T.C.
MİMAR SİNAN GÜZEL SANATLAR ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

LOG-DOĞRUSAL MODELLERİN OLUMSALLIK ÇİZELGELERİNE UYGULANMASI

YÜKSEK LİSANS TEZİ

Nihan ACAR

İstatistik Anabilim Dalı

İstatistik Programı

Tez Danışmanı: Prof. Dr. Aydın ERAR

MAYIS 2011

ÖNSÖZ

Çalışmalarım sırasında bitmek tükenmek bilmeyen sorularıma sabırla cevap veren ve manevi desteğini hep yanımda hissettiğim Sevgili Hocam Prof. Dr. Aydın ERAR'a, çalışmalarım sırasındaki anlayış ve desteğinden dolayı Sayın Bölüm Başkanım Prof. Dr. Nalan CİNEMRE'ye, yeni çalışmalara beni teşvik eden ve veri bulmamda büyük katkısı olan Sayın Hocam Doç. Dr. Nural BEKİROĞLU'na ve hazırladığı veriyi cömertlikle benimle paylaşan Marmara Üniversitesi Radyoloji Bölümü Asistan Doktor Aysun OKAR'a teşekkürü bir borç bilirim. Beraber çalıştığım, bilgi ve tecrübelerinden yararlandığım tüm bölüm hocalarıma ve eğitimime katkısı olan tüm hocalarıma teşekkür ederim.

Yüksek Lisans'a başladığımız günden beri aynı yollarda birbirimize destek olarak omuz omuza ilerlediğimiz sevgili yol arkadaşım Arş. Gör. Bilge ÖZLÜER'e, yoğun çalışma tempomdan dolayı yüzümü görememeye sabırla katlanan anneme, aileme ve arkadaşlarıma, okuma sevgisini bana kazandıran teyzeme ve beni her zaman araştırmaya teşvik eden sevgili ağabeyime bana verdikleri manevi kuvvet ve destek için çok teşekkür ederim. Anneme ve başarılarımı her zaman gururla benimle paylaşan ancak hayatta olmadığı için bu günümü göremeyen babama eğitimime ve kendimi geliştirmeme verdikleri önem, beni bugünlere getiren maddi ve manevi destekleri için teşekkürü bir borç bilirim.

Mayıs 2011 **Nihan ACAR**

İÇİNDEKİLER

ÇİZELGE LİSTESİ

ÖZET

Genelleştirilmiş doğrusal modeller (GDM) yanıt değişkeni ile açıklayıcı değişkenler arasında basit doğrusal bir ilişki bulunmayan, Binom, Poisson ya da Gamma gibi üssel dağılım ailesi üyesi dağılımların analizinde kullanılır. GDM, yanıt değişkeninin bu dağılımlardan birine sahip olması durumunda kullanılan modelleme tekniklerinin bir çerçevesi olarak tanımlanabilir. Özellikle kesikli veriler için analiz yöntemleri içeren GDM'nin bir uzantısı yanıt ve açıklayıcı değişkenler arasında ayrım yapılmadığı zaman kullanılabilen log-doğrusal modellerdir. Log-doğrusal modeller çevrebilim, tıp, bankacılık gibi sektörlerde sıklıkla karşılaşılan kategorik değişkenler arasındaki ilişki örüntülerinin ortaya çıkarılmasını sağlar. Bu modeller genellikle olumsallık çizelgelerinin analizinde kullanılır.

Olumsallık çizelgelerinde beklenen hücre frekanslarının hesaplanması aykırı değerlerin tespiti için önemli bir adım teşkil etmektedir. Bilinen ve sıklıkla kullanılan En Çok Olabilirlik (EÇO) yöntemine alternatif olarak sağlam bir yöntem olan "Medyan Polish (MedPol)" yöntemi de beklenen hücre frekanslarının belirlenmesinde kullanılabilir. Bu çalışmada, sağlam olan "MedPol" ve sağlam olmayan "EÇO" yöntemlerinden elde edilen hücre frekanslarının karşılaştırılması ve EÇO ve MedPol kestirimleri üzerinden "α - aykırı değer bölgesi" yöntemi uygulanarak aykırı hücrelerin belirlenmesi amaçlanmıştır.

Bu çalışmanın amacı, olumsallık çizelgesi şeklinde verilen üç farklı veri kümesine log-doğrusal modellerin uygulanıp ilişki örüntülerinin ortaya çıkarılmasıdır. Bu amaçla çalışmanın ilk uygulamasında, sürücülerin son üç yılda yaptıkları kaza sayıları ile uyku alışkanlıkları arasında bir ilişki olup olmadığı ve sürücülerin verdiği yanıtların güvenilirliği incelenecektir. Bunun yanı sıra takip eden diğer iki uygulamada "Kanser lezyonlarına ait özellikler arasındaki ilişki örüntüleri nelerdir?" ile "Kredi başvurusu yapan kişilerde birikim miktarları, çalışma hayatındaki süreleri ve konut sahibi olma durumları arasında bir ilişki var mıdır?" sorularına yanıt aranacaktır.

SUMMARY

Generalized linear models (GLM) are used in the analysis of exponential distribution families such as Normal, Binomial, Poisson and Gamma distributions where a simple linear relationship is not found between response variable and explanatory variables. Generalized linear models can be defined as a framework of statistical modeling techniques which are used in the case that the response variable comes from one of those distributions. One of the extensions of GLM, that includes analyzing techniques especially for discrete data, is log-linear models which are applicable when there is no distinction between response and explanatory variables. Log-linear models help to reveal association patterns among categorical variables that are widely encountered in sectors such as ecology, medicine and banking. These models are generally used in the analysis of contingency tables.

The calculation of expected cell frequencies has importance in identifying outliers in the analysis of contingency tables. As an alternative to known and widely used Maximum Likelihood (ML) Method, Median Polish (MedPol) Method can also be used in determining expected cell frequencies. In this study, comparing the cell frequencies obtained from the robust method "MedPol" and the non-robust method "ML" and determining outlying cells by using "α - outlier region" method over ML and MedPol estimators were aimed.

The aim of this study is to reveal association patterns among variables in three different data sets given in the form of contingency tables by applying log-linear models. With this aim, in the first application of the study, the association between the number of accidents that drivers made in the last three years and the sleeping habits of drivers and the reliability of the answers of drivers were analyzed. Besides, in the following two applications the answers to the questions: "What are the association patterns among the properties of cancer lesions?" and "Is there any relationship between savings account balance, years working and home ownership of people who made credit applications?" will be sought.

1. GİRİŞ

Bilim ve teknolojide sıkça kullanılan modeller, gerçeğin özet ve basit temsilleridir (Lindsey, 1997). Bir verinin yapısını, rastlantı değişkenleri ve bu değişkenlere bağlı olasılıklar açısından matematiksel denklemler kullanarak açıklayan modellere istatistiksel modeller denir. Bu modellerin en önemli sınıfını ise,

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik} + \varepsilon_i \qquad (1.1)$$

ifadesi ile gösterilen doğrusal modeller oluşturur. Burada, y yanıt (bağımlı) değişkenini, $X_1, X_2, ..., X_k$ açıklayıcı (bağımsız) değişkenleri; $\beta_0, \beta_1, ..., \beta_k$ bilinmeyen parametreleri ve ε hata terimini ifade eder. (1.1) ile verilen bu denklem Y nicel ve sürekli bir değişken iken, alışılagelen ifadesiyle doğrusal regresyon modeli olarak adlandırılır.

Genelleştirilmiş doğrusal modeller (GDM) sınıfı, doğrusal regresyon ve varyans analizi modellerini, ikili yanıtlar için lojit ve probit modellerini, frekanslar için log-doğrusal ve çoklu yanıt modellerini içerir (McCullagh ve diğ., 1988). Genelleştirilmiş doğrusal modeller, normal dağılımı temel alarak (1.1) denklemi ile verilen klasik doğrusal modelleri genelleştirirler.

Klasik doğrusal modeller ilk olarak Gauss ve Legendre tarafından 1809 yılında astronomide uygulanmıştır. Fisher'in 1919'da başladığı deney tasarımı konularındaki çalışmalarının genelleştirilmiş doğrusal modellerin geliştirilmesine katkısı büyüktür. Yine Fisher 1922 yılında Poisson dağılımına sahip organizmaları incelediği seyrelti analizi olarak bilinen çalışmasında, genelleştirilmiş doğrusal modellerde temel kestirim yöntemi olarak kullanılan en çok olabilirlik yöntemini tanımlamıştır (Mc Cullagh ve Nelder, 1983).

Fisher zamanından beri sıkça kullanılan birçok dağılımın, Fisher'in üssel dağılım ailesi (exponential distribution family) olarak adlandırdığı, tek bir aile üyesi olduğu bilinmekteydi. Nelder ve Wedderburn 1972 yılında, tüm bu dağılım ailesi üyelerine aynı şekilde davranılabileceğini ve tüm bu modeller için en çok olabilirlik

kestiricilerinin yinelemeli ağırlıklı en küçük kareler yöntemi adı verilen algoritma ile ele edilebileceğini göstermişlerdir (Lindsey, 1997).

GDM'de yanıt değişkeninin dağılımı ve bağlantı fonksiyonunun türüne göre değişik modeller vardır. Yanıt değişkeninin kesikli olduğu ve Binom, Poisson ya da Poisson dağılımının aşırı yayılım sonucu Negatif Binom dağılımına dönüştüğü durumlarda sırasıyla Lojistik regresyon, Poisson regresyon, Negatif Binom regresyon modelleri kullanılır. Birden fazla kategorik değişkenin çapraz sınıflandırılmasından oluşan olumsallık çizelgelerinin analizinde ise lojit ve log-doğrusal modeller tercih edilir. Değişkenlerden birinin yanıt değişkeni olarak ele alındığı durumda lojit modeller; açıklayıcı değişkenler ile yanıt değişkeni arasında bir ayrım yapılmadığı, birden fazla yanıt değişkeninin olduğu ya da tüm değişkenlerin yanıt değişkeni olarak kabul edildiği durumlarda ise log-doğrusal modeller kullanılır (Baker, 1981; Nelder, 2000). Açıklayıcı değişkenlerin kategorik olduğu durumda lojit modeller log-doğrusal modele dönüşürler (Rao ve Toutenburg, 1999). Poisson regresyonu ve log-doğrusal modeller özellikle frekansların modellenmesinde kullanılır. Açıklayıcı değişkenler hem nitel hem nicel ise Poisson regresyon modeli ya da aşırı yayılım varsa Negatif Binom regresyon modeli tercih edilir. Tüm açıklayıcı değişkenler nitel ise log-doğrusal modeller kullanılır. Bu tip veri kümelerini olumsallık çizelgesi ile ifade etmek de mümkündür.

Olumsallık (contingency) çizelgelerinde log-doğrusal modellerin kullanımı oldukça eskiye dayanır. Log-doğrusal modellerin kullanımı özellikle biyoloji, çevrebilim, sigortacılık, sağlık bilimleri gibi kategorik verilere sıkça rastlanılan alanlarda yaygındır.

Olumsallık çizelgesi deyimi ilk kez 1900'lü yıllarda Karl Pearson tarafından kullanılmıştır. Aynı yıllarda Pearson ve Yule, olumsallık çizelgelerinde kesikli değişkenler arasındaki ilişki yapısını ortaya çıkarmak için ilişki katsayıları bulmuşlardır.

1922 yılında Fisher'in verdiği serbestlik derecesi düzeltmesi ile ki-kare bağımsızlık testi uyum iyiliği ve model seçimi kavramlarını oluşturdu. Buna alternatif olarak da 1930'lu yıllarda Wilks, olabilirlik oran istatistiğini (G^2) önermiştir.

Bartlett (1935) olumsallık çizelgelerinde hücre değerlerinin kestirimi için en çok olabilirlik kestiricilerini geliştirmiştir. Daha sonra bu yöntem Norton tarafından

(1945) 2 x 2 x t boyutlu çizelgelere uygulanmıştır. Hücre değerlerinin kestirimi için önerilen bir başka yöntem ise Deming ve Stephan (1940) tarafından bulunan yinelemeli oransal kestirim (iterative proportional fitting - IPF) yöntemidir. Bu yöntem daha sonra Goodman tarafından (1965) boş hücrelerin bulunduğu veri kümelerine uyarlanmıştır (Fienberg ve Rinaldo, 2007). Daha sonra bu yöntem Bishop tarafından geliştirilmiştir (Bishop ve Fienberg, 1969).

Genelleştirilmiş doğrusal modellerin bir uzantısı olan log-doğrusal modellerin olumsallık çizelgelerine uygulanması konusu 1960'lı yıllarda popüler olmuştur. Teknolojinin gelişmesi ve yüksek hızlı bilgisayarların kullanılmaya başlanmasıyla Good (1963), Goodman (1965, 1970) ve Birch (1963) kategorik veri analizi, özel olarak da log-doğrusal modeller ve En Çok Olabilirlik Kestiricileri üzerine çalışmalar yapmıştır. Birch (1963) n-yönlü olumsallık çizelgelerinde en çok olabilirlik kestiricilerini elde etmiş ve hücre frekansları için Poisson ve bağımsız çok terimli dağılımların denkliğini göstermiştir. Bishop (1969) ise Birch'in çalışmalarından faydalanarak lojit ve log-doğrusal modeller arasındaki ilişkiyi incelemiştir.

Fienberg (1968, 1970a) 2 x 2 ve r x c boyutlu olumsallık çizelgelerinin geometrik yorumunu yapmış ve hücre frekanslarının sıfır ya da boş olduğu durumda EÇO kestiricilerinin elde edilmesi konulu makaleler yayınlamıştır.

Goodman (1970), log-doğrusal modeller ile Wilks'in önerdiği olabilirlik oran istatistiğini kullanarak n-yönlü tabloların analizi için yöntemler sunmuştur. Haberman (1973) olumsallık çizelgelerinde EÇO kestiricilerinin varlığı için gerek ve yeter koşulları göstermiştir (Fienberg ve Rinaldo, 2007).

