değerlendirmek mümkündür. Çoklu konum ve ölçeğin sağlam tahmin edicisine dayalı olarak hesaplanan sağlam uzaklıklar ile, sağlam regresyon sonucu elde edilen standartlaştırılmış artıkların kullanıldığı bu yöntemdeki tahmin ediciler aykırı değerlerden etkilenmediği için veri kümesinin çoğunluğu ile aynı yapıyı göstermeyen gözlemler kolaylıkla saptanabilmektedir. Ancak haritalar oluşturulurken veri kümesine uygulanacak sağlam regresyon yönteminin yüksek kırılma noktasına sahip olması, oluşturulacak haritanın ve belirlenen şüpheli gözlemlerin doğruluğunu arttıracaktır. Bu sebeple hem x hem de y doğrultusunda gözlenecek aykırı değerlerden etkilenmeyen LTS gibi yüksek kırılma noktasına sahip tahmin edicilerin kullanılması tercih edilmelidir. Elbette bir çalışmada, gözlemlerin tek bir tanı aracı ile kesin olarak aykırı değer kabul edilmesi yanıltıcı olabilir. Bu çalışmadaki amaç, araştırmacıya veri kümesinde incelenmesi gereken şüpheli gözlemleri gösterecek, kullanımı kolay bir yöntemi tanıtmak ve hangi durumlarda hangi sağlam tahmin edicinin kullanılmasının daha güvenilir sonuçlar vereceği konusunda karşılaştırmalı bir uygulama sunmaktır.

5. KAYNAKLAR

Croux, C., 2007. An Introduction to Robust Statistics: Mathematics and Practice. Lecture Notes, Faculty of Economics and Management, University Center of Statistics.

Dallal, G.E., Rousseeuw, P. J., 1992. LMSMVE A Program for Least Median of Squares Regression and Robust Distances, Computers and Biomedical Researches, Vol 25, 384-391.

Hubert, M., Rousseeuw, P. J., Aelst, S. V., 2008. High-Breakdown Robust Multivariate Methods, Statistical Science, Vol.23, No.1, 92-119.

Rousseeuw, P. J., Leroy, A. M., 1987. Robust Regression and Outlier Detection, John Wiley & Sons, New York.

Rousseeuw, P. J., Van Zomeren, B. C., 1990. Unmasking Multivariate Outliers and Leverage Points. Journal of the American Statistical Association, Theory and Methods, Vol.85, No.411.

Verboven, S., Hubert, M., 2005. LIBRA: A MATLAB Library for Robust Analysis. http://wis.kuleuven.be/stat/robust/LIBRA.html.

CLASSIFICATION OF THE OBSERVATIONS IN REGRESSION ANALYSIS BY OUTLIER MAP

ABSTRACT

In practice, multidimensional data sets generally contain observations that deviate from the majority of data. One of the important stages of regression analysis is to correctly determine these observations by using residual analysis. However, conventional statistical methods used for this purpose are too much influenced by outliers. Therefore, the outlier analysis techniques based on classical estimators may mislead the investigator. In this study, outlier map which is used to examine observations in multidimensional data sets and generated by robust estimators instead of the classical estimators is briefly explained. The aim of this study is to compare outlier maps of different regression models generated by using different robust estimators and to discuss which robust estimator will create more reliable map.

Keywords: Outliers, Robust regression, Robust estimators, Extreme observation.

Volume: 09 Number: 01 Category: 02 Page: 53-64 ISSN No: 1303-6319 Cilt: 09 Sayı: 01 Kategori: 02 Sayfa: 53-64

COVARIATES OF UNIT NONRESPONSE ERROR BASED ON PROXY RESPONSE FROM HOUSEHOLD SURVEYS

A. Sinan TÜRKYILMAZ* **H. Öztaş AYHAN****

ABSTRACT

Unit nonresponse error and its related covariates are examined from the results of a sample survey. A procedure is proposed to study unit nonresponse when data are from a two stage household sample survey in which household are the units of the first level and individuals are the units of second level. The individual person responses within the sample survey did not contain information on the nonrespondents. Therefore, household schedule variables which are based on proxy person response information are combined with the binary dependent response/nonresponse variable from the individual survey records. The idea is to estimate a logistic model whose dependent variable is the binary unit response indicator and where individual characteristics at the right hand side are approximated by household information collected at the first level. Among other models, a binary logistic regression model is proposed and the results are analyzed and interpreted by the computed odds ratios. The results have indicated several significant covariates for the model of nonresponse.

Keywords: Binary dependent variable, Covariates of nonresponse, Logistic regression, Nonresponse error components, Proxy response.

1. INTRODUCTION

Unit nonresponse is the failure to obtain the minimum required information from an eligible housing unit or person in the sample. Unit nonresponse occurs when the responsents are unable or unwilling to participate; interviewers are unable to locate addresses or respondents, or when other barriers exist for completing the interview.

Covariates of unit nonresponse error have been a concern of survey researchers as a major part of the total survey error. Components of unit nonresponse error are basically associated with the factors related to the reasons of survey non-participation.

In order to have logical causality measures, one has to identify the direct and indirect factors affecting such relations. In many cases, information on such ideal factors may not be available as a survey variable, due to the current objectives of such a survey. Alternative information can be derived from the other existing survey variables which are naturally available due to the survey objective. Consequently, the researchers have to make sense out of such information, because the ideal information which will explain the causality may not be available.

*Assoc. Prof., Hacettepe University, Institute of Population Studies, 06100 Ankara, e-mail: aturkyil@hacettepe.edu.tr

**Prof., Middle East Technical University, Department of Statistics, 06800 Ankara, e-mail: oayhan@metu.edu.tr

With a limited research budget, one can obtain information only on a reasonably small scale. On the other hand, for a large scale survey, additional questions will also bring extra cost, which may not be tolerable by the survey management. Under the circumstances, another alternative may be to utilize the best of the available information.

The examination of the components of unit nonresponse in a demographic survey have been given by Ayhan (1981), and some of the other recent studies have also been evaluated (Ayhan, 1998). The current study examines the issue by taking an alternative approach. The following sections of this paper cover the methodology used, covariates of nonresponse, proposed models and testing, and the conclusions of the findings from this investigation.

2. SURVEY METHODOLOGY

2.1. Sample Design and Implementation

The sample design and sample size of the *Turkey Demographic and Health Survey* (TDHS) – 2003 (HUIPS, 2004) make it possible to perform analyses for Turkey as a whole, for urban and rural areas and for the five demographic regions of the country. A weighted, multistage, stratified cluster sampling approach was used in the selection of the survey sample. The results of the household and individual questionnaire executions are summarized in Table 1.

Table 1. Results of the household and individual interviews in 2003 Turkey Demographic and Health Survey

Outcomes	Urban	Rural	Total
Household interviews:			
Selected sample households	8718	2941	11659
Households interviewed	7956	2880	10836
Household Nonresponse Rate (*HHRR*)	0.087	0.021	0.071
Individual interviews:			
Eligible women selected	6259	2188	8447
Eligible women interviewed	5976	2099	8075
Individual Nonresponse Rate Component (*IRRC*)	0.045	0.041	0.044
Individual Person Nonresponse Rate (*IPNRR*)*	0.128	0.061	0.112

* Computation of the *IPNRR* = [*1 – HHRR * IRRC*]

The target sample size of the TDHS–2003 was set at 13160 dwelling units. This was expected to yield about 11000 completed household interviews. Out of 11659 selected sample households, 10836 number of households were interviewed. Within this, 8447 number of eligible women was present and 8075 was interviewed during the survey operation. Information is provided on the overall coverage of the sample, including household and individual nonresponse rates.

2.2. Questionnaire Design

The data collection for household sample surveys have been executed in two stages; the completion of the household schedule, and the individual survey. The *household schedule* is completed by a selected adult member of the household, as a proxy respondent for the other members of the household, and a self respondent for him/herself.

For the *individual survey*, data are only collected from the eligible women as a self respondent, and no information is available for the non-respondents. On the other hand, household schedule also contains some more additional information about other characteristics of the respondents and non-respondents of the individual survey.

For the responding households, generally the household schedule contains full information on all household members. On the other hand, the selected household member for the individual survey may or may not respond to the individual person's interview. Consequently, we will have two possible groups for the individual survey; respondents and non-respondents.

This study combines the household based proxy information for selected variables, and response-nonresponse outcome information of the individual person's survey from the same household.

3. COVARIATES OF NONRESPONSE

The following household information is obtained from the household schedule by proxy interviews;

A. Independent survey variables: (*Based on household survey information*)

1. ***Stratification variables used as survey variables*****:**
- Region
- Type of place of residence

2. ***Household based proxy individual variables*****:**
- Gender
- Age groups
- Place of birth
- Maternal and paternal survival
- Migration and mobility
- Literacy and education status
- Work status
- Marital status

3. *Housing characteristics*:

- Household ownership
- Safe water access
- Sanitary toilet
- Number of rooms
- Household durability
- Household facilities
- Household income

B. Dependent survey variable: (*Based on individual survey information*)

- Binary nonresponse information

Some of the household based current and generated variables, their response options, and their frequencies are given in Table 2.

4. PROPOSED MODELS AND TESTING

4.1. Search for Models

In the literature, multinomial logistic regression models are grouped into two distinct types as generalized and cumulative logit models. Generalized logit models are usually employed when the response categories are unordered whereas cumulative logit models should be employed when response categories are ordered. Both classical and Bayesian methodologies are available to estimate the model parameters. Moreover, multinomial logistic regression models are developed to analyze categorical response data occuring in matched case-control studies.

For the analysis of data occuring in matched case-control studies, conditional logistic regression likelihood functions are developed to adjust the analysis for the nuisance parameters that are of high dimension. There is a vast literature on multinomial logistic regression models and analysis. For instance Hosmer and Lemeshow (2000) and Agresti (2002) provide the basics, extensions, as well as related special topics including logistic regression analysis for correlated data.

Besides well established multinomial logistic regression models, novel developments emerged in recent years motivated by categorical response data with interesting features that occur especially in epidemiological studies. Of the recent developments, Chatterjee (2004) developed a two stage multinomial logistic regression approach to analyze data with multivariate classification information and derived the asymptotic properties of the test statistics.