1970'lerde tüm bu konular McCullagh ve Nelder tarafından genelleştirilmiş doğrusal modeller başlığı altında incelenmiştir.

Bu çalışmada, Genelleştirilmiş Doğrusal Modellerin kategorik veri analizinde kullanılan bir alt sınıfı olan log-doğrusal modellerin tanıtılması ve bu modellerin değişkenler arası ilişki örüntülerinin ortaya çıkarılması için iki yönlü ve üç yönlü olumsallık çizelgelerine uygulanması amaçlanmıştır.

Tezin birinci kısmında Genelleştirilmiş Doğrusal Modeller tanıtılmış ve yanıt değişkeninin dağılımına göre hangi durumda hangi modelin kullanılmasının daha

uygun olacağı belirtilmiştir. Bu bölümde ayrıca GDM'de parametre kestirim yöntemleri ve model uyumunun belirlenmesi konularına değinilmiştir.

Tezin ikinci kısmında, olumsallık çizelgeleri ve log-doğrusal modeller hakkında genel bilgiler verilmiş ve log-doğrusal modellerin olumsallık çizelgelerinin analizinde kullanımı açıklanmıştır. Olumsallık çizelgelerine ait beklenen hücre frekanslarının söz konusu modele göre EÇO ve Medyan Polish yöntemleri aracılığıyla hesaplanması anlatılmıştır.

Tezin üçüncü ve son kısmında ise, önceki bölümlerde anlatılan bilgiler ışığında trafik, sağlık ve bankacılık alanlarına ait üç ayrı kategorik veri kümesine log-doğrusal modeller uygulanmıştır. Birinci veri kümesinde sürücülerin uyku alışkanlıkları ve yaptıkları kaza sayıları arasındaki ilişkinin araştırılması, EÇO ve Medyan Polish kestirimlerinin karşılaştırılması amaçlanmıştır. İkinci veri kümesinde, kanser lezyonlarına ait önce şekil - vaskülarite, sonra patoloji – yerleşim - vaskülarite değişkenleri arasındaki ilişki örüntüleri araştırılmıştır. Kredi veri kümesinde ise, kişilerin yaptıkları birikim, çalışma hayatında geçirdikleri yıl sayısı ve konut sahibi olup olmadıkları bilgilerinden üç yönlü bir çizelge oluşturularak değişkenler arası ilişkiler incelenmiştir. Bununla birlikte, EÇO ve Medyan Polish yöntemlerinin üç yönlü çizelgelere uyarlanışı ele alınmıştır.

2. GENELLEŞTİRİLMİŞ DOĞRUSAL MODELLER (GDM)

Doğrusal modeller genel olarak, ε hata teriminin bağımsız ve Normal dağılıma $N(0,\sigma^2)$ sahip olduğu varsayımları altında,

$$Y = X\beta + \varepsilon \tag{2.1}$$

basit biçimiyle gösterilebilir.

Y yanıt değişkeninin kategorik olduğu, Normal dağılımdan başka bir dağılıma sahip olduğu ya da yanıt ve açıklayıcı değişkenler arasındaki ilişkinin (2.1) denklemindeki gibi basit doğrusal olmadığı durumlarda genelleştirilmiş doğrusal modeller kullanılır.

GDM'nin ana fikirlerinden biri veri dönüşümünden kaçınmaktır. GDM'nin temel stratejisi, yanıtın ortalamasına bir bağlantı fonksiyonu uygulanarak EÇO yöntemi aracılığıyla bir model kestirilmesidir (McCulloch ve Searle, 2001).

Nelder ve Wedderburn (1972), Normal, Binom, Poisson, Gamma dağılımları gibi sıkça kullanılan birçok dağılımın aslında üssel dağılım ailesi olarak isimlendirilen bir ailenin üyeleri olduğunu ve tek bir biçimde ifade edilebileceğini göstermiştir.

2.1. ÜSSEL DAĞILIM AİLESİ

Üssel dağılım ailesi üyesi dağılımların bir y yanıt değişkeni için sahip olduğu olasılık yoğunluk fonksiyonu $a(\cdot)$, $b(\cdot)$ ve $c(\cdot)$ belirli fonksiyonlar olmak üzere,

$$f(y;\theta,\phi) = \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right\} \tag{2.2}$$

biçimiyle gösterilir. Burada, θ doğal konum parametresi ve ϕ yayılım ya da ölçek parametresi olarak adlandırılır.

Normal, Binom, Poisson, Gamma, Ters normal, Negatif Binom, Geometrik ve Üstel dağılımlar bu ailenin üyeleridir. Bu dağılımlardan en çok kullanılan Normal, Binom ve Poisson dağılımlarının (2.2) eşitlikleri aşağıdaki gibi yazılabilirler:

μ ortalamalı, σ^2 varyanslı Normal dağılıma sahip y yanıt değişkeni için olasılık yoğunluk fonksiyonu,

$$f(y;\mu,\sigma) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left[-\frac{1}{2\sigma^2}(y-\mu)^2 \right]$$

$$= \exp\left[-\frac{y^2}{2\sigma^2} + \frac{y\mu}{\sigma^2} - \frac{\mu^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2) \right] \tag{2.3}$$

olur. Burada, doğal konum parametresi θ=μ dür ve yayılım parametresi ϕ=σ^2 dir. $a(\cdot)$, $b(\cdot)$ ve $c(\cdot)$ belirli fonksiyonları ise sırasıyla $a(\theta)=\sigma^2$, $b(\theta)=\mu^2/2$ ve $c(y,\theta) = -\frac{1}{2}\left[y^2/\sigma^2 + \log(2\pi\sigma^2) \right]$ şeklindedir.

Binom dağılımının olasılık fonksiyonu,

$$f(y;\pi) = \binom{n}{y} \pi^y (1-\pi)^{n-y}$$

$$= \exp\left[y\log\pi - y\log(1-\pi) + n\log(1-\pi) + \log\binom{n}{y} \right] \tag{2.4}$$

biçiminde yazıldığında, $\theta = \log\left(\frac{\pi}{1-\pi}\right), b(\theta) = n\log(1+\exp(\theta))$, $a(\phi)=1$ ve $c(y) = \log\binom{n}{y}$ olacaktır.

Poisson dağılımına sahip kesikli Y rastlantı değişkeni için de olasılık fonksiyonu,

$$f(y;\lambda) = \exp\left[y\log\lambda - \lambda - \log y! \right] \tag{2.5}$$

biçiminde yazılabilir.

Sıkça kullanılan bazı üssel dağılım ailesi üyesi dağılımların özellikleri Çizelge 2.1'de gösterildiği gibidir (Dobson, 1990; Myers ve diğ., 2002).

Çizelge 2.1. Üssel dağılım ailesi üyesi dağılımların özellikleri

	Doğal Konum parametresi (θ)	$a(\cdot)$	$b(\cdot)$	$c(\cdot)$
Normal	μ	σ^2	$\mu^2/2$	$-\frac{1}{2}\left[y^2/\sigma^2+\log(2\pi\sigma^2)\right]$
Binom	$\log\left(\frac{\pi}{1-\pi}\right)$	1	n log $(1+e^{\theta})$	$\log\binom{n}{y}$
Poisson	log λ	1	e^{θ}	-log y!
Üstel	$1/\lambda$	-1	$\ln\lambda$	0
Gamma	$-1/\lambda r$	τ^{-1}	$-\ln(1/\mu)$	$r\ln\lambda-\ln\tau(r)+(r-1)\ln y$

2.2. GENELLEŞTİRİLMİŞ DOĞRUSAL MODELLERİN YAPISI

Genelleştirilmiş doğrusal modeller, regresyon modellerinin, Y yanıt değişkeninin normal olmadığı durumlarda ortalamanın fonksiyonunu modellemek için kullanılan, bir uzantısıdır (Agresti, 2002).

Genelleştirilmiş doğrusal modeller rasgele bileşen, sistematik bileşen ve bağlantı fonksiyonu olarak adlandırılan üç bileşenden oluşur:

Rasgele bileşen, (2.2) denklemi ile verilen bir üssel dağılım ailesi üyesi olan aynı dağılımdan gelen $y_1, y_2, ..., y_n$ yanıt değişkenlerini ve bunlara ait olasılık yoğunluk fonksiyonlarını belirtir.

Sistematik bileşen, $X_1, ..., X_n$ açıklayıcı değişkenlerinin oluşturduğu $\eta_i = \sum_1^p \beta_j x_{ij}$ doğrusal kestirim fonksiyonunu gösterir.

Bağlantı fonksiyonu ise rasgele ve sistematik bileşenleri birbirine bağlayan monoton ve türevlenebilir bir fonksiyondur ve $\mu_i = E(y_i)$ olmak üzere,

$$g(\mu_i) = \eta_i \qquad \text{i=1,2,…,n} \tag{2.6}$$

biçiminde yazılabilir.

$g(\mu) = \mu$ olarak verilen bağlantı fonksiyonuna *birim bağlantı fonksiyonu* denir. Birim bağlantı fonksiyonu, normal dağılımlı Y değişkenine ait regresyon modellemesinde kullanılan bağlantı fonksiyonudur.

Özel olarak $\eta_i = \theta_i$ seçilirse, bu durumda η_i'ye kanonik bağlantı adı verilir. Kanonik bağlantı fonksiyonları Y yanıt değişkeninin dağılımına göre farklılık gösterir. Sık kullanılan bazı dağılımlara ilişkin bağlantı fonksiyonları Çizelge 2.2'de gösterilmiştir. Çizelge 2.2.'de belirtilen bağlantı fonksiyonları dışında,

1) Probit bağlantı: $\eta_i = \Phi^{-1}\left[E(y_i)\right]$

2) Tamamlayıcı log-log bağlantı: $\eta_i = \ln\{\ln[1-\mu_i]\}$

3) Üssel dönüşüm ailesi bağlantı: $\eta_i = \begin{cases} \mu_i^{\lambda}, & \lambda \neq 0 \\ \ln[\mu_i], & \lambda = 0 \end{cases}$

fonksiyonları da genelleştirilmiş doğrusal modellerde kullanılan farklı bağlantı fonksiyonlarıdır.

Dağılıma uygun bağlantı fonksiyonunun seçimi, geçerli ve doğru bir model elde etmek için önemli bir adım teşkil etmektedir.

Çizelge 2.2. Bazı dağılımlar için kanonik bağlantı fonksiyonları

Dağılım	**Bağlantı Adı**	**Kanonik Bağlantı Fonksiyonu**
Normal	Birim Bağlantı	$\eta = \mu$
Binom	Logit Bağlantı	$\eta = \log(\mu/1-\mu)$
Poisson	Log Bağlantı	$\eta = \log \mu$
Üstel	Ters Bağlantı	$\eta = 1/\mu$
Gamma	Ters Bağlantı	$\eta = 1/\mu$

Genelleştirilmiş doğrusal modellerin önemli bir özelliği, klasik regresyon modellerinde olduğu gibi gözlemlerin bağımsızlığı varsayımıdır. Bununla birlikte genelleştirilmiş doğrusal modellerde tek bir hata terimi vardır. Ancak bu modellerde üssel dağılım ailesi üyesi dağılımlarla ilgilenildiğinden klasik regresyonda olduğu

gibi hata teriminin normallik ve sabit varyanslılık özelliklerinin sağlaması gerekmez (McCullagh ve Nelder, 1988).

2.3. GENELLEŞTİRİLMİŞ DOĞRUSAL MODELLER İÇİN PARAMETRE KESTİRİM YÖNTEMLERİ

Parametrelerin istatistiksel kestirimleri için en yaygın olarak kullanılan iki yöntem: En Çok Olabilirlik (EÇO) ve En Küçük Kareler (EKK) yöntemleridir. Genelleştirilmiş doğrusal modellerde parametre kestirimleri, iteratif olarak yeniden ağırlıklandırılmış en küçük kareler metoduna dayalı EÇO yöntemiyle elde edilir (Dobson,1990; Myers ve diğ., 2002).

Üssel dağılım ailesi üyesi bir dağılıma ait $y_1, y_2, ..., y_n$ bağımsız yanıt değişkenleri için en çok olabilirlik yöntemi,

$$L(\theta,\phi;y)=\prod_{i=1}^{n} f(y_i;\theta_i,\phi) \tag{2.7}$$

olarak tanımlanan olabilirlik fonksiyonun en büyüklenerek $\hat{\theta}$ parametre kestirimlerinin bulunmasına dayanır. Bu da olabilirlik fonksiyonunun θ parametre vektörüne göre birinci dereceden türevinin sıfıra eşitlenmesi ile mümkündür. Logaritma fonksiyonu monoton bir fonksiyon olduğundan, genelde $L(\theta,\phi;y)$ olabilirlik fonksiyonunun yerine olabilirlik fonksiyonun logaritması olan,

$$l(\theta,\phi;y)=\log L(\theta,\phi;y)=\sum_{i=1}^{N} l_i=\sum_{i=1}^{N}\left\{\frac{y_i\theta_i-b(\theta_i)}{a(\phi)}+c(y_i;\phi)\right\} \tag{2.8}$$

fonksiyonu kullanılır. $b'(\theta_i)=\partial b(\theta_i)/\partial\theta_i$ ve $b''(\theta_i)=\partial^2 b(\theta_i)/\partial\theta_i^{\,2}$, $l(\theta,\phi;y)$ log-olabilirlik fonksiyonunun θ_i'ye göre birinci ve ikinci dereceden türevleri olmak üzere,

$$\frac{\partial l_i}{\partial\theta_i}=\frac{[y_i-b'(\theta_i)]}{a(\phi)}, \tag{2.9}$$

$$\frac{\partial^2 l_i}{\partial\theta_i^{\,2}}=\frac{-b''(\theta_i)}{a(\phi)}\ , \tag{2.10}$$

şeklindedir. Burada (2.9) eşitliği *skor fonksiyonu* olarak adlandırılır ve U ile gösterilir. (2.9) denkleminin sıfıra eşitlenmesi ile olabilirlik denklem çözümünün

$a(\phi)$ parametresinden bağımsız olduğu açıklık kazanır. Ek 1'de görüldüğü üzere düzenlilik varsayımları altında integral ve türevin sırası değiştirilebilir olduğundan,

$$E\left(\frac{\partial l_i}{\partial \theta_i}\right) = 0 \tag{2.11}$$

$$-E\left(\frac{\partial^2 l_i}{\partial {\theta_i}^2}\right) = E\left(\frac{\partial l_i}{\partial \theta_i}\right)^2 \tag{2.12}$$

dir.