Table 2. Current and generated variables, options and their frequencies

Name of variables	Code and Explanation	Weighted percent
Response and Nonresponse	1 Nonresponse	4.7
	0 Response	95.3
hv017- Number of visits to household	1	79.7
	2	14.9
	3	5.4
v024 – Regions	1 West	40.7
	2 South	12.7
	3 Central	23.1
	4 North	7.3
	5 East	16.2
hv025 - Type of place of residence	1 Urban	71.2
	2 Rural	28.8
hv270 - Wealth index	1 Poorest	15.6
	2 Poorer	18.1
	3 Middle	20.2
	4 Richer	22.4
	5 Richest	23.6
hv102 - Usual resident	0 No	3.6
	1 Yes	96.4
sh26 - Currently working	0 No	75.1
	1 Yes	24.9
SANITATE- Sanitary toilet	0 No	90.7
	1 Yes	9.3
SAFEWAT – Safewater	0 No	92.4
	1 Yes	7.6
CROWD – Number of persons per room	0 less than 3	80.5
	1 more than 3 and over	19.5
Educ – Education level	1 No education / Primary incomplete	22.1
	2 Primary complete/ secondary incomplete	60.7
	3 Secondary +	17.2
hv116 - Marital status	1 Currently married	94.7
	2 Formerly / ever married	5.3
agegroup – Age groups	1 15-19	3.0
	2 20-24	12.9
	3 25-29	18.2
	4 30-34	18.3
	5 35-39	17.5
	6 40-44	16.5
	7 45-49	13.5

In this study, individual survey respondent's related household schedule characteristics are used as possible covariates for the non-response error. The possible covariates are evaluated under several alternative statistical models. For this purpose, several generalized linear models have been examined. As possible alternatives, *loglinear model, logit model, probit model,* and *logistic regression model*s have been evaluated.

After the examination of the current available variables, *multiple logistic regression model* has been selected. Summary measures of goodness-of-fit are provided as output with any fitted model and give an overall indication of the fit of the model (Hosmer and Lemeshow, 1980, and Lemeshow and Hosmer, 1982).

The present model takes non-response as the binary dependent variable which is associated with the other household covariates. In order to test our model, the latest TDHS – 2003 data is used. Questions and topics which are listed in Section 3 were asked during the household interviews. The household survey and individual person's survey data sets are combined under the weighted, stratified cluster design, for the survey analysis. The *SPSS 13.0's "complex samples" feature* were used to perform *binary logistic regression,* where the sample design was naturally taken into account.

4.2. Inferences from Binary Logistic Regression

A binary logistic regression model has been proposed to explain the effect of covariates on survey unit nonresponse for this study. After the regression diagnostics, such as outlier detection and collinearity tests were performed the following model and results were obtained. Some variables were not taken into account, such as work type, since only a portion of women are working. Moreover, only variables available for "all cases" were included to increase the number of cases in model.

The hypothesis to be tested is

$$H_0 : \beta_i = 0 \quad \text{versus} \quad H_a : \beta_i \neq 0 .$$

The binary logistic regression prediction equation for an *S*–shaped curve for the desired probability *p* is

$$p = \exp\left(\hat{\alpha} + \sum_{i=1}^{k} \hat{\beta}_i x_i \right) \Big/ \left[1 + \exp\left(\hat{\alpha} + \sum_{i=1}^{k} \hat{\beta}_i x_i \right) \right] . \qquad (1)$$

Within the *S*–shaped regression model, the probability *p* falls between 0 and 1 for all possible *x* values. Test statistics for the regression model coefficients are

$$t_i = \left(\hat{\beta}_i - \beta_i \right) / \text{se} \left(\hat{\beta}_i \right) . \qquad (2)$$

4.3. The Odds Ratio

The odds ratio ($\boldsymbol{\theta}$) is a measure of association which has found wide use in many disciplines. It approximates how much more likely (or unlikely) it is for the outcome to be present among those with $x = 1$ than among those with $x = 0$ (Lemeshow and Hosmer, 1983). The odds ratio is usually the parameter of interest in a logistic regression due to its ease of interpretation. The interpretation given for the odds ratio is based on the fact that in many instances it approximates a quantity called the *relative risk* (Hosmer and Lemeshow, 2000). Along with the point estimate of a parameter, it is a good idea to use a confidence interval estimate to provide additional information about the parameter value.

The odds ratio is used to interpret the computed coefficients of the binary logistic regression prediction equation, in terms of relative comparative risks. The data layout structure of the odds related variables are given in Table 3, below.

Table 3. The data layout structure for odds

Variables	Nonresponse	Response	Total
Variable *option A*	n_{11}	n_{12}	n_{1+}
Variable *option* A^c	n_{21}	n_{22}	n_{2+}
Total	n_{+1}	n_{+2}	n

The desired (success) probabilities for the two groups are;

π_1 is estimated by $p_1 = n_{11} / n_{1+}$,

π_2 is estimated by $p_2 = n_{21} / n_{2+}$.

In 2×2 contingency tables, the *relative risk* is the ratio of the desired probabilities for the two groups.

$$\text{The Relative Risk} = \pi_1 / \pi_2 \tag{3}$$

The ratio of odds from two rows is given by

$$\theta = \frac{\pi_1 (1 - \pi_1)}{\pi_2 (1 - \pi_2)} = \frac{\pi_{11}\ \pi_{22}}{\pi_{12}\ \pi_{21}} . \tag{4}$$

Sample odds (cross–product) ratio is

$$\hat{\theta} = \frac{p_1 / (1 - p_1)}{p_2 / (1 - p_2)} = \frac{n_{11}\ n_{22}}{n_{12}\ n_{21}} . \tag{5}$$

The odds ratio can be equal to any nonnegative number.

The odds ratio can be interpreted as;

(1) When $1 < \theta < \infty$, the odds of success are higher in row 1 than in row 2.

(2) When X and Y are independent, $\pi_1 = \pi_2$, so that

$$\theta = [\text{odds}_1 / \text{odds}_2] = 1 .$$

(3) When $0 < \theta < 1$, a success is likely in row 1 than in row 2, that is $\pi_1 < \pi_2$.

Generalized linear models yield fitted coefficients that are commonly used to estimate odds ratio or other measures of association. Standard fitting techniques such as maximum likelihood and estimating equation methods yield consistent estimators with

first order asymptotically normal sampling distributions (Cox and Oakes 1984; Agresti 2002; Lyles, Guo and Greenland 2012).

Recently, Allen and Le (2008) introduced the overall odds ratio (OOR) as a new index for quantifying the overall effect size in logistic regression models. The OOR can be interpreted in the same way as the odds ratio of individual independent variables. It is the ratio of the odds of belonging to a category of the dependent variable that a researcher is interested in predicting when the weighted linear combination of the independent variables increases one standard deviation to the odds before such an increase (Le and Marcus 2012).

4.4. Model Based Survey Statistics and Outcomes

Once we have fit a particular multiple (multivariable) logistic regression model, we begin the process of model assessment. The first step in this process is usually to assess the significance of the variables in the model. The likelihood ratio test for overall significance of the p coefficients for the independent variables in the model is performed in exactly the same manner as in the univariate case (Hosmer and Lemeshow, 2000).

Before concluding that any or all of the coefficients are nonzero, we may wish to look at the univariate Wald test statistics. Under the hypothesis that an individual coefficient is zero, these statistics follow the standard normal distribution. In order to obtain the best fitting model while minimizing the number of parameters, the next logical step is to fit a reduced model containing only those variables thought to be significant, and compare it to the full model containing all the variables (Hosmer and Lemeshow, 2000).

The following proposed model is fitted to the TDHS 2003 data.

$$p = \Pr(Y=1) = \exp\left(\hat{\alpha} + \sum_{i=1}^{k} \hat{\beta}_i x_i\right) \Big/ \left[1 + \exp\left(\hat{\alpha} + \sum_{i=1}^{k} \hat{\beta}_i x_i\right)\right] \quad \text{where,} \tag{6}$$

$$\hat{\alpha} + \sum_{i=1}^{k} \hat{\beta}_i x_i = -1.615 + 0.563*\text{hv024(1)} + 0.549*\text{hv024(2)} + 0.470*\text{hv024(3)} + 1.577*\text{hv102(0)} - 0.451*\text{sh26(0)} - 0.656*\text{hv116(1)} - 0.557*\text{agegroup(2)} - 0.433*\text{agegroup(3)} - 0.469*\text{agegroup(4)} - 0.448*\text{agegroup(5)} \tag{7}$$

Information on the related correlation measures are given in Table 4. The Nagelgerke R–square is used as a pseudo R-square of linear regression and measures the power of model in terms of how the model explains the variation in dependent variables by independent variables.

Table 4. Several pseudo R square values for the model

Test statistics	R–square
Cox and Snell	0.021
Nagelgerke	0.066
McFadden	0.056

The Nagelgerke R–square is 0.066 so the power of the model is low but the model is significant (with a *p*-value of 0.000, and Wald statistics value = 7.289, df 1 = 25, df 2 = 322).

The results of the test statistics for the model effects are presented in Table 5. Within the logistic regression model, "the number of visits", "region", "being usual resident", "currently working", "educational level" and "marital status" stands as significant independent variables.

Table 5. Results of the test statistics for model effects

Sources	df 1	df 2	Wald F	Significance	Indicator
(Corrected model)	25	322	7.29	0.00	*
(Intercept)	1	346	54.61	0.00	*
hv017 - Number of visits	2	345	3.12	0.05	*
hv024 – Region	4	343	2.63	0.03	*
hv025 - Type of place of residence	1	346	0.97	0.33	
hv270 - Wealth index	4	343	1.03	0.39	
hv102 - Usual resident	1	346	63.59	0.00	*
sh26 - Currently working	1	346	7.28	0.01	*
SANITATE - Sanitary toilet	1	346	1.09	0.30	
SAFEWAT - Safewater	1	346	0.00	0.96	
CROWD - No of persons per room	1	346	0.30	0.58	
Educ - Education level	2	345	5.43	0.00	*
hv116 – Marital status	1	346	10.35	0.00	*
Age groups	6	341	1.88	0.08	

Finally, the model parameter estimates of the binary logistic regression model are given in detail in Table 6.