(2.9) ve (2.11) denklemlerinden,

$$E(y_i) = \mu_i = b'(\theta_i) \tag{2.13}$$

olarak bulunur.

Benzer şekilde (2.10) ve (2.12) denklemlerinden,

$$\frac{b''(\theta_i)}{a(\phi)} = E\left[\frac{(y_i - b'(\theta_i))^2}{a^2(\phi)}\right] = \frac{\text{var}(y_i)}{a^2(\phi)} \tag{2.14}$$

elde edilir ve $E[y_i - b'(\theta_i)] = 0$ olduğundan,

$$V(\mu_i) = \text{var}(y_i) = b''(\theta_i)a(\phi) \tag{2.15}$$

yazılır.

Y yanıt değişkeninin i'nci gözlemine ait doğrusal kestirim fonksiyonu $\eta_i = x_i^T \beta = \sum_{j=1}^{p} x_{ij}\beta_j$ olsun (i=1,2...n). η tahmin edicilerinin, monoton ve türevlenebilir olan,

$$g(\mu_i) = \eta_i \qquad \text{(i=1,2,...,n)} \tag{2.16}$$

fonksiyonu aracılığıyla E(y)=μ ortalamasına bağlandığını varsayalım. Böylece, θ_i ve β parametreleri (2.13) denklemi dolayısıyla bağlantılıdır ve $\theta_i = \theta_i(\beta)$ olur.

Bu durumda, log-olabilirlik fonksiyonun β parametre vektörüne göre birinci dereceden türevi, zincir kuralına bağlı olarak,

$$\frac{\partial l_i(\beta)}{\partial \beta j} = \frac{\partial l_i}{\partial \theta_i}\frac{\partial \theta_i}{\partial \mu_i}\frac{\partial \mu_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \beta_j} \tag{2.17}$$

biçiminde yazılır. (2.17) denkleminin kısmi türevlerinin Ek2'de hesaplandığı gibi yerine yazılıp, sıfıra eşitlenmesi ile,

$$\frac{\partial l_i}{\partial \beta_j} = \sum_{i=1}^{N} \frac{(y_i - \mu_i)x_{ij}}{(\text{var}\, y_i)} \frac{\partial \mu_i}{\partial \eta_i} = 0, \qquad \text{j=1...,p} \tag{2.18}$$

elde edilir. EÇO kestiricileri bu denklemin çözümünden bulunur. Burada log-olabilirlik fonksiyonu β parametresine göre doğrusal olmadığından, (2.18) denkleminin çözümü iteratif yöntemler gerektirir (Rao ve Toutenburg, 1999).

β parametre vektörünün elemanlarına göre ikinci dereceden türev, (2.12) ve (2.18) denklemleri kullanılarak,

$$E\left(\frac{\partial^2 l_i}{\partial \beta_j \partial \beta_h}\right) = -E\left(\frac{\partial l_i}{\partial \beta_j}\right)\left(\frac{\partial l_i}{\partial \beta_h}\right) = -E\left[\frac{(y_i - \mu_i)(y_i - \mu_i)x_{ij}x_{ih}}{(\text{var}(y_i))^2}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2\right]$$

$$= -\frac{x_{ij}x_{ih}}{\text{var}(y_i)}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2 = E\left(-\frac{\partial^2 l(\beta)}{\partial \beta_j \partial \beta_h}\right) \tag{2.19}$$

elde edilir.

Bu eşitlik, tüm (j,h)-kombinasyonları için,

$$\omega_i = \frac{\left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2}{\text{var}(y_i)} \tag{2.20}$$

ağırlıkları $W = diag(\omega_1 ..., \omega_n)$ matrisinin köşegen elemanlarını oluşturmak üzere matris formunda,

$$I_{jk} = E\left(-\frac{\partial^2 l(\beta)}{\partial \beta_j \partial \beta_k}\right) = X'WX \tag{2.21}$$

şeklinde ifade edilebilir.

Bilgi matrisi olarak da bilinen bu matris, Ek1 (E.6)'da belirtildiği gibi j=1,...,p olmak üzere U_j'lerin varyans kovaryans matrisidir ve Fisher Skorlama Algoritması'nda parametreleri belirlemek amaçlı kullanıldığından, Fisher bilgi matrisi olarak da adlandırılır (Rao ve Toutenburg, 1999).

Genelleştirilmiş doğrusal modellerde parametre kestririmlerinin belirlenmesinde iterasyona dayalı iki yöntem kullanılır: Newton Raphson Yöntemi ve Fisher Skorlama Algoritması.

2.3.1. Newton Raphson Yöntemi

Genel olarak, EÇO kestirimi skor denklemlerinin bir çözümüdür. Skor fonksiyonları doğrusal olmadığından sayısal iterasyonlar aracılığıyla çözülmelidir. Bu fonksiyonların çözümünde en sık kullanılan yöntem, olabilirlik fonksiyonunu en büyükleyecek başlangıç parametre değerinin ön tahminine dayanan Newton Raphson algoritmasıdır.

Newton Raphson algoritması, $U^{(m-1)}$, $\beta=b^{(m-1)}$ noktasındaki $U_j = \partial l / \partial \beta_j$ birinci dereceden türevlerinin vektörü; $\left[\frac{\partial^2 l}{\partial \beta_j \partial \beta_k}\right]_{\beta=b^{(m-1)}}$, $\beta=b^{(m-1)}$ noktasında l fonksiyonunun ikinci dereceden türevlerinin oluşturduğu Hessian matrisi olmak üzere Taylor serisi yaklaşımı kullanılarak,

$$b^{(m)} = b^{(m-1)} - \left(\left[\frac{\partial^2 l}{\partial \beta_j \partial \beta_k}\right]_{\beta=b^{(m-1)}}\right)^{-1} U^{(m-1)} \qquad (2.22)$$

şeklinde ifade edilir. Buradan EÇO kestiricisi $m \to \infty$ iken $b^{(m)}$'in limiti olarak elde edilir (Rao ve Toutenburg, 1999).

Bu iteratif algoritma, maksimum noktasında bir değişiklik kaydedilmeyinceye kadar devam eder.

2.3.2. Fisher Skorlama Algoritması

EÇO kestiriminin belirlenmesinde Newton Raphson algoritmasına alternatif bir yöntem de skorlama algoritmasıdır. Bu yöntem, ilk kez R.A. Fisher tarafından probit modellerin EÇO kestirimlerinin bulunmasında kullanıldığından Fisher skorlama algoritması adını alır (Agresti, 2007).

Regresyondaki ağırlıklandırılmış en küçük kareler yöntemine benzeyen bu algoritmada, Newton Raphson yönteminden farklı olarak Hessian matrisi yerine

Fisher bilgi matrisi olarak bilinen ikinci türevlerin oluşturduğu Hessian matrisinin beklenen değeri kullanılır ve her bir adımda ağırlıklar yeniden belirlenir.

$\beta^{(k)}$, EÇO kestiricisi $\hat{\beta}$'nın k'ncı dereceden bir yaklaşımı, $q^{(k)}(\beta) = \partial l(\beta)/\partial\beta$ (2.18)'de ifade edilen $\beta^{(k)}$ için birinci dereceden türevlerin vektörü ve $W^{(k)}$ (2.20) ifadesi ile belirtilen köşegen elemanlarından oluşan matris olmak üzere Fisher skorlama algoritması,

$$(X'W^{(k)}X)\beta^{(k+1)} = (X'W^{(k)}X)\beta^{(k)} + \mathbf{q}^{(k)} \tag{2.23}$$

denklemiyle ifade edilir. Bu ifadenin sağ tarafındaki vektör daha önce belirtilen (2.18) ve (2.19) bileşenleri türünden yazılabilir:

$$\sum_h \left[\sum_i \frac{x_{ij}x_{ih}}{\text{var}(y_i)} \left(\frac{\partial\mu_i}{\partial\eta_i} \right)^2 \beta_h{}^{(k)} \right] + \sum_i \frac{\left(y_i - \mu_i{}^{(k)}\right)x_{ij}}{(\text{var}\, y_i)} \frac{\partial\mu_i}{\partial\eta_i} \qquad \text{(j=1,...,p)} \tag{2.24}$$

Bu ifade,

$$z_i{}^{(k)} = \sum_{j=1}^{p} x_{ij}\beta_j{}^{(k)} + (y_i - \mu_i{}^{(k)}) \left(\frac{\partial\eta_i{}^{(k)}}{\partial\mu_i{}^{(k)}} \right) = \eta_i{}^{(k)} + (y_i - \mu_i{}^{(k)}) \left(\frac{\partial\eta_i{}^{(k)}}{\partial\mu_i{}^{(k)}} \right) \tag{2.25}$$

iken matris biçiminde,

$$X'W^{(k)}\mathbf{z}^{(k)} \tag{2.26}$$

olarak gösterilebilir. Bundan dolayı, (2.23) ile verilen algoritma

$$(X'W^{(k)}X)\beta^{(k+1)} = X'W^{(k)}\mathbf{z}^{(k)} \tag{2.27}$$

olarak da yazılabilir. EÇO kestiricisi $\hat{\beta}$, $\text{k} \to \infty$ iken,

$$(X'W^{(k)}X)\beta^{(k+1)} = X'W^{(k)}\mathbf{z}^{(k)} \tag{2.28}$$

denkleminin limitidir.

$g(\mu_i) = \theta$ kanonik bağlantı foksiyonu için Newton Raphson algoritması ve Fisher skorlama algoritması aynı sonucu verir (Rao ve Toutenburg,1999).

2.4. İSTATİSTİKSEL ÇIKARSAMALAR

İstatistiksel modelleme, model belirleme, parametre kestirimi ve istatistiksel çıkarsamalar olmak üzere üç adım içerir (Dobson,1990).

Genelleştirilmiş doğrusal modeller için model belirlenmesi ve parametre kestirimi aşamalarından daha önceki kısımlarda bahsedilmişti. Bu bölümde istatistiksel çıkarsamalar başlığı altında yer alan güven aralıkları, hipotez testleri, model geçerliliği ve uyum iyiliği konularına değinilecektir.

2.4.1. Modelin Uyum İyiliği

Model seçimi sürecinde parametre sayısına göre farklı modeller söz konusudur. İstatistiksel analizlerde en sık kullanılan iki model doygun model ve sıfır modelidir. Bu modelleri aşağıdaki gibi özetleyebiliriz:

Doygun model: Gözlem sayısı kadar parametreye sahip, veriyi en iyi açıklayan ancak basitleştirmeden uzak olduğu için yorumu en zor olan modeldir. Veri kümesine ait maksimum sayıda parametreye sahip olan model olduğundan maksimal model olarak da adlandırılır.

Sıfır modeli (Null model): Tek parametreli en basit modeldir. Tüm y değerleri için tek bir ortalama yanıt değeri atar. Bu model y'lerdeki tüm değişimi sistematik bileşene atar.

n gözlemli bir veri kümesine en fazla n parametreye sahip bir model uydurulabilir. Bir veriye model uydurmak, y değerlerinin olabildiğince az sayıda parametre içeren bir modelden türetilen ortalama yanıt değerleri $\hat{\mu}$'lar ile yer değiştirmesi olarak tanımlanabilir. Genel olarak, $\hat{\mu}$'lar tam olarak y değerlerine eşit olmayacağından, bu iki değerin birbirinden ne kadar farklı olduğu sorusu ortaya çıkar. Bu iki değer arasındaki düşük orandaki farklılık göz ardı edilebilirken, yüksek orandaki farklılıklar için aynı şey söylenemez (McCullagh ve Nelder, 1988).

Bir modelin uyumunu belirlemek için ilgilenilen modelin olabilirlik fonksiyon değeri ile aynı dağılım ve bağlantı fonksiyonuna sahip, gözlem sayısı kadar parametre içeren doygun (tam) modelin olabilirlik fonksiyonu değeri karşılaştırılır.

Doygun model ve ilgilenilen modelin en çok olabilirlik kestiricileri $b_{\max}$ ve b değerlerindeki olabilirlik fonksiyonları, sırasıyla, $L(b_{\max}; y)$ ve $L(b; y)$ olsun. Eğer ilgilenilen model veriyi iyi tanımlıyorsa, $L(b; y)$ değeri yaklaşık olarak $L(b_{\max}; y)$ değerine eşit olmalıdır. Eğer model zayıfsa $L(b; y)$ değeri $L(b_{\max}; y)$ değerinden daha küçük olacaktır (Dobson,1990).

Bir modele ait uyum iyiliğini ölçmek için,

$$\lambda = \frac{L(b_{\max}; y)}{L(b; y)} \tag{2.29}$$

olarak tanımlanan *olabilirlik oran istatistiği* ya da buna denk olarak,

$$\log \lambda = l(b_{\max}; y) - l(b; y) \tag{2.30}$$

şeklinde ifade edilen log-olabilirlik fonksiyonları arasındaki fark kullanılabilir; $\log \lambda$ 'nın değeri ne kadar büyükse uyum o denli zayıftır.

Genelleştirilmiş doğrusal modeller için başka bir uyum iyiliği ölçütü de Nelder ve Wedderburn (1972) tarafından *sapma (deviance)* olarak da adlandırılan *log-olabilirlik oran istatistiği*dir. Bu istatistik,

$$D = 2\log \lambda = 2\left[l(b_{\max}; y) - l(b; y)\right] \tag{2.31}$$

olarak tanımlanır ve

$$D = 2\left\{\left[l(b_{\max}; y) - l(\beta_{\max}; y)\right] - \left[l(b; y) - l(\beta; y)\right] + \left[l(\beta_{\max}; y) - l(\beta; y)\right]\right\} \tag{2.32}$$

şeklinde yeniden yazılabilir.

(2.32) ifadesinin sağ yanındaki ilk terim doygun model gözlem sayısı n kadar parametreye sahip olduğundan n serbestlik dereceli χ_n^2 dağılımına sahiptir. Benzer şekilde, ilgilenilen model p parametreye sahip olduğundan ikinci terim de χ_p^2 dağılımına sahiptir. Üçüncü terim ise, eğer ilgilenilen p parametreli model, veriyi maksimal model kadar iyi açıklıyorsa sıfıra eşit olacak pozitif bir sabittir. Bu durumda, söz konusu model yeteri kadar iyi bir model ise üçüncü terim yaklaşık olarak sıfır olacağından modele ait log-olabilirlik oran istatistiği D, n gözlem sayısı ve p parametre sayısı olmak üzere n-p serbestlik dereceli ki-kare dağılımına yakınsar; ya da,

$$D \sim \chi^2_{n-p} \tag{2.33}$$

yazılır. Bunun tersine, eğer model uyumu zayıfsa (2.32) denkleminin sağ yan eşitliğindeki üçüncü terim değeri artacağından D değeri de χ^2_{n-p} ile tahmin edilenden daha büyük olacaktır.

Log-olabilirlik oran istatistiği, dağılıma ait D değeri uygun ki-kare dağılımı ile karşılaştırılarak model geçerliliğinin araştırılmasında da kullanılabilir.

İyi bir model için D değerinin dağılım ortalamasına yakın olması beklenir. Ki-kare dağılımına sahip bir rastlantı değişkenin beklenen değeri dağılımın serbestlik derecesine eşit olduğundan, n gözlemli bir veri kümesi için iyi açıklama sağlayan p parametreli model,

$$D \cong n - p \tag{2.34}$$

eşitliğini sağlayan modeldir (Dobson, 1990).

Bazı özel dağılımlara ait sapma ölçütleri Çizelge 2.3.'de verildiği gibidir.

Çizelge 2.3. Bazı özel dağılımlara ait sapma ölçütleri

Dağılım	**Sapma ölçütü**
Normal	$\sum (y-\hat{\mu})^2$
Binom	$2\sum \{[y \ln(y/\hat{\mu}] + (n-y)\ln[(n-y)/(n-\hat{\mu})]\}$
Poisson	$2\{\sum [y \ln(y/\hat{\mu}) - (y-\hat{\mu})]\}$
Gamma	$2\sum [-\ln(y/\hat{\mu}) + (y-\hat{\mu})/\hat{\mu}]$
Ters Normal	$\sum (y-\hat{\mu})^2/(\hat{\mu}^2 y)$

Çizelge 2.3'den de görüldüğü gibi Normal dağılım için sapma, artık kareler toplamına eşittir; Poisson dağılımı için Bishop ve diğerleri (1975) tarafından G^2 olarak adlandırılan bir istatistiktir. Ayrıca Poisson ve Gamma dağılımlarına ait ifadelerin ikinci terimleri genellikle sıfıra eşittir (Nelder ve Wedderburn, 1972).

Genelleştirilmiş doğrusal modellerde bir başka uyum iyiliği ölçütü de,

$$\chi^2 = \sum (y-\hat{\mu})^2 / V(\hat{\mu}) \tag{2.35}$$

biçiminde verilen genelleştirilmiş Pearson ki-kare istatistiğidir. Burada $V(\hat{\mu})$ ilgilenilen dağılıma ilişkin varyans fonksiyonunu ifade eder (McCullagh ve Nelder, 1988).

Veriye en uygun modele karar vermek için H_0 ve H_a hipotezleri aracılığıyla biri diğerinin daha genel biçimi olan iki ayrı model karşılaştırılır. H_0, M_0 gibi bir modeli, H_a ise M_a gibi daha genel bir modeli temsil etsin. G_0 ve G_a sırasıyla M_0 ve M_a

modellerine ait uyum istatistikleri olmak üzere sıfır hipotezi ve alternatif hipotez aşağıdaki gibidir.

$$H_0 : G_a = G_0$$

$$H_a : G_a \neq G_0$$

H_0 hipotezi kabul edilirse M_0 modeli, reddedilirse M_a modeli daha iyi bir model olarak kabul edilir.

2.4.2. Hipotez Testleri ve Güven Aralıkları

İstatistiksel çıkarımların önemli bir aşamasını da parametrelerin anlamlılığının test edilmesi ve parametrelere ait güven aralıklarının belirlenmesi oluşturur. Genelleştirilmiş doğrusal modellerde parametrelere ilişkin çıkarımların yapılmasında Ek 3'te açıklanan, En Çok Olabilirlik (EÇO) kestiricilerinin örneklem dağılımına dayanan Wald istatistiği kullanılır.

β parametreleri ile ilgili hipotezler, b kestiricilerinin örneklem dağılımı $N(\beta,I^{-1})$ ya da buna denk olarak $(b-\beta)^T I(b-\beta)$ ile verilen Wald istatistiği kullanılarak test edilir (Dobson, 1990).

Elimizde N gözlemli bir veri seti için q ve p parametreli iki model olsun. Bu modellerden hangisine ait parametrelerin modeli daha iyi açıkladığını test etmek isteyelim. Bu iki modele ilişkin sıfır ve seçenek hipotezler $q < p < N$ olmak üzere,

$$H_0 : \beta = \beta_0 = \begin{bmatrix} \beta_1 \\ \dots \\ \beta_q \end{bmatrix}$$

$$H_A : \beta = \beta_A = \begin{bmatrix} \beta_1 \\ \dots \\ \beta_p \end{bmatrix}$$

biçiminde yazılır.

D_0 ve D_1, sırasıyla, q ve p parametreli modellere ilişkin log-olabilirlik oran istatistiklerini göstermek üzere H_0 hipotezi, $\Delta D = D_0 - D_1$ farkı alınarak test edilebilir.

D_0 ve D_1 istatistikleri modeli iyi temsil ediyorsa, sırasıyla, χ^2_{n-q} ve χ^2_{n-p} dağılımlarına sahiptir. Bundan dolayı da, $\Delta D \sim \chi^2_{p-q}$ olacaktır. Eğer bu fark, χ^2_{p-q} dağılımına sahip ise daha basit olduğundan H_0 hipotezinde belirtilen model tercih edilir. Eğer ΔD, dağılımın kritik bölgesinde yer alıyorsa H_0 hipotezi reddedilir. Bu da β_A parametresinin modeli daha iyi açıkladığını gösterir (Dobson, 1990).

2.5. ARTIK İNCELEMESİ

Bir modelin bir veri kümesine uygunluğunun araştırılması, veri kümesindeki anormal değerlerin varlığının ve model için gerekli olan bazı varsayımların sağlanıp sağlanmadığının saptanması, artıkların incelenmesini gerektirir.

Üç tür genelleştirilmiş artık bulunur: Pearson, Anscombe ve Sapma (Deviance) artıkları. GDM'de artık incelemesinde genellikle sapma artıkları kullanılır. McCullagh ve Nelder'in (1988) belirttiği gibi bu artıklar aşağıdaki gibi tanımlanabilir:

Pearson Artıkları:

Pearson artığı y'nin standart sapması ile ile ölçeklendirilen bir artıktır ve

$$r_p = \frac{y - \mu}{\sqrt{V(\mu)}}$$

biçiminde tanımlanır.

Anscombe Artıkları:

Pearson artıklarının dezavantajı, Normal olmayan dağılımlar için r_p'lerin dağılımının çarpık olmasıdır. Bu nedenle, Normal olmayan dağılımlar söz konusu olduğunda, y'ler yerine A(y)'lerin kullanılarak tanımlanan, A(y)'lerin dağılımını olabildiğince normal kılacak bir A(.) fonksiyonunun seçimine dayanan Anscombe artıkları tercih edilir. A(.) fonksiyonu,

$$A(\cdot) = \int \frac{d\mu}{V^{\frac{1}{3}}(\mu)}$$

biçiminde tanımlanır.

Farklı dağılımlar için farklı şekiller alan Anscombe artıkları, Poisson dağılımı için,

$$r_A = \frac{\frac{3}{2}\left(y^{\frac{2}{3}} - \mu^{\frac{2}{3}} \right)}{\mu^{1/6}},$$

Gamma dağılımı için,

$$r_A = \frac{3\left(y^{1/3} - \mu^{1/3} \right)}{\mu^{1/3}}$$

ve Ters Normal dağılım için,

$$r_A = \left(\ln y - \ln \mu\right)/\mu$$

olarak tanımlanır (McCullagh ve Nelder, 1988).

2.6. GENELLEŞTİRİLMİŞ DOĞRUSAL MODEL TÜRLERİ

Genelleştirilmiş doğrusal modeller (GDM) rasgele bileşenin sahip olduğu dağılıma, bağlantı fonksiyonunun türüne ve sistematik bileşeni oluşturan bağımlı değişkenlerin kategorik ya da sürekli olma durumuna göre farklı adlar alırlar:

Çizelge 2.4. Genelleştirilmiş doğrusal model türleri

Rasgele Bileşen	**Bağlantı Fonksiyonu**	**Sistematik Bileşen**	**Model**
Normal	Birim	Sürekli	Regresyon
Normal	Birim	Kategorik	Varyans Analizi
Normal	Birim	Karma	Kovaryans Analizi
Binom	Lojit	Karma	Lojistik Regresyon
Multinomial	Lojit	Karma	Çoklu Yanıt
Poisson	Log	Karma	Poisson Regresyon
Negatif Binom	Log/Birim	Karma	Negatif Binom Regresyon
Gamma	Ters	Karma	Gamma Regresyon

İkili yanıt değişkenleri için kullanılan lojistik regresyon ve yanıt değişkenin Poisson dağılımına sahip olduğu durumda uygulanan Poisson regresyon modelleri genelleştirilmiş doğrusal modeller ailesinin en önemli iki üyesidir.

Çizelge 2.4'te de görüldüğü gibi bağımsız değişkenin Binom dağılımına sahip olduğu durumda lojit bağlantı fonksiyonu, Poisson ya da Negatif Binom dağılımlarına sahip olduğu durumda ise log bağlantı fonksiyonu kullanılır. Genel olarak, lojit bağlantı fonksiyonuna sahip modeller *lojit model olarak* adlandırılır.

Genelleştirilmiş doğrusal modeller frekansların modellenmesi için de kullanılır. Burada hücre frekanslarının dağılımına bakılır. Genelde bu dağılım Poisson dağılımına ya da Çok Terimli dağılıma uymaktadır. Hücre frekansları modellenirken sistematik bileşen sürekli ve kategorik değişkenlerden oluşuyorsa Poisson regresyon modeli, sadece kategorik değişkenlerden oluşuyorsa log-doğrusal modeller tercih edilir.

Log-doğrusal modellerin Poisson regresyon modelinden başka bir farkı da log-doğrusal modellerde bağımlı-bağımsız değişken ayrımı yapılmamasıdır.

Bir sonraki bölümde log-doğrusal modellerin yapısı, özellikleri ve olumsallık çizelgelerine uygulanması konularına değinilecektir.

3. OLUMSALLIK ÇİZELGELERİ VE LOG-DOĞRUSAL MODELLER

Bu bölümde kategorik verilerin gösterilmesinde sıkça yararlanılan olumsallık çizelgeleri ve olumsallık çizelgelerinin analizinde kullanılan log-doğrusal modeller anlatılacaktır.

3.1. OLUMSALLIK ÇİZELGELERİ

Olumsallık çizelgeleri, bir veri kümesine ait gözlemlerin değişken kategorilerine göre sınıflandırılmasından oluşan frekans çizelgeleridır. Bu tablolar çapraz sınıflandırma çizelgeleri olarak da adlandırılır. Olumsallık tablosu terimi ilk olarak 1904 yılında Karl Pearson tarafından kullanılmıştır (Agresti, 2002).

Kolay anlaşılır olması, yorum kolaylığı sağlaması ve nonparametrik olduğundan dolayı fazla varsayım gerektirmemesi nedeniyle olumsallık çizelgeleri en sık tercih edilen istatistiksel araçlardan biridir. Bu tablolar, özellikle sosyal bilimlerde sıkça karşılaşılan kategorik verilerin gösterilmesinde kullanılır.

Olumsallık çizelgeleri genellikle,

1. İki bağımsız değişkenin birleşik dağılımını ifade eden tablolar (örn: boy ve ağırlık),
2. Bir bağımlı değişken ile bir ya da birden fazla açıklayıcı değişken arasındaki neden- sonuç ilişkisini ifade etmek üzere kullanılan tablolar (örn: sigara içme ve akciğer kanseri arasındaki ilişkinin ortaya konması),
3. İki bağımlı değişken arasındaki ilişkinin ortaya konması için kullanılan olumsallık çizelgeleri

olarak üç şekilde kullanılır (Powers and Xie, 1999).

İki ya da daha çok düzeyli iki kategorik değişkenin çapraz sınıflandırılmasından oluşan tablolara iki yönlü olumsallık çizelgeleri; üç değişkenin çapraz sınıflandırılmasından oluşan tablolara üç yönlü olumsallık çizelgeleri; üçten fazla

değişkenin çapraz sınıflandırılmasından oluşan tablolara ise çok yönlü olumsallık çizelgeleri adı verilir.

Değişkenlerin kategori sayıları, tablonun satır ve sütun sayılarını gösterir. r kategorili bir X satır değişkeni ile s kategorili bir Y sütun değişkeninin çaprazlanmasından elde edilen tablonun satır sayısı r, sütun sayısı s olacak ve bu tablo r x s hücreden oluşacaktır.

X ve Y değişkenlerinin bileşik olasılıkları,

$$P(X = i, Y = j) = \pi_{ij} \tag{3.1}$$

olmak üzere, r satır ve s sütundan oluşan olumsallık tablosu Çizelge 3.1'de verildiği gibi gösterilsin. X ve Y değişkenlerinin marjinal olasılık toplamları sırasıyla, $\{\pi_{+1} \quad \pi_{+2} \quad \dots \quad \pi_{+s}\}$ ve $\{\pi_{1+} \quad \pi_{2+} \quad \dots \quad \pi_{r+}\}$ biçiminde olsun.