Table 6. Binary logistic regression model parameter estimates

Variables	Category	$\hat{\beta}_i$	$se(\hat{\beta}_i)$	t_i	df	p-value	$deff$	$\hat{\theta}$	Indicator
Intercept		-1.615	0.560	-2.885	346	0.00	1.54	0.20	*
hv017- Number of visits	1	-0.284	0.282	-1.004	346	0.32	1.69	0.75	
	2	0.192	0.296	0.650	346	0.52	1.74	1.21	
	3	0	.	.	.	.	.	1.00	
hv024 – Region	1 West	0.563	0.201	2.803	346	0.01	1.11	1.76	*
	2 South	0.549	0.238	2.309	346	0.02	1.21	1.73	*
	3 Central	0.470	0.224	2.098	346	0.04	1.19	1.60	*
	4 North	0.190	0.284	0.671	346	0.50	1.01	1.21	
	5 East	0	.	.	.	.	.	1.00	
hv025 - Type of place of residence	1 Urban	0.170	0.173	0.983	346	0.33	1.43	1.19	
	2 Rural	0	.	.	.	.	.	1.00	
hv270 - Wealth index	1 Poorest	-0.238	0.277	-0.859	346	0.39	1.76	0.79	
	2 Poorer	-0.358	0.206	-1.735	346	0.08	1.24	0.70	
	3 Middle	-0.264	0.210	-1.258	346	0.21	1.50	0.77	
	4 Richer	-0.343	0.197	-1.739	346	0.08	1.49	0.71	
	5 Richest	0	.	.	.	.	.	1.00	
hv102 - Usual resident	0 No	1.577	0.198	7.974	346	0.00	1.30	4.84	*
	1 Yes	0	.	.	.	.	.	1.00	
sh26 - Currently working	0 No	-0.451	0.167	-2.699	346	0.01	1.83	0.64	*
	1 Yes	0	.	.	.	.	.	1.00	
SANITATE- Sanitary toilet	0 No	-0.280	0.268	-1.042	346	0.30	1.69	0.76	
	1 Yes	0	.	.	.	.	.	1.00	
SAFEWAT - Safewater	0 No	-0.011	0.243	-0.045	346	0.96	1.53	0.99	
	1 Yes	0	.	.	.	.	.	1.00	
CROWD – no of persons per room	0 less than 3	-0.114	0.208	-0.548	346	0.58	1.68	0.89	
	1 more than 3 and over	0	.	.	.	.	.	1.00	
Educ – education level	1 No education/ Primary incomplete	0.335	0.245	1.366	346	0.17	1.58	1.40	
	2 Primary complete/ secondary incomplete	-0.198	0.178	-1.114	346	0.27	1.42	0.82	
	3 Secondary +	0	.	.	.	.	.	1.00	
hv116 - marital status	1 Currently married	-0.656	0.204	-3.217	346	0.00	1.25	0.52	*
	2 Formerly/ ever married	0	.	.	.	.	.	1.00	
Age Group	1 15-19	0.136	0.369	0.368	346	0.71	1.63	1.15	
	2 20-24	-0.557	0.234	-2.384	346	0.02	1.26	0.57	*
	3 25-29	-0.433	0.192	-2.253	346	0.02	1.15	0.65	*
	4 30-34	-0.469	0.197	-2.374	346	0.02	1.20	0.63	*
	5 35-39	-0.448	0.215	-2.083	346	0.04	1.47	0.64	*
	6 40-44	-0.379	0.216	-1.754	346	0.08	1.55	0.68	
	7 45-49	0	.	.	.	.	.	1.00	

For the coefficients of this model, the following results can be summarized in terms of odds ratios. The probabilities of being "non-responder" women are 1.76, 1.73 and 1.60 times higher for women who are in West, South and Central regions when compared to women in East region. Temporary members of the household are 4.84 times more likely to be "non-responders" than the usual members of the household. Non-working women are 1.56 (=1 / 0.64) times better responders compared to working women. Similarly,

currently married women are 2 (=1 / 0.52) times better responders. Excluding the youngest age group of reproductive women aged 15-19, all other age groups are about 1.5 times better responders compared to the oldest age group of 45-49.

5. CONCLUSIONS

Since the number of independent variables is limited to questions asked in the household questionnaire and some of them are not included into the model due to small number of cases, the number of significant independent variables is few. However, as expected, the number of visits, the region where the woman lives are significant and the "East" region of Turkey gives smaller odds value; meaning that the response rates are higher than the other regions. In addition, naturally "being a usual resident" and "currently working" are also significant and usual residents and non-working women are better responders. "Being a currently married women" and "middle age women within the reproductive age groups of 15-49" are also significant.

As it is stated earlier the variables that are included into the regression model are based on proxy information and limited to the information collected by household questionnaire. This model can be thought as an indirect way of examining the covariates of non-response when it is not possible to measure the non-response by a well-defined independent module added to the study and applied to non-responders directly. If the number of proxy information is increased, future models may include more independent variables to the model and the power of model may be higher.

6. REFERENCES

Agresti, A., (2002). Categorical Data Analysis. 2nd edition. John Wiley, Hoboken, NJ.

Allen, J. and Le, H., (2008). An Additional Measure of Overall Effect Size for Logistic Regression Models. Journal of Educational and Behavioral Statistics 33, 416 – 441.

Ayhan, H. Ö., (1981). Sources of Nonresponse Bias in 1978 Turkish Fertility Survey. Turkish Journal of Population Studies 2–3, 104–148.

Ayhan, H. Ö., (1998). Survey Nonresponse Models and Applications in Turkey. In Official Statistics in a Changing World. Stockholm: Statistics Sweden Press, pp. 155–158.

Chatterjee, N., (2004). A Two Stage Regression Model for Epidemiological Studies with Multivariate Disease Classification Data. Journal of the American Statistical Association 99, 127 – 138.

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

Hosmer, D. W., and Lemeshow, S., (1980). A Goodness-of-Fit Test for the Multiple Logistic Regression Model. Communications in Statistics A 10, 1043–1069.

Hosmer, D. W., and Lemeshow, S., (2000). Applied Logistic Regression (Second Edition). New York: John Wiley and Sons, Inc.

HUIPS., (2004). Turkey Demographic and Health Survey, 2003. Hacettepe University Institute of Population Studies, Ministry of Health General Directorate of Mother and Child Health and Family Planning, State Planning Organization and European Union. Ankara, Turkey. 309 pp.

Le, H. and Marcus, J., (2012). The Overall Odds Ratio as an Intuitive Effect Size Index for Multiple Logistic Regression: Examination of Further Refinements. Educational and Psychological Measurement 76(6), 1001 – 1014.

Lemeshow, S., and Hosmer, D. W., (1982). The Use of Goodness-of- Fit Statistic in the Development of Logistic Regression Models. American Journal of Epidemiology 115, 92–106.

Lemeshow, S., and Hosmer, D. W., (1983). Estimation of Odds Ratio with Categorical Scaled
Covariates in Multiple Logistic Regression Analysis. American Journal of Epidemiology 119, 147–151.

Lyles, R. H., Guo, Y., and Greenland, S., (2012). Reducing Bias and Mean Squared Error Associated with Regression-Based Odds Ratio Estimators. Journal of Statistical Planning and Inference 142, 3235 – 3241.

HANEHALKI ARAŞTIRMALARINDA YERİNE CEVAPLAYICIDAN ELDE EDİLEN BİRİM CEVAPLANMAMA HATASI ORTAK DEĞİŞKENLERİNİN BİLEŞENLERİ

ÖZET

Birim cevaplanmama hatası ve ortak değişkenlerinin bileşenleri, yapılan bir örneklem araştırmasının sonuçlarına dayanarak incelenmiştir. Birinci aşaması hanehalkı ve ikinci aşaması kişi düzeyinde gerçekleşen iki aşamalı bir çalışmanın verilerde birim cevaplanmama hatasını çalışmak için bir prosedür önerilmiştir. Bu çalışmadaki kişi düzeyinde cevaplanmama ile ilgili bilgiler bulunmamaktadır. Bu nedenle, hanehalkı araştırmasında bulunan seçilmiş değişkenlerle ilgili bilgiler yerine cevaplayıcıdan elde edilmiş ve bu bilgiler aynı kişiye ait olan kişi araştırmasının sonuçlarındaki ikili cevaplama/cevaplamama bağımlı değişkeniyle birleştirilmiştir. Düşünce, cevaplanmama göstergelerini açıklamak için lojistik regresyon modeli geliştirilmesi ve modelin sağ tarafı kişi özelliklerinin ilk aşamada toplanan hanehalkı bilgileriyle yakınsamalarıdır. Diğer modellerin yanında,bir lojistik regresyon önerilmiş ve sonuçlar hesaplanan ihtimaller oranı ile analiz edilmiş ve yorumlanmıştır. Elde edilen sonuçlar,cevaplanmama modelini etkileyen bazı önemli ortak değişkenlerin mevcut olduğunu göstermektedir.

Anahtar Kelimeler: Cevaplanmama hatası bileşenleri, Cevaplanmama ortak değişkenleri, Kesikli bağımlı değişken, Lojistik regresyon, Yerine cevaplama.

Volume: 09 Number: 01 Category: 02 Page: 65-71 ISSN No: 1303-6319 Cilt: 09 Sayı: 01 Kategori: 02 Sayfa: 65-71

ÇOKGEN ALANLARDA İKİ DEĞİŞKENLİ BİRİKİMLİ DAĞILIM FONKSİYONUNUN BULUNMASI

Orhan KESEMEN * **Fatma Zehra DOĞRU** **

ÖZET

İki değişkenli olasılık yoğunluk fonksiyonundan birikimli dağılım fonksiyonunu hesaplamak için genellikle dikdörtgen alan kullanılır. Ancak uygulamada dikdörtgen olmayan birçok alan mevcuttur. Bu çalışmada, bu alanlar çokgenlerle yaklaşım yapılarak hesaplandı. Hesaplamada iki tür yöntem kullanıldı. İlk yöntem sürekli fonksiyonlar için geliştirildi, ancak bu yöntem yalnızca düzgün dağılım için uygulandı. İkinci yöntem ise ayrık fonksiyonlar için geliştirildi ve herhangi bir olasılık yoğunluk fonksiyonu için kullanılabilir bir yöntemdir.

Anahtar Kelimeler: Birikimli dağılım fonksiyonu, Çokgen tabanlı olasılık yoğunluk fonksiyonu, İki değişkenli dağılımlar, Kapalı bölgede iki değişkenli dağılım fonksiyonları.

1. GİRİŞ

Olasılık fonksiyonları uygulamada genellikle tek değişkenli bir fonksiyon olarak kullanılmasına rağmen, iki değişkenli kullanımları da bulunmaktadır (Martinez vd., 2002). İki değişkenli olasılık fonksiyonlarının uygulama alanları olarak bir göletteki balık popülasyonu, bir şehirdeki kirlilik oranı, bir ormandaki bir ağaç türünün yoğunluğu, bir bölgedeki trafik akışı, bir hava sahasındaki uçuş yoğunluğu, bir bölgedeki yaban hayatın çeşitliliği, bir şehirdeki suç işleme oranı vb. verilebilir. İki değişkenli olasılık yoğunluk fonksiyonları genelde dikdörtgen şekilli veya fonksiyonu bilinen bir alan içerisinde incelenmektedir (Whitt, 1976; Kay, 2006). Ancak gerçek yaşamda olasılık yoğunluk fonksiyonuna temel teşkil eden alanların istenen şekilde olması olasılığı oldukça düşüktür. Bu alanlar değişik geometrik şekillerde olabilmektedir. Bu çalışmada her türlü şekle sahip alanlar, hesaplamada kolaylık olması açısından çokgen yaklaşımıyla gösterilir. Bu gösterimden yola çıkarak elde edilen çokgen alanların (kapalı alan) dağılım fonksiyonlarının hesaplanmasında geometrik yaklaşımlar kullanılmıştır.