Çizelge 3.1. X ve Y değişkenlerine ait bileşik ve marjinal olasılık dağılımları

	Y	
X	$\pi_{11} \quad \pi_{12} \quad \dots \quad \pi_{1s}$ $\pi_{21} \quad \vdots \quad \vdots \quad \pi_{2s}$ $\vdots \quad \vdots \quad \vdots \quad \vdots$ $\pi_{r1} \quad \pi_{r2} \quad \dots \quad \pi_{rs}$	π_{1+} π_{2+} $\vdots$ π_{r+}
	$\pi_{+1} \quad \pi_{+2} \quad \dots \quad \pi_{+s}$	1

Eğer X değişkeni de Y gibi rasgele bir yanıt değişkeni ise bu durumda bileşik olasılık değerleri X ve Y değişkenleri arasındaki ilişkiyi verir. X değerlerine karşılık Y değişkenine ait koşullu olasılıklar,

$$\pi_{j/i} = \frac{\pi_{ij}}{\pi_{i+}}, \qquad \forall i, j \tag{3.2}$$

şeklindedir.

X ve Y değişkenlerinin bağımsızlığı durumunda tüm i ve j değerleri için,

$$\pi_{ij} = \pi_{i+}\pi_{+j} \tag{3.3}$$

koşulu gerçeklenir.

Eğer X ve Y değişkenleri bağımsız ise X değişkeni i değerini alırken Y değişkenine ait koşullu olasılık değeri, Y değişkenine ait marjinal olasılık değerine eşittir.

$$\pi_{j/i} = \frac{\pi_{ij}}{\pi_{i+}} = \frac{\pi_{i+}\pi_{+j}}{\pi_{i+}} = \pi_{+j} \tag{3.4}$$

(Rao ve Toutenburg, 1999).

3.1.1. Olumsallık Çizelgeleri için Örnekleme Dağılımları

İki ya da daha çok boyutlu olumsallık çizelgeleri için üç ana dağılım söz konusudur: Poisson dağılımı, Çok Terimli (Multinomial) dağılım ve Bağımsız Çok Terimli dağılım. y, N hücreli bir olumsallık tablosuna ait y_i hücre frekanslarını göstermek üzere bu dağılımlar aşağıda belirtildiği gibidir.

3.1.1.1. Poisson Dağılımı

Poisson dağılımı belirli bir zaman aralığında birbirinden bağımsız olarak gözlenen olay sayıları için kullanılır. Gözlem sayısı için herhangi bir sınır yoktur. Poisson modeli, y_i rastlantı değişkenleri ve $E(y_i) = \mu_i$ parametre olmak üzere bileşik olasılık fonksiyonu,

$$f(\mathbf{y};\boldsymbol{\mu}) = \prod_{i=1}^{N} \mu_i^{y_i} e^{-\mu_i} \Big/ y_i! \tag{3.5}$$

olan modeldir.

Olumsallık çizelgeleri için $\boldsymbol{\mu}$, μ_i beklenen hücre frekanslarının oluşturduğu bir vektördür. *ixj* boyutlu bir çapraz sınıflama tablosu için (i=1,…r ve j=1,…,s) (3.5) eşitliği, y_{ij} hücre frekanslarının oluşturduğu vektör $\mathbf{y}$ ve m_{ij} beklenen hücre frekanslarının oluşturduğu vektör $\mathbf{m}$ olmak üzere,

$$f(\mathbf{y};\mathbf{m}) = \prod_{i,j=1}^{r,s} m_{ij}^{y_{ij}} e^{-m_{ij}} \Big/ y_{ij}! \tag{3.6}$$

olarak da yazılabilir (Dobson, 1990; Rao ve Toutenburg, 1999).

3.1.1.2. Çok Terimli (Multinomial) Dağılım

Çok terimli dağılım, binom dağılımının genelleştirilmiş biçimidir (Powers and Xie, 1999). Bu dağılım, satır ve sütun toplamlarının sabit olmayıp örneklem büyüklüğünün sabit olduğu durumlarda kullanılır. Çok terimli dağılıma ait olasılık yoğunluk fonksiyonu,

$$f(\mathbf{y};\boldsymbol{\pi}/n) = n!\prod_{i=1}^{N} \pi_i^{y_i} \Big/ y_i! \tag{3.7}$$

biçimindedir. Burada, $n = \sum_{i=1}^{N} y_i$ ve $\sum_{i=1}^{N} \pi_i = 1$ dir ve π_i hücre olasılıklarını belirtir. Bu durumda, $E(Y_i) = n\pi_i$'dir.

3.1.1.3.Bağımsız Çok Terimli Dağılım

Satır toplamlarının sabit olduğu durumda söz konusu olan *bağımsız çok terimli (product multinomial) dağılım* için de olasılık yoğunluk fonksiyonu,

$$\frac{y_i!}{\prod_j y_{ij}} \prod_j \pi_{j/i}^{y_{ij}} \tag{3.8}$$

olarak verilir.

Üç yönlü bir olumsallık tablosu için bu dağılıma ait olasılık yoğunluk fonksiyonu ,

$$f(\mathbf{y};\boldsymbol{\pi}/y_{j.l}, i = 1,...,J; l = 1,...,L) = \prod_{j=1}^{J}\prod_{l=1}^{L} y_{j.l}! \prod_{k=1}^{K} \pi_{jkl}^{y_{jkl}} \Big/ y_{jkl}! \tag{3.9}$$

şeklindedir (Dobson, 1990).

Poisson dağılımının en önemli özelliği dağılımın ortalamasının dağılımın varyansına eşit olmasıdır. Ancak bazen dağılımın varyansı ortalamadan daha yüksek olabilmektedir. Bu durum *aşırı yayılım* olarak adlandırılır. Aşırı yayılım durumunda Poisson dağılımı Negatif Binom (Hipergeometrik) dağılımına dönüşür.

3.1.2. Olumsallık Çizelgeleri için Beklenen Hücre Frekansı Kestirim Yöntemleri

3.1.2.1. En Çok Olabilirlik (EÇO) Kestiricileri

Olumsallık çizelgelerinde verilen bir örneklem modeli için EÇO fonksiyonu, gözlenen hücre frekansları y_i'lere ait (i=1,…,N) hücre olasılıkları olarak tanımlanır

(Rao ve Toutenburg, 1999). EÇO kestiricileri ise gözlenen veriye ait EÇO fonksiyonunun maksimum değerini almasıyla elde edilir.

Bir önceki kısımda belirtilen olumsallık çizelgelerinde karşılaşılan her üç dağılıma ilişkin olabilirlik fonksiyonları da benzer olduğundan, bu dağılımlara ait EÇO kestiricileri özdeştir (Powers and Xie,1999).

X ve Y değişkenlerinin bağımsızlığı varsayımı altında EÇO kestiricileri, örneklem olasılıkları "p" ve satır ve sütun toplamları, sırasıyla, y_{i+} ve y_{+j} olmak üzere,

$$\hat{\pi}_{ij} = p_{i+}p_{+j} = \frac{y_{i+}y_{+j}}{n^2} \tag{3.10}$$

biçiminde hesaplanır.

Buradan, beklenen hücre frekansları da,

$$\hat{m}_{ij} = n\hat{\pi}_{ij} = n\pi_{i+}\pi_{+j} = \frac{y_{i+}y_{+j}}{n} \tag{3.11}$$

eşitliğinden kestirilir.

Olumsallık çizelgeleri için en EÇO fonksiyonunun maksimizasyonu ve bu fonksiyon aracılığıyla beklenen hücre frekanslarının elde edilişi Ek 4'te ayrıntılı olarak açıklanmıştır.

Olumsallık çizelgelerinde bazı hücrelerin boş olması ya da bazı hücrelerin sıfır değerini alması durumunda bağımsız EÇO kestiricileri, (3.10) denklemiyle elde edilemez. Bu durumda kullanılmak üzere Goodman, 1965 yılında iteratif bir yöntem vermiştir. Bu yöntem daha sonra Bishop tarafından geliştirilmiştir (Bishop ve Fienberg, 1969).

3.1.2.2. "Medyan Polish" Kestiricileri

Olumsallık çizelgeleri için EÇO kestiricileri dışında Medyan Polish kestiricileri de mevcuttur. Bu kestiriciler ilk olarak Mosteller ve Parunak (1985) tarafından kullanılmıştır. Medyan Polish kestirim yöntemi satır ve sütun medyanlarına dayanarak elde edildiğinden sağlam bir kestirim yöntemidir. Bu yöntem temel olarak beş aşamadan oluşur:

1. Önce olumsallık tablosunun her bir satırına ait ortanca değerleri hesaplanır.

2. Hesaplanan satır ortancaları sırasıyla her bir satır elemanından çıkarılır.
3. Her bir sütunun ortanca değeri hesaplanır.
4. Hesaplanan sütun ortancaları sırasıyla her bir sütun elemanından çıkarılır.
5. Bu algoritma satır veya sütun ortancanın değeri yaklaşık olarak sıfır bulununcaya dek sürdürülür (Hoaglin, Mosteller ve Tukey, 1983).

Son aşamada elde edilen değerler hücre artık değerlerini verir. Bu değerlerin 1. aşamadaki gözlenen hücre değerlerine eklenmesiyle Medyan Polish kestiricileri bulunmuş olur.

Medyan Polish kestirimi sağlam bir kestirim yöntemi olduğundan olumsallık çizelgelerinde özellikle aykırı değer incelemesinde tercih edilmektedir (Kuhnt ve Pawlitschko, 2005; Kuhnt, 2010).

3.1.2. Olumsallık Çizelgeleri için Uyum İyiliği

Olumsallık çizelgelerinde uyum iyiliği için iki istatistik kulanılır. Bunardan en sık kullanılanı 1900 yılında Pearson tarafından verilen, gözlenen ve beklenen hücre frekansları arasındaki farka dayalı olarak,

$$\chi^2 = \sum_{i,j}^{r,s} \frac{(y_{ij} - \hat{m}_{ij})^2}{\hat{m}_{ij}} \tag{3.12}$$

biçiminde ifade edilen Pearson ki-kare istatistiğidir. Bu istatistiği kullanırken dikkat edilmesi gereken husus hücre frekanslarının en az beş olması gerekliliğidir (Fienberg ve Rinaldo, 2007).

Uyum iyiliğinin belirlenmesinde kullanılan ikinci bir istatistik de

$$G^2 = 2\sum_{i,j}^{r,s} y_{ij} \log\left(\frac{y_{ij}}{\hat{m}_{ij}}\right) \tag{3.13}$$

olarak tanımlanan en çok olabilirlik oran istatistiğidir. Bu istatistik 1935 yılında Wilks tarafından ortaya konulmuştur. H_0 hipotezi, G^2 istatistiğinin büyük değerleri için red, küçük değerleri için kabul edilir. Örneklem büyüklüğünün yeteri kadar büyük olması durumunda yukarıda belirtilen her iki istatistik de, N hücre sayısı ve p bağımsız parametre sayısı olmak üzere N-p serbestlik dereceli χ^2 dağılımına yakınsar.

Örneklem büyüklüğünün yeteri kadar büyük olmadığı durumda uyum iyiliğinin belirlenmesinde hangi istatistiğin kullanılması gerektiği konusunda bir netlik yoktur (Fienberg, 1970b).

3.2. LOG-DOĞRUSAL MODELLER

Kategorik veri analizinde önemli bir araç olan log-doğrusal modeller 1900'lü yıllarda geliştirilmiştir. Tıp, mühendislik ve sosyal bilimlerde çok popüler olan log-doğrusal modeller değişkenlerin kategorik olduğu durumda değişkenler arası ilişkilerin araştırılması ve modelleme amaçlı kullanılır.

İkinci bölümde değinilen genelleştirilmiş doğrusal modellerin Poisson dağılımlı veriler için kullanılan özel bir durumu olan log-doğrusal modeller iki yönlü olumsallık çizelgelerinde kategorik değişkenler arasındaki ilişki örüntüsünün araştırılması ve hücre frekanslarının modellenmesi amaçları ile kullanılır. Yanıt değişkeni ve açıklayıcı değişkenler arasında ayrım yapılamadığı durumlarda bu yöntemin kullanılması daha uygundur (McCulloch ve Searle, 2001).

Log-doğrusal modellemenin en önemli noktası kestirim değerlerinin nasıl elde edileceğidir. Olumsallık tablosu şeklinde ifade edilebilen verilerde 3.1.2.'de değinilen kestirim yöntemleri kullanılır.

Log-doğrusal modelleme işlemi başlıca beş adımdan oluşur:

1. Veriyi açıklayan bir model önerilir.

2. Modelin uygunluğu varsayımı altında beklenen hücre değerleri hesaplanır.

3. Gözlenen hücre değerleri, uyum iyiliği ölçütleri olan ki-kare ya da olabilirlik oran istatistiği kullanılarak beklenen hücre değerleri ile karşılaştırılır.

4. Modelin kabul edilip edilmediği saptanılır.

5. Model kabul ediliyorsa sonuçlar yorumlanır; model reddediliyorsa başka bir modelin uygunluğu denenir; yani 1. adıma geri dönülür (Burnett, 1983).

Log-doğrusal modeller olumsallık çizelgelerine uygulanırken, gözlenen olay sayısının, hücre sayısının en az 5 katı olmasına dikkat edilmelidir. Örneğin, 2x2x3 boyutlu bir tablo için olay sayısının en az 60 olması gerekmektedir (Jeansonne, 2007).

Log-doğrusal modeller genelde iki yanıt değişkeninin olması ya da bağımlı ve bağımsız değişkenler arasında ayrım yapılamadığı durumlarda tercih edilir. İki kategorili tek yanıt değişkeni olması durumunda log-doğrusal modeller lojit modele dönüşür (Nelder, 2000). Lojit modeller ile ilgili ayrıntılı bilgi Dobson (1990), Christensen (1997), Rao ve Toutenburg (1999), Nelder (2000), Agresti (2002, 2007) kaynaklarında bulunabilir. Bu bölümde, yanıt değişkeninin ikiden fazla kategoriye sahip olduğu durumda kullanılan log-doğrusal modeller ve bu modellerin olumsallık çizelgelerinin analizinde kullanımı ele alınacaktır.

3.2.1. İki Yönlü Olumsallık Çizelgeleri için Log-Doğrusal Modeller

Log-doğrusal modeller, GDM'nin özel bir durumudur. Bölüm 2'de yanıt değişkeni Poisson dağılımına sahip olan verilerin modellenmesinde log-doğrusal modellerin kullanıldığından bahsedilmişti. Olumsallık çizelgeleri için de hücre frekanslarının Poisson dağılımı başta olmak üzere Bölüm 3.2.1.'de belirtilen dağılımlardan herhangi birine sahip olması durumunda log-doğrusal modeller kullanılır.

Olumsallık çizelgelerinde rastlanılan Çok Terimli ve Poisson dağılımları birbirleri türünden ifade edilebilir (EK 5). Bundan dolayı, her iki dağılım için de GDM'de Poisson dağılımı için uygun olan log bağlantı fonksiyonu yani doğal logaritma fonksiyonu kullanılabilir. Bu fonksiyonun kullanılmasıyla, $E(y_i)$, parametre vektörü ve sistematik bileşenlere ait vektörün transpozunun çarpımı olarak,

$$\eta_i = \log E(y_i) = x_i^T \beta \,, \qquad \text{i} = 1, \ldots, \text{N} \quad (3.14)$$

şeklinde ifade edilebilir. (3.14) ile verilen tüm bu genelleştirilmiş doğrusal modeller log- doğrusal modeller olarak adlandırılır.

Poisson dağılımına sahip tek bir yanıt değişkeni varsa log-doğrusal model, Poisson regresyona eşdeğerdir. Poisson regresyon modelinin log-doğrusal modelden en önemli farkı, log-doğrusal modellemede kullanılan satır, sütun ve marjinal toplamların rolünü Poisson regresyon yönteminde açıklayıcı nicel değişkenlerin alıyor olmasıdır (Lawal, 2003).

Çok terimli dağılımda, iki değişken bağımsız ise bu değişkenlere ait bileşik olasılık değerinin marjinal olasılıkların çarpımı olarak (3.3)'teki gibi yazılabileceği bilinmektedir. Hücre olasılıkları π_{ij} ve beklenen hücre değerleri $\mu_{ij} = n\pi_{ij}$ olmak

üzere, $E(y_{ij}) = \mu_{ij}$ eşitliğini gerçekleyen bağımsız y_{ij} gözlemlerinden oluşan $N = rs$ hücreli bir olumsallık tablosu için, log bağlantı fonksiyonunun (3.11) ile verilen beklenen hücre frekansları üzerine uygulanmasıyla,

$$\log E(y_{ij}) = \log(\hat{m}_{ij}) = \log n + \log \pi_{i+} + \log \pi_{+j} \tag{3.15}$$

elde edilir.

Hücre olasılıkları üzerinden elde edilen (3.15) denklemini gözlenen satır ve sütun toplamları olan y_{i+} ve y_{+j} değerlerini temel alarak yazmak da mümkündür. Bu durumda, yine (3.11) eşitliği kullanılarak,

$$\log E(y_{ij}) = \log y_{i+} + \log y_{+j} - \log n \tag{3.16}$$

olur.

(3.15) ve (3.16) ile verilen eşitliklerden yola çıkarak her hücre için beklenen frekansların logaritmalarını üç terimin toplamı olarak ifade etmek mümkündür: Örneklem büyüklüğüne dayalı sabit bir terim i satırlarının (i = 1,..,r) marjinal olasılıkları ile ilişkili bir terim ve j sütunlarının (j = 1,...,s) marjinal olasılıkları ile ilişkili bir terim (Burnett, 1983).

Tüm $\hat{m}_{ij}$ hücre frekanslarının pozitif olduğu varsayıldığında, boş olmayan hücreler için $\hat{m}_{ij}$ frekansları (3.17) deki gibi gösterilebilir.

$$\hat{\eta}_{ij} = \log \hat{m}_{ij} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} \tag{3.17}$$

Burada,

$$\sum_{i=1}^{r} \alpha_i = \sum_{j=1}^{s} \beta_j = \sum_{k=1}^{r} (\alpha\beta)_{kj} = \sum_{k=1}^{s} (\alpha\beta)_{ik} = 0 \tag{3.18}$$

koşulları gerçeklenir (Bishop ve Fienberg, 1969).

(3.17) denklemi etkileşim teriminin de dahil olduğu tam modeli verir. Etkileşim terimi $(\alpha\beta)_{ij}$'nin dahil olmadığı model, toplamsal (bağımsız) log-doğrusal model olarak adlandırılır. Bu modeldeki ortalama, satır ve sütun etkilerinin kestirimleri sırasıyla,

$$\hat{\mu} = \frac{1}{N} \sum\sum \hat{\eta}_{ij} = \frac{1}{r} \sum \log y_{i+} + \frac{1}{s} \sum \log y_{+j} - \log n \tag{3.19}$$

$$\hat{\alpha}_i = \log y_{i+} - \frac{1}{I}\sum \log y_{i+} \tag{3.20}$$

$$\hat{\beta}_j = \log y_{+j} - \frac{1}{J}\sum \log y_{+j} \tag{3.21}$$

denklemlerinden bulunur.

Bu ilişkinin genelleştirilmesine dayanarak, iki yönlü bir olumsallık tablosu için toplamsal log-doğrusal model, μ ortalama etki α_i satır ve β_j sütun etkileri ya da başka bir deyişle ana etkiler olmak üzere,

$$\log(\hat{m}_{ij}) = \mu + \alpha_i + \beta_j \tag{3.24}$$

şeklinde yazılabilir.

X ve Y değişkenlerinin bağımsız olmaması durumunda (3.24) ile ifade edilen denkleme bir de etkileşim terimi eklenerek, tam model

$$\log(\hat{m}_{ij}) = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} \tag{3.25}$$

biçimini alır.

İki yönlü bir olumsallık tablosu için bağımsız log-doğrusal modele ait parametre sayısı 1 + (I-1) + (J-1) = I + J - 1 bulunurken; tam modele ait parametre sayısı 1 + (I-1) + (J-1) + (I-1) (J-1) =I J dir.

İki yönlü olumsallık çizelgeleri için en çok kullanılan log-doğrusal modeller Çizelge 3.2.'deki gibidir (Dobson, 2002).

Çizelge 3.2. İki yönlü olumsallık çizelgeleri için log-doğrusal modeller

Log-doğrusal model	***Poisson Dağılımı***	***Çok Terimli Dağılım (n sabit)***	***Çarpımsal Çok Terimli Dağılım ($y_{i.}$ sabit)***
JK parametreli tam model $\mu+\alpha_i+\beta_j+(\alpha\beta)_{ij}$	$E(Y_{ij})=\mu_{ij}$	$E(Y_{ij})=n\pi_{ij}$	$E(Y_{ij})=y_{i.}\pi_{ij}$
J+K–1 parametreli bağımsız model $\mu+\alpha_i+\beta_j$	$E(Y_{ij})=\alpha_i\beta_j$	$E(Y_{ij})=n\pi_{i.}\pi_{.j}$	$E(Y_{ij})=y_{i.}\pi_{.j}$

3.2.2. Üç Yönlü Olumsallık Çizelgeleri için Log-Doğrusal Modeller

Sırasıyla r, s ve t kategorili X, Y, Z gibi üç değişkenin çapraz sınıflandırılmasından oluşan çizelgeler üç boyutlu olumsallık çizelgeleri olarak adlandırılır. Üç yönlü olumsallık çizelgelerinde, iki yönlü olumsallık çizelgelerinde bulunan X satır ve Y sütun değişkenlerinin yanı sıra bir de Z katman değişkeni bulunur.

Log-doğrusal modeller, değişkenler arası ilişkilerin ortaya çıkarılmasının daha zor olduğu üç ya da daha çok boyutlu olumsallık çizelgelerinin modellenmesinde önemli rol oynar.

Üç yönlü tablolarda değişken sayısına bağlı olarak üç tip bağımsızlık söz konusudur: karşılıklı bağımsızlık, bileşik bağımsızlık ve koşullu bağımsızlık. Her bağımsızlık türü için hücre olasılıkları ve oluşturulan model farklılık gösterir.

1. Yukarıda belirtilen üç durum için hücre olasılıkları tüm i, j, k değerleri için aşağıda X,Y,Z değişkenlerinin karşılıklı bağımsızlığı durumu:

$$\pi_{ijk} = \pi_{i++}\pi_{+j+}\pi_{++k} \tag{3.24}$$

2. Bileşik bağımsızlık durumu:

$$\pi_{ijk} = \pi_{i+k}\pi_{+jk} \tag{3.25}$$

3. Koşullu bağımsızlık durumu:

$$\pi_{ijk} = \frac{\pi_{i+k}\pi_{+jk}}{\pi_{++k}} \tag{3.26}$$

Eğer X, Y, Z değişkenleri birbirlerinden bağımsız ise $\hat{m}_{ijk}$ hücre frekansları aşağıdaki şekilde modellenir:

$$\log(\hat{m}_{ijk}) = \mu + \alpha_i + \beta_j + \gamma_k \tag{3.27}$$

Bu durum, *karşılıklı bağımsızlık* durumu olarak adlandırılır.

Eğer değişkenlerden biri diğer iki değişkenin olası birleşimlerinden bağımsız ise bu duruma *bileşik dağılımından bağımsızlık* durumu denilir. Z değişkeni ile X ve Y değişkenlerinin r x s birleşimleri arasında bağımsızlık olduğu model,

$$\log(\hat{m}_{ijk}) = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} \tag{3.28}$$

olarak gösterilir.

Değişkenlerden biri sabit olmak üzere diğer iki değişkenin *koşullu bağımsızlığı* altında olası log-doğrusal modellerden biri,

$$\log(\hat{m}_{ijk}) = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\beta\gamma)_{jk} \tag{3.29}$$

olarak gösterilebilir. (3.29) ile verilen bu modelde Y değişkeni sabit iken X ve Z koşullu bağımsızdır.

Log-doğrusal bağımsızlık modelleri Çizelge 3.3'teki gibi özetlenebilir (Agresti, 2002).

Çizelge 3.3. Log-doğrusal bağımsızlık modelleri

Model	**Modele ait etkileşim terimi**	**Yorum**
$\log(\hat{m}_{ijk}) = \mu + \alpha_i + \beta_j + \gamma_k$	Yok	Değişkenler karşılıklı bağımsız
$\log(\hat{m}_{ijk}) = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\gamma)_{ik}$	$(\alpha\gamma)_{ik}$	Y, X ve Z değişkenlerine bağımlı
$\log(\hat{m}_{ijk}) = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk}$	$(\alpha\gamma)_{ik} + (\beta\gamma)_{jk}$	Z değişkeni sabitken, X ve Y bağımsız

Üç yönlü olumsallık çizelgeleri için en genel (tam) model ise

$$\log(m_{ijk}) = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + (\alpha\beta\gamma)_{ijk} \qquad (3.30)$$

biçimindedir. Burada $(\alpha\beta\gamma)_{ijk}$ ile gösterilen son terim *üç faktörlü etkileşim terimi*dir. Tam modele ait toplam parametre sayısı 1 + (I-1) + (J-1) + (K-1) +(I-1) (J-1) +(I-1) (K-1) + (J-1) (K-1) + (I-1) (J-1) (K-1) = I J K' dır.

Bu modelde ortalamadan sapmayı ifade eden tüm etkileşim parametreleri ve ana faktör etkileri için,

$$\sum_{i=1}^{r}(\alpha\beta)_{ij} = \sum_{j=1}^{s}(\alpha\beta)_{ij} = \cdots = \sum_{k=1}^{t}(\alpha\beta\gamma)_{ijk} = 0 \qquad (3.31)$$

$$\sum_{i=1}^{r}\alpha_i = \sum_{j=1}^{s}\beta_j = \sum_{k=1}^{t}\gamma_k = 0 \qquad (3.32)$$

kısıtları gerçeklenir (Rao ve Toutenburg, 1999).

Log-doğrusal modellerde hiyerarşi ilkesi söz konusudur. Yani, yüksek dereceli bir etki modele katılıyorsa anlamlı olup olmadığına bakılmaksızın, tüm alçak dereceli etkiler de modele dahil edilmelidir. Örneğin, bir model $(\alpha\beta)_{ij}$ etkileşim parametresini içeriyorsa α_i ve β_j parametrelerini de içermelidir.

Üç yönlü olumsallık çizelgeleri için modele dahil edilen değişkenlere göre pek çok olası log-doğrusal model bulunur. Bu olası modelleri değişkenlere bağlı olarak simgelerle göstermek mümkündür. Bazı olası model türleri ve bu modellerin gösterilişleri Çizelge 3.4'te verilmiştir.

Çizelge 3.4. Üç Yönlü olumsallık çizelgeleri için log-doğrusal modeller

Log-Doğrusal Model	Sembol
$\log(\hat{m}_{ijk}) = \mu + \alpha_i + \beta_j + \gamma_k$	(X, Y, Z)
$\log(\hat{m}_{ijk}) = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij}$	(XY, Z)
$\log(\hat{m}_{ijk}) = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\beta\gamma)_{ik}$	(XY, YZ)
$\log(\hat{m}_{ijk}) = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk}$	(XY, YZ, XZ)
$\log(\hat{m}_{ijk}) = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + (\alpha\beta\gamma)_{ijk}$	(XYZ)

3.2.3. Üç Yönlü Olumsallık Çizelgeleri İçin Beklenen Hücre Frekansı Kestirim Yöntemleri

3.2.3.1. En Çok Olabilirlik (EÇO) Kestiricileri

Çizelge 3.4'te verilen farklı etkileşim terimlerinin modele katıldığı durumlarda, beklenen hücre frekansları farklı algoritmalar kullanılarak elde edilir.