2. YÖNTEM

Sınırlı bir bölgede iki değişkenli olasılık yoğunluk fonksiyonu, genelde sınırları belli dörtgen alanlar içerisinde tanımlanmaktadır. Bu iki değişken bağımsız olması durumunda her birinin olasılık yoğunluk fonksiyonlarının çarpımı biçiminde yazıldığı bilinmektedir (Yates vd., 2005). İki değişkenli birikimli dağılım fonksiyonu, iki değişkenin karma olması durumunda çift katlı integralle hesaplanabilmektedir.

*Yrd. Doç. Dr., Karadeniz Teknik Üniversitesi, Fen Fakültesi, İstatistik ve Bilgisayar Bilimleri Bölümü, e-posta: okesemen@gmail.com

**Karadeniz Teknik Üniversitesi, Fen Fakültesi, İstatistik ve Bilgisayar Bilimleri Bölümü, e-posta: fatmazehradogru@gmail.com

Bu integral,

$$F(u,v) = \int_{-\infty}^{v}\int_{-\infty}^{u} f(x,y)\,dx\,dy \tag{1}$$

biçiminde gösterilmektedir. Burada $f(x,y)$ iki değişkenli olasılık yoğunluk fonksiyonudur. Öte yandan bu fonksiyon farklı koordinat sistemlerine çevrilerek farklı geometrik sınırlamalar için hesaplanabilir (Nelsen, 1993).

2.1. Düzgün Dağılımlı Çokgen Alanlarda Dağılım Fonksiyonu

Bu fonksiyonun düzgün dağılıma sahip olduğu varsayıldığında $x \in [a,b]$ ve y $\in [e,f]$ biçiminde sınırlandırılmış bir bölge için birikimli dağılım fonksiyonu,

$$F(u,v) = c\int_{e}^{v}\int_{\Omega}^{u} dx\,dy \tag{2}$$

eşitliğiyle verilir. Ancak alan dikdörtgen şekilli değilse integralin değerini bulmak sorun olacaktır. Olasılık yoğunluk fonksiyonun tanımlı ($f(x,y) \geq 0$) olduğu bölge N köşeden $\Omega = \{p_i = (x_i, y_i), i = 1,2,\dots,N\}$ oluşan bir çokgen alan Şekil 1(a) ise bu alan üzerinden integralin değeri,

$$F(u,v) = c\iint_{\Omega} dx\,dy = 1 \tag{3}$$

biçiminde verilir. Burada integralin değerinin ikinci olasılık aksiyomuna ($P(\Omega) = 1$) göre 1'e eşit olması için c ölçek değerinin bulunması gerekir. Kapalı çokgen alanın hesaplanması için çokgen alan yöntemine göre Ω kapalı çokgen alanın değeri,

$$A_{\Omega} = \frac{1}{2}\left|\sum_{i=1}^{N}(x_{i+1}\,y_i - x_i y_{i+1})\right| \tag{4}$$

eşitliğiyle bulunur (Wikipedia, 2011). Burada $i = \{1, \quad i+1 > N$ biçiminde kullanılır. Bu alan değeri eşitlik (2)'de verilen kapalı integral yerine konur, c değişkeni yalnız bırakılırsa,

$$c = \frac{1}{A_{\Omega}} = \frac{2}{\left|\sum_{i=1}^{N}(x_{i+1}\,y_i - x_i y_{i+1})\right|} \tag{5}$$

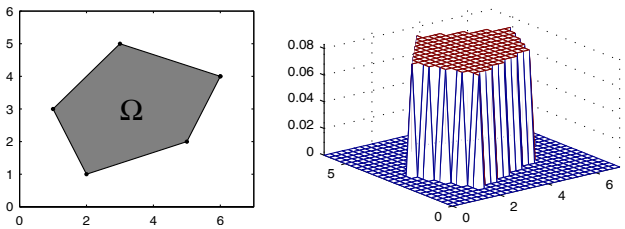
eşitliği elde edilir. Burada,

$$c = f(x,y) = \begin{cases} \dfrac{1}{A_{\Omega}}, & x,y \in \Omega \\ 0, & x,y \notin \Omega \end{cases} \tag{6}$$

parçalı fonksiyon ile tanımlanmaktadır. Ω kapalı çokgen alanının olasılık yoğunluk fonksiyonun üç boyutlu grafiği Şekil 1(b)'de verilmektedir. (5) eşitliği (2) eşitliğinde yerine konur sınırlar düzenlenirse,

$$F(u,v)=\frac{2}{\left|\sum_{i=1}^{N}(x_{i+1}\,y_i-x_i\,y_{i+1})\right|}\int_{-\infty}^{v}\int_{-\infty}^{u}dx\,dy \tag{7}$$

eşitliği elde edilir. Burada, olasılık dağılım fonksiyonu, $y \le v$ ile $x \le u$ ifadeleri ile tanımlanan bölge ile çokgenin kesişiminden oluşan ara kesit bölgesinin (Q) alanı yardımıyla bulunabilir. Ara kesit bölgesi $Q_{u,v}=\{q_j=(\hat{x}_j,\hat{y}_j), j=1,2,\ldots,M\}$ çokgeni q_i köşe noktaları yardımıyla yeniden tanımlanır.



Şekil 1. Olasılık yoğunluk fonksiyonları; (a) olasılık yoğunluk fonksiyonun sınırlı olduğu alan; (b) üç boyutlu olasılık yoğunluk fonksiyonu.

Yeni ara kesit bölgesinin dağılım fonksiyonu aşağıdaki eşitlikte gösterildiği gibi Q bölgesinin alanının Ω bölgesinin alanına oranıyla hesaplanır,

$$F_Q(u,v)=\frac{A_Q}{A_\Omega}=\frac{\left|\sum_{i=1}^{M}(\hat{x}_{i+1}\,\hat{y}_i-\hat{x}_i\,\hat{y}_{i+1})\right|}{\left|\sum_{i=1}^{N}(x_{i+1}\,y_i-x_i\,y_{i+1})\right|}. \tag{8}$$

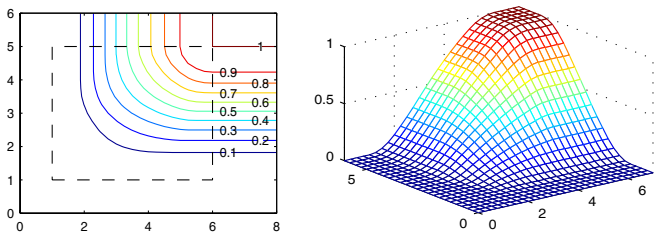
Bu dağılım fonksiyonun değişik bölgelerdeki değerleri,

$$F(u,v)=\begin{cases}0, & u<x_{min}\ \vee\ v<y_{min}\\ F_Q(u,v), & x_{min}\le u\le x_{max}\ \wedge\ y_{min}\le v\le y_{max}\\ 1, & u>x_{max}\ \wedge\ v>y_{max}\end{cases} \tag{9}$$

biçiminde verilmektedir. Burada $\wedge$ simgesi mantıksal VE, $\vee$ simgesi ise mantıksal VEYA işlecine karşılık gelmektedir. (9) eşitliğine göre değişim bölgesinin olasılık dağılım fonksiyonu hesaplamak için (7) eşitliğindeki sınırlar aşağıdaki gibi yeniden düzenlenirse,

$$F(u,v)=\frac{2}{\left|\sum_{i=1}^{N}(x_{i+1}\,y_i-x_i\,y_{i+1})\right|}\int_{y_{min}}^{v}\int_{x_{min}}^{u}dx\,dy, \tag{10}$$

eşitliği elde edilir. Bu eşitlikte $v\in[y_{min},y_{max}]$ ve $u\in[x_{min},x_{max}]$ verilen aralıklarında hesaplanır. Diğer bölgelerin değişimi ise Şekil 2 ve (9) eşitliğinde verildiği gibidir.

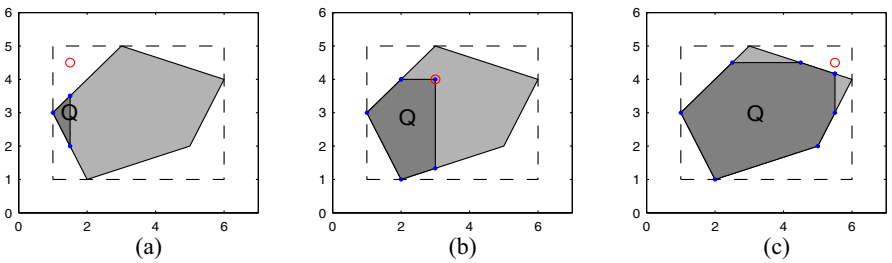


Şekil 2. Birikimli dağılım fonksiyonları; (a) eş yükselti eğrileri ile birikimli dağılım fonksiyonu; (b) üç boyutlu birikimli dağılım fonksiyonu.

Şekil 2(a)'da verilen eş yükselti grafiği birikimli dağılım fonksiyonun grafiğini göstermektedir. Burada kesikli çizgiyle gösterilen alan hesaplama bölgesi olarak tanımlanmaktadır. Bu bölgenin solunda, altında ve sol-altındaki dağılım fonksiyonu değerleri 0'dır. Üst ve sağ tarafındaki değerler için, hesaplama bölgesine en yakın değer alınarak o noktanın dağılım fonksiyonunun değeri hesaplanabilir. Bu hesaplama bir çeşit değerlerin yinelenmesi tekniği olarak kullanılabilir. Hesaplama bölgesinin sağ-üst köşesinde tüm değerler 1'e eşittir.

2.2. Q Bölgesinin Tanımlanması

Dağılım fonksiyonun istenen aralıktaki her (u, v) noktası için Q bölgesinin yeniden tanımlanması gerekir. Q bölgesi, Ω bölgesinin bir alt bölgesi olduğundan, $x \leq u$ ve $y \leq v$ bölgeleri ile Ω bölgesinin kesişim bölgesi olarak tanımlanır. Q bölgesinin değişik durumlardaki görünümü Şekil 3'te verilmektedir. Buradaki yuvarlak nokta birikimli dağılım fonksiyonun hesaplandığı (u, v) noktasını göstermektedir. Diğer küçük noktalar ise Q bölgesinin köşe konumlarını vermektedir. Yuvarlak noktalar çokgen alan dışındayken bu noktalar Q bölgesinin köşe konumlarına dahil edilmez.



Şekil 3. Q bölgesinin hesaplanması, (a) kesilen doğrular yalnızca bir noktadan kesilmesi ve (u, v) noktasının çokgenin dışında olduğu durum; (b) (u, v) noktası çokgenin içinde olduğunda; (c) (u, v) noktasının çokgenin dışında olması ve bazı kenarları iki noktadan kesmesi.