Tam bağımsızlık modeli altında beklenen hücre frekansları, i=1,…,r, j=1,…,s ve k=1,…,t olmak üzere marjinal toplamlar aracılığıyla,

$$\hat{m}_{ijk} = \left(\frac{y_{i++}}{N}\right) \times \left(\frac{y_{+j+}}{N}\right) \times \left(\frac{y_{++k}}{N}\right) \times N \tag{3.33}$$

denkleminden elde edilir.

(3.28) eşitliğindeki gibi iki faktörlü tek bir etkileşim teriminin bulunduğu, değişkenlerden birinin diğer ikisinden bağımsız olduğu durumda tüm i,j,k değerleri için beklenen hücre frekansları,

$$\hat{m}_{ijk} = \left(\frac{y_{ij+}}{N}\right)\left(\frac{y_{++k}}{N}\right) N \tag{3.34}$$

kullanılarak hesaplanır. (3.34) denklemi, Z değişkeninin X ve Y değişkenlerinden bağımsız olduğu (3.29) modelinin en çok olabilirlik kestiricilerini ifade eder.

İki etkileşim teriminin modele katılmasıyla (3.28)'e dönüşen modele ait en çok olabilirlik kestiricileri tüm i,j,k değerleri için,

$$\hat{m}_{ijk} = \frac{\left(y_{ij+}\right)\left(y_{i+k}\right)}{y_{i++}} \tag{3.35}$$

şeklinde hesaplanır.

Tüm ikili etkileşim terimlerini içeren

$$\log(\hat{m}_{ijk}) = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} \qquad (3.36)$$

modeli için EÇO kestiricilerini tek aşamada elde etmek mümkün değildir; yinelemeli yöntem gerektirir. Bu durumda şöyle bir yol izlenir:

1. Beklenen hücre değerlerini bulmak için gerekli olan y_{ij+}, y_{i+k} ve y_{+jk} marjinal toplamları elde edilir.
2. Her hücrenin başlangıç değeri, $\hat{m}_{ijk}^{(0)} = 1$ olarak kabul edilir.
3. 2.adımda elde edilen başlangıç hücre değerleri y_{ij+} marjinal toplamı ile çarpımsal işleme girerek 1. aşama beklenen hücre frekansları bulunur.

$$\hat{m}_{ijk}^{(1)} = \hat{m}_{ijk}^{(0)} . \left(\frac{y_{ij+}}{\hat{m}_{ij+}^{(0)}} \right) \qquad (3.38)$$

4. 3. adımda elde edilen hücre değerleri y_{i+k} marjinal toplamı ile işleme girerek 2. aşama beklenen hücre frekansları bulunur.

$$\hat{m}_{ijk}^{(2)} = \hat{m}_{ijk}^{(1)} . \left(\frac{y_{i+k}}{\hat{m}_{i+k}^{(1)}} \right) \qquad (3.39)$$

5. 4. adımda elde edilen değerler son olarak y_{+jk} marjinal toplamı ile işleme girer ve son aşamada beklenen hücre frekansları (3.40)'taki gibi elde edilir.

$$\hat{m}_{ijk}^{(3)} = \hat{m}_{ijk}^{(2)} . \left(\frac{y_{+jk}}{\hat{m}_{+jk}^{(2)}} \right) \qquad (3.40)$$

6. 3, 4 ve 5 numaralı adamlar beklenen hücre frekanslarının elde edilmesinde birinci döngüyü oluşturur. Bu işlem bir önceki adımdan elde edilen değerlerin bir sonraki adımda başlangıç değeri olarak kullanılmasıyla devam eder. Her döngünün sonunda elde edilen değer bir önceki adımda elde edilen değerle karşılaştırılır ve değişim miktarına bakılır. Eğer değişim yeteri kadar azsa döngüye son verilir, aksi halde döngü değişim miktarı azalana kadar devam eder.

Bu döngü Deming ve Stephan (1940) tarafından "Census" adlı veri kümesine uygulanarak tanıtılmıştır (Fienberg, 1970b).

3.2.3.2. "Medyan Polish" (MedPol) Kestiricileri

Üç yönlü olumsallık çizelgelerinde de iki yönlü olumsallık çizelgelerindekine benzer şekilde Medyan Polish kestiricilerini elde etmek mümkündür. Üç yönlü olumsallık çizelgeleri için MedPol kestiricileri iki yönlü olumsallık çizelgelerindekinden biraz daha farklı şekilde elde edilir.

μ ortalama etkiyi, α_i, β_j ve γ_k sırasıyla X,Y,Z değişkenlerinin model üzerindeki etkilerini belirtmek üzere,

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + \varepsilon_{ijk} \tag{3.41}$$

toplamsal modeline sahip üç yönlü bir olumsallık çizelgesini ele alalım.

$r_{ijk}^{(q)}$, q. aşamada elde edilen hücre artık değerleri ve $a_i^{(q)}$, $b_j^{(q)}$ ve $c_k^{(q)}$ sırasıyla X, Y, Z değişkenleri için q. aşamada kestirilen etki değerleri olsun. Tüm i,j,k değerleri için başlangıç artık ve etki değerleri sıfıra eşittir:

$$a_i^{(0)} = b_j^{(0)} = c_k^{(0)} = 0 \tag{3.42}$$

$$r_{ijk}^{(0)} = 0 \tag{3.43}$$

(3.41) ile verilen değişkenlerin bağımsız olduğu model için Medyan Polish kestiricileri aşağıdaki algoritma uygulanarak elde edilir:

1. Sütun değişkeninin her kategorisi için ortanca değerleri (med) bulunur.

$$a_i^{(1)} = \underset{jk}{med}\{r_{ijk}^{(0)}\}, \qquad i = 1,...,r \tag{3.44}$$

2. Bulunan ortanca değerleri, bulundukları sütuna ait hücre değerlerinden çıkarılarak yeni hücre değerleri elde edilir:

$$r_{ijk}^{(0)} - \underset{jk}{med}\{r_{ijk}^{(0)}\} \tag{3.45}$$

3. Satır değişkenlerinden biri sütun değişkeni ile yer değiştirecek biçimde tablo yeni elde edilen hücre frekansları üzerinden tekrar düzenlenir.

4. Düzenlenen yeni tabloya ait sütun değişkeninin kategorilerine göre sütun ortancaları alınır ve bulundukları sütun hücre değerlerinden çıkarılır ve yeni hücre değerleri elde edilir.

$$b_j^{(1)} = \underset{ik}{med}\{r_{ijk}^{(0)} - a_i^{(1)}\}, \qquad j = 1,...,s \tag{3.46}$$

5. Elde edilen yeni hücre değerleri geriye kalan son değişken sütun değişkeni olacak şekilde oluşturulan tabloya yerleştirilir.

6. Yeni sütun değişkenine göre 1. adımın uygulanması ile $c_k^{(1)}$ ortanca değerleri bulunur:

$$c_k^{(1)} = \underset{ij}{med}\{r_{ijk}^{(0)} - a_i^{(1)} - b_j^{(1)}\}, \quad k = 1,...,t \tag{3.47}$$

7. 6. adımda bulunan hücre değerlerinin gözlenen hücre değerlerinden çıkarılması ile 1. aşama sonunda elde edilen hücre artık değerleridir:

$$r_{ijk}^{(1)} = r_{ijk}^{(0)} - a_i^{(1)} - b_j^{(1)} - c_k^{(1)} \tag{3.48}$$

Ortanca değerlerinin yaklaşık olarak sıfır bulunduğu aşamaya dek bu algoritmaya devam edilir. Son aşamada elde edilen değerler hücre artık değerleridir. Bu artıkların gözlenen hücre frekanslarına eklenmesiyle "*medyan polish" kestiricileri* elde edilmiş olur.

Birinci aşama sonundaki etki kestirim değerlerini elde etmek için ise önce her bir değişken için farklı adımlarda elde edilen etki değerleri kümesinin ortancaları alınarak merkezileştirme yapılır:

$$m_a^{(1)} = med\{a_i^{(1)}\} \tag{3.49}$$

$$m_b^{(1)} = med\{b_j^{(1)}\} \tag{3.50}$$

$$m_c^{(1)} = med\{c_k^{(1)}\} \tag{3.51}$$

Buradan, 1. aşama sonundaki etki kestirim değerleri,

$$\hat{\alpha}_i^{(1)} = a_i^{(1)} - m_a^{(1)} \tag{3.52}$$

$$\hat{\beta}_j^{(1)} = b_j^{(1)} - m_b^{(1)} \tag{3.53}$$

$$\hat{\gamma}_k^{(1)} = c_k^{(1)} - m_c^{(1)} \tag{3.54}$$

$$\hat{\mu}^{(1)} = m_a^{(1)} + m_b^{(1)} + m_c^{(1)} \tag{3.55}$$

olarak bulunur.

Bu algoritmaya n. döngüye kadar devam edilirse,

$$a_i^{(n)} = \underset{jk}{med}\{r_{ijk}^{(n-1)}\}$$

$$b_j^{(n)} = \underset{ik}{med}\{r_{ijk}^{(n-1)} - a_i^{(n)}\}$$

$$c_k^{(n)} = \underset{ij}{med}\{r_{ijk}^{(n-1)} - a_i^{(n)} - b_j^{(n)}\}$$

ve

$$A_i^{(n)} = a_i^{(1)} + a_i^{(2)} + ... + a_i^{(n)} \tag{3.56}$$

$$B_j^{(n)} = b_j^{(1)} + b_j^{(2)} + ... + b_j^{(n)} \tag{3.57}$$

$$C_k^{(n)} = c_k^{(1)} + c_k^{(2)} + ... + c_k^{(n)} \tag{3.58}$$

değerleri elde edilir. Burada, $A_i^{(n)}$, $B_j^{(n)}$ ve $C_k^{(n)}$ sırasıyla X,Y ve Z değişkenlerine ait merkezileştirilmemiş etki değerleridir.

t.döngüde algoritmanın sıfıra yakınsadığını varsayarsak, son adımdaki merkezileştirilmiş etki kestirim değerleri,

$$\hat{\alpha}_i^{(t)} = A_i^{(t)} - \underset{i}{med}\{A_i^{(t)}\}, \qquad i = 1,...,r \tag{3.59}$$

$$\hat{\beta}_j^{(t)} = B_j^{(t)} - \underset{j}{med}\{B_j^{(t)}\}, \qquad j = 1,...,s \tag{3.60}$$

$$\hat{\gamma}_k^{(t)} = C_k^{(t)} - \underset{k}{med}\{C_k^{(t)}\}, \qquad k = 1,...,t \tag{3.61}$$

denklemleri aracılığıyla hesaplanır.

Bu yöntem aşağıda belirtilen koşulların gerçeklenmesini gerektirir (Cook, 1985):

$$\underset{jk}{med}\{r_{ijk}\} = 0 \text{ , tüm i değerleri için}$$

$$\underset{ik}{med}\{r_{ijk}\} = 0 \text{ , tüm j değerleri için}$$

$$\underset{ij}{med}\{r_{ijk}\} = 0 \text{ , tüm k değerleri için} \tag{3.62}$$

$$\underset{i}{med}\{\hat{\alpha}_i\} = 0$$

$$\underset{j}{med}\{\hat{\beta}_j\} = 0$$

$$\underset{k}{med}\{\hat{\gamma}_k\} = 0$$

Toplamsal model için MedPol kestiricilerini elde etmenin bir başka yolu da veri setinin iki yönlü bir çizelge şeklinde ele alınıp, algoritmanın Bölüm 3.1.2.2.'de verildiği şekilde uygulanmasına dayanır. Bu yöntem sıklıkla yukarıda belirtilen hesaplamaların daha güç olduğu çok yönlü çizelgelerin analizinde tercih edilir. Yalnız bu yöntemin bir dezavantajı, (3.62) ile verilen koşullardan, artık değerlerinin ortancalarının sıfır olma koşulunun her zaman gerçeklenmemesidir.

MedPol kestiricilerinin, etkileşim terimli modeller için hesaplanışı toplamsal model için hesaplanışından biraz daha karmaşıktır. Çizelge 3.4'te verilen üç yönlü olumsallık çizelgelerine ait olası modellerden tüm iki yönlü etkileşim terimlerinin dahil olduğu,

$$\log(\hat{m}_{ijk}) = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk}$$

modelini ele alalım.

Toplamsal modellere Medyan Polish yöntemi uygulanırken, satır etkilerinin bulunması için tüm sütun ve katmanlar için satırların ortancası alınıyordu. Etkileşim terimlerinin dahil olduğu model için ise, önce tüm sütunlar üzerinden ilk satır ve katmanlar için ortanca değerleri elde edilir. Bu işlem sırasıyla tüm satır ve katmanlara, satır ve sütunlara ve en son olarak da sütun ve katmanlara uygulanır.