Çokgen alan hesaplama yöntemi, çokgenin noktalarının sırasıyla yazılması koşuluyla hesaplanır. Bu sıralama ister saat yönünde, isterse saatin tersi yönünde hesaplanabilir (Beyer, 1987). Bu çalışmada hesaplamalar saatin tersi yönünde gerçekleştirilmiştir. Q bölgesinin köşe noktalarının belirlenmesi için geliştirilen yöntem Algoritma 1'de verilmiştir.

Algoritma 1. Q bölgesinin köşe noktalarının belirlenmesi

Adım 1. Ω bölgesinin i inci noktası Q bölgesinin köşe noktalarına eklenir,

Adım 2. $x = u$ ve $y = v$ doğrularının (p_i, p_{i+1}) doğru parçasını kestiği noktalar bulunur. Eğer,

- Kesilen nokta sayısı bir tane ise, kesişim noktası Q bölgesine köşe diye eklenir.
- Kesilen nokta sayısı iki tane ise, kesişim noktalarından hangisi p_i noktasına yakınsa öncelikle o nokta Q bölgesi köşesi diye eklenir, sonra diğer nokta eklenir.

Adım 3. Son doğru parçasına gelindiyse 5. adıma geçilir. Eğer değilse, Ω bölgesini bir sonraki doğru parçasına geçilerek, 1. adıma gidilir ve işlemlere devam edilir.

Adım 4. Ω bölgesinin tüm köşe noktaları ve bu köşe noktalarının oluşturduğu tüm doğru parçaları ile $x = u$ ve $y = v$ doğrularının tüm kesişim noktaları sırasıyla Q bölgesinin köşe noktaları diye eklenir. Sonra, (u, v) noktası Ω bölgesinin içindeyse bu noktada son olarak Q bölgesine köşe diye eklenir.

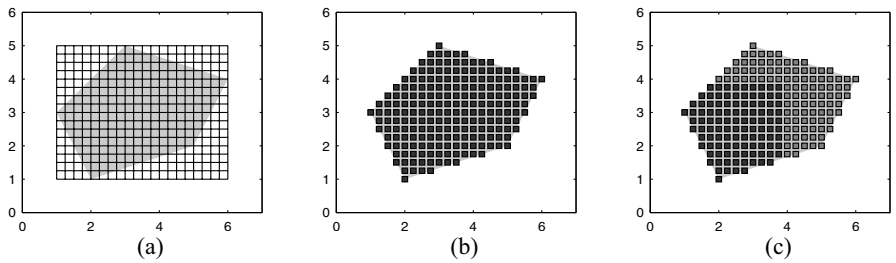
Adım 5. Q bölgesini tanımlayan tüm noktalardan $x_j > u$ veya $y_j > v$ koşulunu sağlayanlar köşe noktaları dizisinden silinir. Geriye kalan tüm noktalar Q bölgesinin kesin köşe noktalarıdır.

2.3. Çokgen Alanlarda Düzgün Olmayan Dağılım Fonksiyonu

Çokgen içindeki olasılık yoğunluk fonksiyonu düzgün olmayan bir dağılıma sahipse hesaplama için sayısal yöntemler (sonlu elemanlar) kullanılır. Sonlu elemanlar yöntemi için dikdörtgen ızgara düzeneği kullanılarak dağılım fonksiyonu,

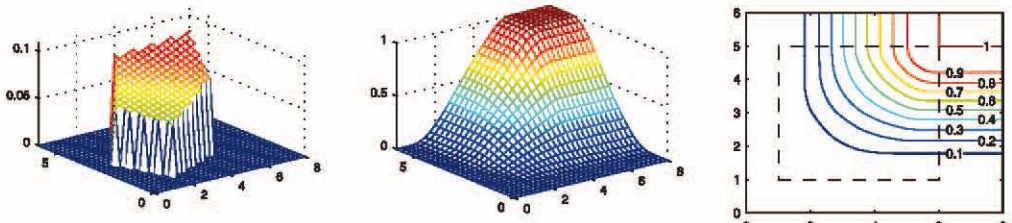
$$F(u,v) = (\Delta x \cdot \Delta y) \sum_{i=1}^{x_i \le u} \sum_{j=1}^{y_j \le v} f(x_i, y_j), \qquad \{f(x_i, y_j) > 0, \quad x_i, y_i \in \Omega\} \tag{11}$$

biçiminde hesaplanır. Burada $x_i = x_{min} + i\Delta x$ ve $y_i = y_{min} + j\Delta y$ olarak verilmektedir. Δx ve Δy Şekil 4'teki ızgara çizgileriyle gösterilen alanın x ve y yönündeki birim alan ızgara genişliğidir.



Şekil 4. Hesaplama bölgesinin ızgaraya bölümlemesi, (a) Ω bölgesinin sonlu elemanlara bölümlemesi; (b) Ω bölgesinin kare biçimli sonlu elemanlara bölümlemesi; (c) Q bölgesi $x < 4$ ve $y < 4$ için hesaplanacak sonlu elemanlar koyu kareler olarak gösterilmiştir.

Şekil 4(c)'de verilen her bir sonlu eleman bir dikdörtgen prizma yaklaşımıyla ele alınır. Prizmada her elemanın orta noktasının olasılık yoğunluk fonksiyonu Şekil 5(a) tüm prizma yüzeyinin olasılık yoğunluk fonksiyonu olarak alınıp prizmanın yüksekliği bulunur. Prizmanın yüksekliği ile prizma hacmi hesaplanıp birikimli dağılım fonksiyonu hesaplanır Şekil 5(b), (c).



Şekil 5. Çokgen alanlarda düzgün olmayan dağılım fonksiyonu, (a) olasılık yoğunluk fonksiyonu; (b) birikimli dağılım fonksiyonu; (c) birikimli dağılım fonksiyonun eş yükselti grafiği.

3. SONUÇLAR

Bu çalışmada değişkenlerin sınırlandığı alanın keyfi bir çokgen olduğu zaman dağılım fonksiyonunun bulunması için kullanışlı bir yöntem önerildi. Çokgen içindeki iki değişkenli olasılık yoğunluk fonksiyonun düzgün dağılama sahip durumunda birikimli dağılım fonksiyonu sürekli bir fonksiyon olarak hesaplanabilmektedir. Ancak çokgen alan ile sınırlandırılmış bölgedeki olasılık yoğunluk fonksiyonu düzgün olmayan bir dağılıma sahip olduğunda birikimli dağılım fonksiyonunun hesaplanması için sonlu elemanlar yöntemi önerilmiştir. Bu yöntem herhangi iki rastgele değişkenin belirli bir noktadan itibaren istenen dağılım fonksiyonunu bulmaktadır. Dolayısıyla uygulamada hesaplanamayan dağılımlar bu yöntem ile kolayca elde edilebilmektedir.

4. KAYNAKLAR

Martinez, W. L., Martinez, A. R., 2002. Computational Statistics Handbook with MATLAB, Chapman & Hall/Crc, NY.

Whitt, W., 1976. Bivariate Distributions with Given Marginals, The Annals of Statistics, 6 (4) 1280-1289.

Kay, S. M., 2006. Intuitive Probability and Random Processes Using Matlab®, Springer, NY.

Yates, R. D. and Goodman, D. J., 2005. Probability and Stochastic Processes, John Wiley & Sons, Inc., USA.

Nelsen, R. B., 1993. Some Concepts of Bivariate Symmetry, Journal of Nonparametric Statistics, 3(1), 95-101.

Wikipedia, 2011. Polygon, http://en.wikipedia.org/wiki/Polygon, 3 April 2011.

Beyer, W. H., 1987. CRC Standard Mathematical Tables, 28th ed. Boca Raton, FL: CRC Press, p. 123.

FINDING CUMULATIVE DISTRIBUTION FUNCTIONS OF TWO VARIABLES IN POLYGONAL AREAS

ABSTRACT

In general, rectangular area is used to calculate cumulative distribution function from bivariate probability density function. However, in practice a many areas are available that aren't rectangular. In this study, these areas were calculated by polygons approach. Two types of methods were used in the calculation. First method was developed for continuous functions; however this method was applied only for uniform distribution. The second method was developed for discrete functions and can be used for any probability density function.

Keywords: Cumulative distribution function, Polygonal based probability density function, Bivariate distributions, Bivariate distributions functions over bounded domain.

Cilt: 09 Sayı: 01 Kategori: 05 Sayfa: 72-80 ISSN No: 1303-6319 Volume: 09 Number: 01 Category: 05 Page: 72-80

DEMODEX SAYISININ ÇEŞİTLİ DEĞİŞKENLERE GÖRE TANIMLANMASINDA SIFIR AĞIRLIKLI VE HURDLE REGRESYON MODELLERİNİN İNCELENMESİ

Esra PAMUKÇU[*] Cemil ÇOLAK[**] Sinan ÇALIK[***] Ülkü KARAMAN[****]

ÖZET

Bu çalışmada, demodex sayısının çeşitli değişkenlere göre tanımlanmasında sıfır ağırlıklı regresyon modelleri ile hurdle regresyon modellerinin incelenmesi ve uyum ölçütleri kullanılarak, portör muayenesi için gelen hastalardan elde edilen bilgiler ile demodex sayısının yaşa, cinsiyete ve mesleğe göre etkilerinin tanımlanmasında en iyi sonuçları verecek modelin belirlenmesi amaçlanmıştır. Çalışmadaki veriler, Haziran 2007-Haziran 2009 tarihleri arasında Malatya Halk Sağlığı laboratuarına portör bakısı için gelen 156 kişiyi kapsamaktadır. Demodex sayısı cevap değişkeni; hastaların yaşı, cinsiyeti ve mesleği açıklayıcı değişkenler olarak belirlenmiştir. İstatistiksel analizde R 2.11.1 yazılım programı kullanılmıştır. Cevap değişkeni olarak ele alınan demodex sayısının %62.8'i sıfır değerli olduğu için sıfır ağırlıklı ve hurdle regresyon modelleri kullanılmıştır. Uygulama sonucunda modellerin birbirlerine göre üstünlüklerini belirlemede kullanılan AIC değerleri ZIP için 731.18, ZAP için 731.49, ZINB için 531.73 ve ZANB için 531.11 olarak elde edilmiştir. En düşük AIC değerine sahip olan model, ZANB olarak elde edilmiştir. Sıfır ağırlıklı ve hurdle regresyon modellerinde, sıfır değerlerinin önemli bir etkiye sahip olup olmadığı test edilmektedir. Eğer sıfır değerlerinin etkisi önemli değilse, bu durumda sıfır ağırlıklı regresyon modelleri ve hurdle regresyon modellerinin sonuçları ile Poisson regresyonu ve negatif binomiyal regresyon analiz sonuçları benzer olacaklardır. Bu çalışmada veri setinde var olan sıfır değerlerinin önemli bir etkiye sahip oldukları belirlenmiştir.

Anahtar Kelimeler: Demodex, Hurdle modeller, Sıfır ağırlıklı regresyon modelleri.

1. GİRİŞ

Demodex sp, Arachnida sınıfının Prostigmata takımının Demodicidae ailesinden bir akar olup pilosebase (kıl+yağ bezi) ünitelerde bulunmaktadır (Karaman vd., 2008, 343; Karaman vd., 2008, 5). *Demodex* türleri arasından *D. folliculorum* ve *D. brevis*'in insanda en sık rastlanan kalıcı ektoparazit olup çoğunlukla yüzdeki kıl folikülünde bulunan pilosebase ünitelerine yerleştiği tespit edilmiştir. *D. folliculorum*'un foliküler açıklıklarda tek veya gruplar halinde yaşadığı, *D. brevis*'in ise sebaseöz bezlerinin derinliklerinde tek olarak yaşadığı ve akarların ince uzun yapılarının bu yerlere uygun olduğu belirtilmiştir (Karaman vd., 2008, 343; Wesolowska et al., 2005).

[*]Arş. Gör., Fırat Üniversitesi Fen Fakültesi İstatistik Bölümü, e-posta: esra_pamukcu@hotmail.com
[**]Doç. Dr., İnönü Üniversitesi Tıp Fakültesi Biyoistatistik Anabilim dalı, e-posta: cemilcolak@yahoo.com
[***]Yrd. Doç. Dr., Fırat Üniversitesi Fen Fakültesi İstatistik Bölümü, e-posta: scalik@firat.edu.tr
[****]Yrd. Doç. Dr., Ordu Üniversitesi Sağlık Yüksekoukulu, e-posta: ulkukaraman44@hotmail.com

Bu makalenin özeti, poster çalışması olarak 7. Uluslararası İstatistik Kongresi'nde tebliğ edilmiştir.

Ayrıca *Demodex* türleri insanda yanak, burun, kirpik dipleri, çene, alın, dış kulak yolu, meme ucu, sırt, penis ve kalça gibi vücudun değişik bölgelerinde de rapor edilmiştir (Karaman vd., 2008, 343; Karaman vd., 2008, 5). İnsandan insana yakın temasla bulaşarak rosacea, akne vulgaris, perioral dermatit, seboreik dermatit, mikropapüler-kaşıntılı dermatit ve blefarit gibi rahatsızlıkların sebebi ve gelişiminde önemli rol oynadıkları araştırıcılar tarafından bildirilmiştir (Sheals, 1973; Mathieu and Wilson, 2000; Markell et al., 1992). Parazitin pilosebase foliküllerde bulunmasını zararsız olarak değerlendirenlerin yanı sıra yüzde oluşan bazı deri hastalıklarının sebebi ve gelişiminde bu parazitin rolü olduğunu savunanlar da bulunmaktadır (Morsy et al., 2000; Pena and Andrade Filho, 2000; Baima and Sticterling 2002).

Histolojik incelemelerde parazitin mononükleer, perifoliküler inflamatuar infiltrasyona neden olabileceği bildirilmiştir. Oluşan infiltrasyonun CD4+ T lenfositlerinden ve CD8+ T hücrelerinden oluştuğu gözlenmiştir. Ayrıca enfeste folikül çevresinde CD1a+ makrofajlar bulunduğu saptanmıştır (Baima and Sticterling 2002). Yine Volmer (1996) *demodex* içeren folliküllerin %83'ünde folikülit tanımlamıştır.

Tanısında da selofan bant, deri kazıntısı, punch biyopsisi ve standart yüzeysel deri biyopsisi (SYDB) gibi yöntemler kullanılmaktadır (Erbağcı ve Özgöztaşı, 1998; Marks and Dawber, 1971). *Demodex* parazitliğinin bütün dünyada yaygın olduğu, ırk ve cinsiyet farkı göstermediği ancak parazitliğinin yaşla doğru orantılı olarak arttığı bildirilmiştir (Özçelik, 1997; Yazar vd., 2008). Parazitin hareketinin saatte ortalama 8-16 cm yol alabileceği ve bütün dönemlerinin negatif fototaksik reaksiyon gösterdiği bildirilmiştir. Isı ve kuruluğa karşı ise dirençlerinin az olduğu saptanmıştır (Clifford and William, 1972).

Bu çalışmada, parazitliğin yaşla birlikte artması ve insandan insana yakın temasla bulaşabilmesi nedeniyle, yaş ve meslek gruplarının bulaşımda etkili olabileceği düşünülmüştür. Bu nedenle sağlıklı, herhangi bir sağlık problemi görülmeyen sıhhı ve gayri sıhhı müessese yönetici ve çalışanların portör muayeneleri ile birlikte yüzlerinde de *demodex* türlerinin varlığının; meslek grupları, yaş ve cinsiyet arasındaki etkilerinin tanımlanmasında en uygun regresyon modelinin tespit edilmesi amaçlanmıştır.

2. GEREÇ VE YÖNTEM

2.1. Sıfır Ağırlıklı Poisson (ZIP) ve Sıfır Ağırlıklı Negatif Binomiyal Regresyon (ZINB) Modelleri

Poisson regresyon analizi sayıma dayalı veri setlerini analiz etmek için standart bir çerçeve oluşturmaktadır. Bununla birlikte uygulamalarda sayma verileri poisson dağılımına göre aşırı yayılım göstermektedirler. Aşırı yayılımın sık bir tezahürü ise poisson dağılımı için beklenenden daha fazla sıfır değerlerinin var olmasıdır (Ridout et al., 1998).

Veri setinde var olan sıfırlar iki şekilde ortaya çıkabilmektedirler. Bu çalışmanın cevap değişkeni olan *demodex* sayısının veri yapısı incelendiğinde, %62.8'inin sıfır değerli olduğu tespit edilmiştir. Burada portör bakısı alınan kişiler ya gerçekten hasta değillerdir ya da hastalığa karşı bir dirence sahiplerdir. Her iki durumda da *demodex* sayısı sıfır çıkabilmektedir. Bu nedenle sıfır değerleri, kaçınılmaz bir şekilde ortaya çıkan yapısal sıfırlar ve tesadüfi olarak ortaya çıkan örneksel sıfırlar olarak ikiye ayrılmaktadır.

Lambert (1992) tarafından tanımlanan ZIP regresyonunda yapısal sıfırların, sıfır olma olasılığını ifade eden bir φ parametresi ile Bernoulli sürecini takip ettiği ve rasgele sayıların da λ parametreli bir Poisson sürecini takip ettiği varsayılır. Bu durumda aşırı sıfır değerine sahip olan bir Y cevap değişkeni için ZIP dağılımı

$$Y_i \sim \begin{cases} 0 & ,\varphi \text{ olasılığı ile} \\ Poisson\,(\lambda_i) & ,1-\varphi \text{ olasılığı ile} \end{cases} \quad (1)$$

$$Y_i \sim \begin{cases} \varphi + (1-\varphi)e^{-\lambda} & ,y = 0 \\ (1-\varphi)\frac{\lambda^y}{y!}e^{-\lambda} & ,y > 0 \end{cases} \quad (2)$$

şeklindedir.

Sıfır ağırlıklı poisson regresyon modellerinde, Poisson sürecini takip eden rasgele sayılar için aşırı yayılım olması durumunda, Greene (1994) tarafından geliştirilen sıfır ağırlıklı negatif binomial regresyon (zero-inflated negative binomial regression ZINB) yöntemi kullanılabilmektedir. ZINB regresyonunda yapısal sıfırların φ parametresi ile bir Bernoulli sürecini takip ettiği ve rasgele sayılarında μ ve k parametreli negatif binomial dağılımı takip ettiği varsayılır. Bu durumda aşırı sıfır değerine sahip olan bir Y cevap değişkeni için ZINB dağılımı

$$Y_i \sim \begin{cases} 0 & ,\varphi \text{ olasılığı ile} \\ Binomial\,(\mu_i, k_i) & ,1-\varphi \text{ olasılığı ile} \end{cases} \quad (3)$$

$$Y_i \sim \begin{cases} \varphi + (1-\varphi)\left(\frac{k}{k+\mu}\right)^k & ,y = 0 \\ (1-\varphi)\binom{y+k-1}{y}\left(\frac{\mu}{\mu+k}\right)^y\left(\frac{k}{k+\mu}\right)^k & ,y = k \end{cases} \quad (4)$$

şeklindedir (Yee, 2008; Hall, 2000).

2.2. Hurdle Poisson (ZAP) ve Hurdle Negatif Binomial (ZANB) Regresyon Modelleri

Hurdle modeller, sıfır değerlerinin çok olduğu veri kümelerinin analizinde kullanılabilen diğer bir yöntemdir. Hurdle model iki kısımdan oluşmaktadır. Birinci kısım; sıfır değerleri (0) ve pozitif değerleri (1) olarak gösteren binary cevapların oluşturduğu binomial olasılık modelidir. Pozitif sonuçları tanımlayan kesilmiş sayıma dayalı veriler ise ikinci kısım olarak modellenmektedir. Bu kısım, poisson dağılımı kullanılarak modellendiğinde Hurdle Poisson model (ZAP), negatif binomial dağılım kullanılarak modellendiğinde ise Hurdle Negatif binomial model (ZANB) olarak adlandırılmaktadır. $Y_i, i = 1,2,\ldots,n$ için birbirinden bağımsız sayıma dayalı olarak elde edilen bağımlı değişken için gözlem değerleri olmak üzere ZAP dağılımı

$$Y_i \sim \begin{cases} \varphi & ,y = 0 \\ (1-\varphi)\dfrac{e^{-\lambda}\lambda^y}{y!} & ,y > 0 \end{cases} \quad (5)$$

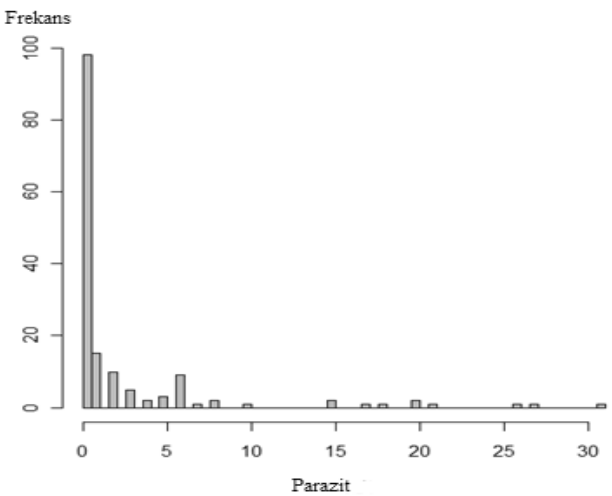
şeklindedir. Y_i, φ olasılığı ile sıfır değerini alırken, $1-\varphi$ olasılığı ile pozitif değerleri alır ve bu durumda $Y_i \sim truncated Poisson\ (\lambda)$ dağılımına sahiptir. Heilbron (94) tarafından gösterilen bu formül sıfır sonuçlarının olasılıklarını artırır ve tüm olasılıklar toplamı 1 olacak şekilde kalan olasılıkları hesaplar. ZANB dağılımı ise

$$Y_i \sim \begin{cases} \varphi & ,y = 0 \\ \dfrac{(1-\varphi)}{1-\left(\frac{k}{k+\mu}\right)^k}\binom{y+k-1}{y}\left(\frac{\mu}{\mu+k}\right)^y\left(\frac{k}{k+\mu}\right)^k & ,y > 0 \end{cases} \quad (6)$$

şeklindedir. Y_i, φ olasılığı ile sıfır değerini alırken, $1-\varphi$ olasılığı ile pozitif değerleri alır ve bu durumda $Y_i \sim truncated Negbin(\mu, k)$ dağılımına sahiptir (McDowell, 2003; Dalrymple et al., 2003).

3. BULGULAR VE TARTIŞMA

Çalışmadaki veriler, Haziran 2007-Haziran 2009 tarihleri arasında Malatya Halk Sağlığı laboratuvarına portör bakısı için gelen 156 kişiyi kapsamaktadır *Demodex* sayısı cevap değişkeni Y, hastaların yaşı X_1, cinsiyeti X_2 ve mesleği X_3 açıklayıcı değişkenler olarak belirlenmiştir. Meslekler, unlu mamuller sektörü (fırıncı, pideci, pastacı), hizmet sektörü (turizm, bar, restoran çalışanları, garsonlar), et ürünleri sektörü (kasap, beyaz et üreticisi, balıkçı) ve gıda perakende sektörü (bakkal, market, vb.) şeklinde alt gruplara ayrılmıştır. Cevap değişkeni olarak ele alınan *demodex* sayısının %62.8'i sıfır değerlidir. Bu durum aşağıdaki histogram grafiğinde de görülmektedir . İstatistiksel analizde R 2.11.1 yazılım programı kullanılmıştır.



Şekil 1. Demodex parazit sayısının frekans dağılımı

Veri setini modelleyebilmek için uygulanan analizler ve sonuçları aşağıdaki tablolarda verilmiştir. Öncelikle sayıma dayalı veri setini modelleyebilmek için, standart prosedür olarak kullanılan poisson regresyonu kullanılmıştır. Analiz sonucunda açıklayıcı değişkenlerden yaş, cinsiyet kadın, meslek hizmet, et ve gıda sektörü değişkenleri anlamlı değişkenler olarak elde edilmiştir. Ancak poisson modelinin veri setine uygunluğunun bir göstergesi olarak aşırı yayılım parametresi hesaplandığında Φ=10.93 olduğu belirlenmiştir. Poisson regresyonunda varyansın ortalamadan büyük çıkması olarak tanımlayabileceğimiz aşırı yayılım problemi için genellikle gözlenemeyen heterojenlik, gözlemler arasındaki korelasyon, zamana bağlı gözlemler arasındaki bağımlılık veya veri setinde beklenenden fazla sıfır değerleri neden olarak gösterilebilir (Böhning et al., 1999).

Tablo 1. Poisson modeller için parametre tahminleri

Değişkenler	Poisson modeller				
	PR	ZIP		ZAP	
		Lojit	Poisson	Lojit	Poisson
Sabit	-0.26(0.254)	2.02(0.822)*	1.43(0.25)***	-1.979(0.795)*	1.428(0.251)***
X_1	0.023(0.005)***	-0.049(0.02)*	-0.002(0.005)	0.047(0.019)*	-0.002(0.005)
X_2 (kadın)	0.383(0.116)***	0.394(0.396)	0.683(0.120)***	-0.353(0.394)	0.681(0.12)***
X_2(erkek)	-	-	-	-	
X_3 (un)	-	-	-	-	
X_3(hizmet)	0.355(0.162)*	-0.035(0.505)	0.414(0.165)*	0.054(0.499)	0.412(0.165)*
X_3(Et)	0.601(0.188)**	-0.139(0.616)	0.636(0.198)**	0.18(0.61)	0.631(0.197)**
X_3 (Gıda)	-0.551(0.218)*	0.158(0.593)	-0.402(0.232) .	-0.24(0.576)	-0.396(0.231) .
AIC ve Φ	1247.4 10.93	731.18	2.23	731.49	2.23

p: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Cevap değişkeni olarak ele alınan *demodex* sayısının %62.8'i sıfır değerli olduğu için veri setine sıfır ağırlıklı poisson regresyon modellerinin uygulanmasına karar verilmiştir. ZIP regresyon yönteminin uygulanması sonucunda, Poisson dağılan sıfır değerleri ve sıfırdan büyük sayı değerleri ile yapılan tahminlerde; sabit, cinsiyet kadın, meslek et ürünleri, meslek hizmet sektörü anlamlı değişkenler olarak bulunmuştur. Ancak ekstra sıfır değerleri (yapısal sıfır) ile birlikte binomial dağılım kullanılarak yapılan tahminlerde sadece yaş değişkeni anlamlı bulunmuştur. Yayılım parametresi 2.23 ve AIC uyum kriteri 731.18 olarak elde edilmiştir. Sıfır ağırlıklı regresyon modellerinden bir diğeri olan ZAP regresyonu uygulandığında ise, pozitif sonuçları tanımlayan kesilmiş sayıma dayalı veriler için poisson ile tahmin yapıldığında; sabit, cinsiyet kadın, meslek et ürünleri, meslek hizmet sektörü anlamlı değişkenler olarak bulunmuştur. Sıfır değerleri (0) ve pozitif değerleri (1) olarak gösteren binary cevapların oluşturduğu binomial olasılık modeli ile tahmin yapıldığında ise; sabit ve yaş anlamlı bulunmuştur. Yayılım parametresi 2.23 ve AIC uyum kriteri 731.49 olarak elde edilmiştir. Birden büyük olarak elde edilen yayılım parametresi aşırı yayılım probleminin hala var olduğu göstermiştir.

Aşırı yayılım problemi standart hatanın küçümsenmesine ve regresyon parametrelerinin çıkarsamalarında yanılmalara yol açmaktadır (Long, 1997). Bu problemi giderebilmek için alternatif olarak kullanılan yöntemlerden biri negatif binomiyal regresyon yöntemidir. Bu nedenle analizlerin ikinci aşaması olarak negatif binomial regresyon, sıfır ağırlıklı negatif binomial regresyon, hurdle negatif binomial regresyon yöntemlerinin kullanılmasına karar verilmiştir.

Tablo 2. Negatif binomiyal modeller için parametre tahminleri

Değişkenler	Negatif Binomial Modeller				
	NB	ZINB		ZANB	
		Lojit	Kesik	Lojit	Kesik
Sabit	-0.5(0.895)	1.189(1.429)	0.81(0.986)	-1.979(0.795)*	1.034(1.074)
X_1	0.032(0.022)	-0.067(0.033)*	-0.00007(0.002)	0.047(0.019)*	-0.007(0.027)
X_2 (kadın)	0.496(0.437)	0.804(0.617)	0.729(0.471	-0.353(0.394)	0.718(0.507)
X_2(erkek)	-	-	-	-	-
X_3 (un)	-	-	-	-	-
X_3(hizmet)	0.178(0.569)	0.571(1.088)	0.608(0.557)	0.054(0.498)	0.523(0.602)
X_3(Et)	0.618(0.702)	0.476(1.256)	0.852(0.645)	0.180(0.610)	0.762(0.690)
X_3(Gıda)	-0.574(0.658)	0.32(1.218)	-0.322(0.636)	-0.240(0.575)	-0.410(0.689)
AIC ve Φ	523.2 0.78	531.73	0.96	531.11	0.91

p: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

NBR uygulanması sonucunda, açıklayıcı değişkenlerin hiçbiri anlamlı olarak bulunamazken yayılım parametresi de 0.78 olarak elde edilmiştir. ZINB regresyonu analizi kullanıldığında ise, negatif binomial dağılan sıfır değerleri ve sıfırdan büyük sayı değerleri ile yapılan tahminlerde hiçbir değişken anlamlı bulunamazken, ekstra sıfır değerleri ile birlikte binomial dağılım kullanılarak tahmin yapıldığında yaş değişkeni anlamlı bulunmuştur. Pearson uyum istatistiği modelin negatif binomial dağılım için uygun olduğunu gösterirken (p=0.610), yayılım parametresi 0.96, yani 1'e çok yakın bir değer olarak elde edilmiştir.. Veriye ZANB uygulandığında ise, kesik veriler için negatif binomial ile yapılan tahminlerde hiçbir değişken anlamlı bulunamazken, sıfır değerleri ile tahmin yapıldığında sabit ve yaş değişkenleri anlamlı bulunmuştur. Bu durum yaşla orantılı olarak parazitin görülme yüzdesinin artığını desteklemektedir.

4. SONUÇ VE ÖNERİLER

Demodex türlerinin çocuklarda görülemeyeceği ergenlikten itibaren artarak ileri yaşlarda en yüksek orana ulaşabildiği bildirilmiştir (Sibenge and Gawkrodger, 1992). Parazitin görülmesi ile yaş gruplarının karşılatıldığı çalışmalarda Baysal vd. (1997)'de, 11-15 yaş grubunda 1 (%8,3)'inde, 16-20 yaş grubunun ise 7 (%12.7)' sinde pozitiflik saptamışlardır. Yine Aycan vd. (2007)'de, ≤20 yaş grubunun 5 (%20)'inde, 21 ve üstündeki yaş grubunun ise 92 (%53,5)' sinde *Demodex* pozitifliği bildirmişlerdir. Benzer olarak Kaya vd. (2010), lisede yabancı uyruklu olan yaşları 15 ile 21 (yaş ortalaması: 17.52±1.36) arasında değişen, 347 erkek öğrenci incelemiş ve 9 (%2.7)'unda *Demodex* türleri saptamışlardır. Parazitin görülme oranı 19 yaş ve üzeri öğrencilerde 18 yaş altı öğrencilere göre daha yüksek bulunmakla birlikte, yaş ile parazit görülmesi arasındaki farkın istatistiksel olarak anlamlı olmadığını bildirmişlerdir. Ancak araştırma kapsamındaki 17-25 yaş grubu ve 46-55 yaş grubu bireylerin %90'ının *demodex* görülme durumunun pozitif olması yaşla orantılı olarak parazitin arttığı sonucunu çıkarmıştır. Ayrıca parazitin insandan insana yakın temasla bulaşabilmesi nedeniyle, gayri sıhhi müesselelerde çalışanların portör olabileceği kanısına varılmıştır. Çalışmada Halk Sağlığı Laboratuarına portör bakısı için gelenlerin diğer rutin bakılarla birlikte *demodex* pozitifliğinin de araştırılması gerektiği önerisi sunulmuştur.

Sayıma dayalı olarak elde edilen veri setlerine Poisson regresyonunun uygulanabilirliği, ortalama ve varyansın eşitliği olarak tanımlanan eş yayılım varsayımına bağlıdır. Ancak bir çok uygulamada, veri setleri ortalamayı aşan bir varyansa sahip olabilmekte ve bu durumda aşırı yayılım problemi ortaya çıkmaktadır. Veri setinde beklenenden fazla sıfır değerlerinin varlığı aşırı yayılımın kaynaklarından biri olarak gösterilebilmektedir. Bu tip verilerin analizinde sıfır değerlerini dikkate alarak modelleme yapan, sıfır ağırlıklı regresyon modellerinin veya hurdle regresyon modellerinin kullanımı tavsiye edilebilir. Sıfır ağırlıklı regresyon modellerinde, sıfır değerlerinin önemli bir etkiye sahip olup olmadığı test edilmektedir. Eğer sıfır değerlerinin etkisi önemli değilse, bu durumda sıfır ağırlıklı regresyon modellerinin sonuçları ile Poisson regresyonu ve negatif binomiyal regresyon analiz sonuçları benzer olacaklardır (Yeşilova, 2007). Bu çalışmada veri setinde var olan sıfır değerlerinin önemli bir etkiye sahip oldukları belirlenmiştir. Uygulama sonucunda modellerin birbirlerine göre üstünlüklerini belirlemede kullanılan AIC değerleri ZIP için 731.18, ZAP için 731.49, ZINB için 531.73 ve ZANB için 531.11 olarak elde edilmiştir. En düşük AIC değerine sahip olan model, ZANB olarak elde edilmiştir.

Sonuç olarak, parazitlerin varlığının araştırıldığı ve pozitifliğine etki edebileceği düşünülen parametrelerin karşılaştırıldığı çalışmalarda, cevap değişkeninde var olabilecek sıfır değerlerinin önemli etkiye sahip olabileceklerinin de göz önünde bulundurulması ve sıfır ağırlıklı regresyon modellerinin veya hurdle regresyon modellerinin kullanılabileceği sonucuna varılmıştır.

5. KAYNAKLAR

Aycan, Ö. M., Otlu, G. H., Karaman, Ü., Daldal, N., Atambay, M., 2007. Çeşitli Hasta ve Yaş Gruplarında *Demodex* sp. Görülme Sıklığı. Türkiye Parazitol Derg.. 31(2): 115 -118.

Baima, B., Sticterling, M., 2002. Demodicidosis Revisited. Acta Derm Venereol, 82:3-6.

Baysal, V., Aydemir, M., Yorgancıgil, B., Yıldırım, M., 1997. Akne Vulgaris Etyopatogenezinde *D. folliculorum*'ların Rolünün Araştırılması. Türkiye Parazitol Derg, 21: 265-268.

Böhning, D., Dietz, E., Schlattmann, P., 1999. The Zero Inflated Poisson Model and Decayed, Missing and Filled Teeth Index in Dental Epidemiology. Journal of Royal Statistical Society Series A, 162(2): 195-209.

Clifford, D., William, B. N., 1972. *Demodex folliculorum* (Simon) and *D. brevis* Akbulatova of Man: Redescription and Reevaluation. The Journal of Parasitology, 58, (1): 169-177.

Dalrymple M. L., et.al, 2003. Finite Mixture, Zero Inflated Poisson and Hurdle Models with Application to SIDS. Computational Statistics & Data Analysis, 41: 491-504.

Erbağcı, Z., Özgöztaşı, O., 1998. The significance of *Demodex folliculorum* density in Rosacea. Int J Dermatol, 39:743-745.

Greene, W. H., 1994. Accounting for Excees Zeros and Sample Section in Poisson and Negative Binomial Regression Models. NYU Working Paper No. EC-94-10, NewYork: Stern School of Business, http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1293115##

Hall, D. B., 2000. Zero-Inflated Poisson and Binomial Regression with Random Effects: A Case Study. Biometrics , 56: 1030-1039.

Heilbron, D. C., 1994. Zero-Altered and Other Regression Models for Count Data with Added Zeros. Biometrical Journal, 36(5): 531-547.

Karaman, U., Çelik, T., Çalık, S., Sener, S., Aydın, N. E., Daldal, Ü. N., 2008. Saçlı Deri Biyopsi Örneklerinde demodex spp. Türkiye Parazitoloji Dergisi, 32(4): 343-345.

Karaman, U., Sener, S., Çelik, T., Atambay, M., Aydın, N. E., Daldal, Ü. N., 2008. Derinin Enfeksiyoz ve Benign Durumlarında Histopatolojik Yöntemle *demodex* spp. Araştırılması. Parazitoloji Dergisi,15(1): 5-7.

Kaya, M., Hamamcı, B., Çetinkaya, U., Yaman, O., Yaza, S., 2010. Bir Lisede Öğrenim Gören Yabancı Uyruklu Erkek Öğrencilerde Selofan-Bant Yöntemi ile *Demodex* sp. Araştırılması. Türk Hijyen ve Deneysel Biyoloji Dergisi, 67: 73-7.

Lambert, D., 1992. Zero-Inflated Poisson Regression, With an Application to Defects in Manufacturing. Technometrics February, Vol: 34 (1): 1-14.

Long, J. S. ,1997. Regression Models for Categorical and Limited Dependent Variables, Sage Publications, USA, s.217-249.

Markell, E. K., Voge, M., John, D. T., 1992. Medical Parasitology, 7th ed, W.B Saunders Company, U.S.A. s.348.

Marks, R., Dawber, R. P. R., 1971. Skin Surface Biopsy: An Improved Technique for the Examination of the Horny Layer. Br J Dermatol, 84: 117-123.

Mathieu, E. M., Wilson, B. B., 2000. Mites (Including Chiggers). Mandell, Douglas and Bennett's Principles and Practice of Infectious Diseases. L. M. Gerald, E. B. John, D. Raphel (eds.), 50 th. ed. Vol: II., U.S.A. s.2980.

McDowell, A., 2003. From The Help Desk: Hurdle Model. The Stata Journal, 3(2):178-184.

Morsy, T. A., Fayad, M. E., Morsy, A. T., Afify, E. M., 2000. Demodex Folliculorum Causing Pathological in Immunocompetent Children. J Egypt Soc Parasitol, 30: 851-4.

Özçelik, S., 1997. Allerjik ve Dermatit Nedeni Olabilen Akarlar, Parazitoloji'de Arthropod Hastalıkları ve Vektörler. M. A. Özcel, N. Daldal (eds.), Türkiye Parazitoloji Derneği Yayınları No: 13, s.349-353.

Pena, G. P., Andrade Filho, J. S., 2000. Is Demodex Folliculorum Really Non-Pathogenic?. Rev Inst Med Trop Sao Paulo, 42:171-3.

Ridout, M., Demetrio, C. G. B., Hinde, J., 1998. Models for count data with many zeros. International Biometric Conference, Cape Town, http://www.kent.ac.uk/IMS/personal/msr/webfiles/zip/ibc_fin.pdf

Sheals, J.G., 1973. Arachnida. Insects and Another Arthropods of Medical Importance. The Trustees of the British Museum (Natural History). Smith K.G.V., (ed.) London. s:17, 462.

Sibenge, S., Gawkrodger, D. J., 1992. Rosacea: A Study of Clinical Patterns, Blood Flow, and the Role of D. Folliculorum. J Am Acad Dermatol., 26: 590-593.

Vollmer, R. T., 1996. Demodex-Associated Folliculitis. Am J Dermatopathol, 18(6): 589-91.

Wesołowska, M., Baran, W., Szepietowski, J., Hirschberg, L., Jankowski, S., 2005. Demodicidosis in Humans As A Current Problem in Dermatology. Wiad Parazytol, 51(3): 253-256.

Yazar, S., Özcan, H., Çetinkaya, Ü., 2008. Üniversite Öğrencilerinde Selofan-Bant Yöntemi ile Demodex sp. Araştırılması. Türkiye Parazitol Derg. 32(3): 238-240.

Yee, T. W., 2008. VGAM Family Functions for Positive Zero-Altered and Zero-Inflated Discrete Distributions, http://www.stat.auckland.ac.nz/~yee/VGAM/doc/Positive.pdf

Yeşilova, A., et. al., 2007. Sıfır Değer Ağırlıklı Verilerin Modellenmesi, http://zootekni2007.yyu.edu.tr./pdfler/IST1.pdf

INVESTIGATION OF ZERO-INFLATED AND HURDLE MODELS IN DESCRIBING DEMODEX COUNTS BY VARIOUS VARIABLES

ABSTRACT

It is aimed to investigate the zero inflated regression models and hurdle regression models for describing demodex counts by various variables. It is also aimed to determine the model that will give the best results in describing demodex counts by different variables such as age, sex and profession with the information obtained from patients who come for porter examination. Data were obtained from 156 patients who came to Malatya Public Health Laboratory for porter examination between the dates of June 2007 - June 2009. Demodex count was determined as response variable and age, sex and profession of patients were determined as explanatory variables. R.2.11.1 software program was used for statistical analysis. Since 68 percent of demodex counts, which was considered as response variable, were zero, zero-inflated regression and hurdle models were used. As a result of applications, AIC values, which were used to determine the superiority of models relative to each other, were obtained as 731.18, 731.49, 531.73 and 531.11 for ZIP, ZAP, ZINB and ZANB models respectively. The model that has smallest AIC value was ZANB. In zero inflated and hurdle regression models, it is tested whether zero values have a significant impact. If the impact of zero values is not significant, the results of zero inflated and hurdle regression models and the results of Poisson regression and negative binomial regression models will be same. In this study, the impact of zero values in the existing data set was determined to be significant.

Keywords: Demodex, Hurdle Models, Zero-Inflated Regression Models.

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	Çankaya Üniversitesi
Mehmet KAYTAZ	Işık Üniversitesi
Meltem DAYIOĞLU TAYFUR	Orta Doğu Teknik Üniversitesi
Metin TOPRAK	Rekabet Kurumu
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
Reşat KASAP	Gazi Üniversitesi
Savaş ALPAY	SESRTCIC
Seyfettin GÜRSOY	Galatasaray Üniversitesi
Süleyman GÜNAY	Hacettepe Üniversitesi
Turan EROL	Ankara Strateji Enstitüsü
Ümit OKTAY FIRAT	Marmara Üniversitesi
Yasin AKTAY	Selçuk Üniversitesi
Yılmaz AKDİ	Ankara Üniversitesi