Algoritma matematiksel olarak aşağıda gösterilmiştir:

$$r_{ijk}^{(0)} = y_{ijk}$$

$$(ab)_{ij}^{(1)} = \underset{k}{med}\{r_{ijk}^{(0)}\}, \quad \text{tüm i ve j değerleri için}$$

$$(ac)_{ik}^{(1)} = \underset{j}{med}\{r_{ijk}^{(0)} - (ab)_i^{(1)}\}, \quad \text{tüm i ve k değerleri için}$$

$$a_i^{(1)} = \underset{jk}{med}\{(ab)_{ij}^{(1)}\}, \quad \text{tüm i değerleri için}$$

$$(bc)_{jk}^{(1)} = \underset{i}{med}\{r_{ijk}^{(0)} - (ab)_{ij}^{(1)} - (ac)_{ik}^{(1)}\}, \quad \text{tüm j ve k değerleri için}$$

$$c_k^{(1)} = \underset{i}{med}\{(ac)_{ik}^{(1)}\}, \quad \text{tüm k değerleri için}$$

$$b_j^{(1)} = \underset{i}{med}\{(ab)_{ij}^{(1)} - a_i^{(1)}\}, \quad \text{tüm j değerleri için}$$

$$m_a^{(1)} = med\{a_i^{(1)}\}.$$

Birinci döngünün sonunda elde edilen artık ve etki kestirim değerleri,

$$r_{ijk}^{(1)} = r_{ijk}^{(0)} - (ab)_{ij}^{(1)} - (ac)_{ik}^{(1)} - (bc)_{jk}^{(1)}$$

$$= r_{ijk}^{(0)} - \hat{\alpha}_i^{(1)} - \hat{\beta}_j^{(1)} - \hat{\gamma}_k^{(1)} - (\alpha\beta)_{ij}^{(1)} - (\alpha\gamma)_{ik}^{(1)} - (\beta\gamma)_{jk}^{(1)}$$

$$(\alpha\beta)_{ij}^{(1)} = (ab)_{ij}^{(1)} - a_i^{(1)} - b_j^{(1)}$$

$$(\alpha\gamma)_{ik}^{(1)} = (ac)_{ik}^{(1)} - c_k^{(1)}$$

$$(\beta\gamma)_{ik}^{(1)} = (bc)_{jk}^{(1)}$$

$$\hat{\alpha}_i^{(1)} = a_i^{(1)} - m_a^{(1)}$$

$$\hat{\beta}_j^{(1)} = b_j^{(1)}$$

$$\hat{\gamma}_k^{(1)} = c_{kj}^{(1)}$$

$$\hat{\mu}^{(1)} = m_a^{(1)}$$

olarak bulunur. Tüm iki faktör etkileşim terimlerini içeren modele sahip üç yönlü olumsallık çizelgesine medyan polish uygulanması sonucu elde edilen etki kestirim ve artık değerleri Çizelge 3.5.'te görüldüğü gibidir (Cook, 1985).

Çizelge 3.5. İki faktör etkileşim terimlerini içeren tam model için Medyan Polish kestiricileri.

		X				
Y	**Z**	X_1	X_2	$\cdots$	X_r	
Y_1	Z_1 $\vdots$ Z_t					
Y_1	Z_1 $\vdots$ Z_t		Artıklar (r_{ijk})			Y x Z-etkisi
$\vdots$	$\vdots$					
Y_1	Z_1 $\vdots$ Z_t					
	Z_1 $\vdots$ Z_t		X x Z-etkisi			Z-etkisi $(\hat{\alpha}_i)$
Y_1 $\vdots$ Y_s			X x Y-etkisi			Y-etkisi $(\hat{\gamma}_k)$
			X-etkisi $(\hat{\beta}_j)$			Toplam Değer $(\hat{\mu})$

Bir sonraki bölümde log-doğrusal modellerin, iki ve üç yönlü olumsallık çizelgelerine uygulanışı ele alınacaktır.

4. UYGULAMALAR

Bu bölümde log-doğrusal modellerin iki ve üç yönlü olumsallık çizelgelerine uygulanışı 3 farklı veri kümesi üzerinden ele alınacaktır. Log-doğrusal modellerde EÇO yöntemine alternatif olarak Medyan Polish algoritması da uygulanacak ve iki yöntemden elde edilen sonuçlar karşılaştırılacaktır. Analiz ve hesaplamalarda SAS, MATLAB ve Excel programları kullanılmıştır.

4.1. Sürücülerin Uyuma Alışkanlıkları ve Kaza Yapma Sayıları Üzerine Log-Doğrusal Model lerin Uygulanması

Sürücülerin uyku alışkanlıklarının ve son 3 yıl içinde yaptıkları kaza sayılarının kaydedilmesi sonucu 124 gözlemden oluşan bir veri kümesi elde edilmiştir. Kaza ve uyku değişkenleri,

Kaza: **0 –** Sürücü son üç yıl içinde hiç kaza yapmamış

1 – Sürücü son üç yıl içinde sedece bir kez kaza yapmış

2 – Sürücü son üç yıl içinde iki kez kaza yapmış

Uyku: **0** – Sürücü hiç uykusuz araba kullanmıyor

1 – Sürücü nadiren uykusuz araba kullanıyor

2 - Sürücü çoğunlukla uykusuz araba kullanıyor

olarak kodlanmıştır (Acar, 2010). Kodlanan bu değişkenlerin çapraz sınıflandırılması sonucu oluşan iki-yönlü olumsallık çizelgesi Çizelge 4.1'de görüldüğü şekildedir. BestFit programı kullanılarak hücre frekanslarının yaklaşık olarak Poisson dağılımına sahip olduğu gözlenmiştir.

Çizelge 4.1. Kaza x Uyku çizelgesi

	Uyku		
Kaza	**0**	**1**	**2**
0	32	25	10
1	18	12	5
2	8	4	10

Kaza ve uyku değişkenlerinin bağımsız olup olmadığını belirlemek için,

$$H_0 : (\alpha\beta)_{ij} = 0$$

hipotezi,

$$H_A : (\alpha\beta)_{ij} \neq 0$$

hipotezine karşılık test edilmelidir. Eğer H_0 hipotezi kabul edilirse kaza ve uyku değişkenleri bağımsızdır ve

$$\log \hat{m}_{ij} = \hat{\mu} + \alpha_i + \beta_j + \varepsilon_{ij} \tag{4.1}$$

şeklinde gösterilen toplamsal model; H_0 reddedilirse kaza ve uyku değişkenleri bağımlıdır ve etkileşim terimli,

$$\log \hat{m}_{ij} = \hat{\mu} + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ij} \tag{4.2}$$

modeli geçerli olacaktır.

Log-doğrusal modellerin bu veri kümesine uygulanması sonucu "Sapma" ve "Ki-kare" değerleri yaklaşık olarak $D = 11.80$ ve $\chi^2 = 13.85$ olarak bulunmuştur. Modele ait D değerinin modelin serbestlik derecesi olan 4 serbestlik dereceli ki-kare değeriyle ($\chi^2_{0.01,4}$=13.84) karşılaştırılması sonucu % 99 güven düzeyinde H_0 kabul edilir (p=0.018). Bu durumda, iki değişkenin birbirinden bağımsız olduğunu ve (4.1) ile verilen toplamsal modelin geçerli olduğunu söyleyebiliriz.

Modeldeki değişkenlerin anlamlı olup olmadığı Çizelge 4.2.'den incelenebilir. Buradan da görüldüğü üzere, her iki değişkene ait p değerleri 0.01 değerinden küçük olduğundan anlamlıdır.

Çizelge 4.2. Kaza uyku değişkenlerinin ana etkilerinin anlamlılığı

Değişken	**df**	**p**
Kaza	2	<.0001
Uyku	2	0.0002

En çok olabilirlik ve Medyan Polish yöntemleri aracılığıyla elde edilen beklenen hücre frekansları Çizelge 4.3. ve Çizelge 4.4.'de verilmiştir.

Çizelge 4.3. EÇO kestirim değerleri

	Uyku		
Kaza	**0**	**1**	**2**
0	31.3	22.1	13.5
1	16.3	11.5	7.0
2	10.2	7.3	4.4

Çizelge 4.4. MedPol kestirim değerleri

	Uyku		
Kaza	**0**	**1**	**2**
0	33	25	2
1	18	12	5
2	6	4	23

Çizelge 4.5. EÇO artık değerleri

	Uyku		
Kaza	**0**	**1**	**2**
0	0.7	2.9	***-3.5***
1	1.7	0.5	-2.0
2	-2.2	-3.3	***5.6***

Çizelge 4.6. MedPol artık değerleri

	Uyku		
Kaza	**0**	**1**	**2**
0	-1	0	***8***
1	0	0	0
2	2	0	***-13***

Yukarıda her iki yöntem için artık değerlerini veren 4.5. ve 4.6. numaralı çizelgeler incelendiğinde Medyan Polish yöntemi ile elde edilen hücre frekanslarının orijinal veriye daha iyi uyum sağladığı görülmüştür. Her iki çizelgede de (0,2) ve (2,2) kodlu hücrelerdeki artık değerleri diğer hücrelerin artık değerlerinden yüksek çıkmıştır. Bu hücrelerin aykırı değer olmasından şüphelenilebilir. Medyan Polish yöntemi ile elde edilen beklenen hücre frekansları genelde veriye daha iyi uyum sağlamıştır; (0,2) ve (2,2) numaralı hücrelere ait artık değerleri EÇO yöntemindekinden çok daha yüksek çıkmıştır. Bundan dolayı, Medyan Polish yönteminin EÇO yöntemine göre aykırı hücreleri daha iyi belirlediği düşünülebilir.

Artık değerleri çok yüksek çıkmış olan bu hücrelerin aykırı değer olup olmadığını belirlemek için Davies ve Gather (1993) tarafından önerilen "α - Aykırı Değer Bölgesi (α_i - outlier region)" yöntemi kullanılabilir. Normal dağılıma sahip veri kümeleri için aykırı değer belirlemek amaçlı ortaya çıkan bu yöntem Gather ve diğ. (2003) tarafından üssel dağılım ailesinin genel durumlarına genişletilmiştir. Kuhnt (2005) ise bu yöntemi genelleştirilmiş doğrusal modellerin bir alt sınıfı olan Poisson yanıt dağılımına sahip log-doğrusal modellere uygulamıştır.

Bu yöntem kısaca şu şekilde anlatılabilir:

P üssel dağılım ailesini ve $P_i \in P$, λ parametreli Poi(λ) dağılımını temsil etmek üzere tüm $\lambda \in R^+$ ve $\alpha_i \in (0,1)$ değerleri için α_i -aykırı değer bölgesi,

$$\text{out}(\alpha_i, \text{Poi}(\lambda))=\{x \in \text{supp}(\text{Poi}(\lambda)): f_i(x) < \delta(\alpha_i)\}$$

olarak tanımlanır. Burada,

$$\delta(\alpha_i)=\sup\{\delta>0: \text{Poi}(\{x: f_i(x) < \delta\}) \leq \alpha_i$$

dir. $x \in \text{out}(\alpha_i, \text{Poi}(\lambda))$ olan her değer Poisson dağılımına göre "α_i –outlier" olarak tanımlanır (Kuhnt ve Pawlitschko, 2005). Aykırı değer bölgesinin dışı, "iç değer bölgesi" olarak bilinir.

$$\text{inl}(\alpha_i, \text{Poi}(\lambda))= \text{supp}(\text{Poi}(\lambda))/ \text{out}(\alpha_i, \text{Poi}(\lambda))$$

olarak gösterilir. Bu yöntemde gözlenen hücre frekansları aykırı değer bölgesi içinde ise "aykırı", değil ise "iç" değer olarak tanımlanır (Kuhnt, 2004).

EÇO ve MedPol yöntemleri ile elde edilen kaza-uyku çizelgesine ait beklenen hücre frekanslarına bu yöntem uygulandığında 0.05-iç bölgeler Şekil 4.7 ve Şekil 4.8'deki gibi elde edilmiştir.

Çizelge 4.7. EÇO kestiricileri için 0.05-iç bölgeler

	Uyku		
Kaza	**0**	**1**	**2**
0	{19,...,41}	{11,..,30}	{6...,20}
1	{7,...,23}	{3,...,17}	{2,...,12}
2	{4,...,16}	{2,...,12}	{1,...,8}

Çizelge 4.8. MedPol kestiricileri için 0.05-iç bölgeler

	Uyku		
Kaza	**0**	**1**	**2**
0	{21,...,43}	{9,..,33}	{0,...,5}
1	{8,...,25}	{5,...,18}	{1,...,9}
2	{1,...,10}	{1,...,8}	{12,...,31}

Görüldüğü gibi son üç yılda iki kez kaza yapmış olan ve çoğunlukla uykusuz araba kullandığını söyleyen sürücüleri sayısını gösteren son hücre (2,2) hem EÇO hem MedPol yöntemi tarafından aykırı değer olarak elde edilmiştir. Çoğunlukla uykusuz araba kullandığını ve üç yıldır hiç kaza yapmadığını söyleyen sürücülerin sayısını ifade eden (0,2) numaralı hücre için ise durum daha farklıdır. Bu hücre yöntemi MedPol yöntemi ile aykırı değer olarak bulunurken EÇO yönteminde aykırı değer olarak bulunmamıştır.

Çizelge 4.9. EÇO kestiricileri için 0.01-iç bölgeler

	Uyku		
Kaza	**0**	**1**	**2**
0	{16,...,45}	{10,..,34}	{5...,23}
1	{4,...,26}	{3,...,20}	{1,...,14}
2	{2,...,18}	{1,...,14}	{0,...,10}

Çizelge 4.10. MedPol kestiricileri için 0.01-iç bölgeler

	Uyku		
Kaza	**0**	**1**	**2**
0	{17,...,47}	{11,..,37}	{0,...,6}
1	{8,...,29}	{4,...,21}	{0,...,11}
2	{0,...,12}	{0,...,9}	{11,...,35}

Güven düzeyi %95'ten %99'a çıkarıldığında bulunan 0.01-iç bölgeler Çizelge 4.9'da ve Çizelge 4.10'da verildiği gibidir. Gözlenen hücre frekansları ile EÇO kestiricileri ile elde edilen iç bölgeler kıyaslandığında hiçbir hücrenin aykırılığından şüphe duyulmadığı görülür. MedPol kestiricileri ile elde edilen 0.01-iç bölgeler incelenirse, %95 güven düzeyinde olduğu gibi, (0,2) ve (2,2) numaralı hücreler aykırı değer olarak tespit edilir.

MedPol sağlam bir aykırı değer bulma yöntemi olduğundan burada EÇO kestiriminin güvenilirliğinden şüphe duyulabilir. Bununla birlikte, güven düzeyi artışı MedPol yöntemi ile aykırı değer tespitinde bir değişiklik yaratmazken, EÇO yönteminde aykırı değer tespitini etkilemiştir.

4.2. Kanser Lezyonlarına Ait Özellikler Arasındaki İlişki Örüntülerinin Araştırılması Üzerine Log-Doğrusal Modellerin Uygulanması

Kanser hastalarında bulunan lezyon yapılarına ait değişik özellikler arasındaki ilişki örüntülerinin ortaya çıkarılması amacıyla Marmara Üniversitesi Hastanesi Radyoloji Anabilim Dalı'na memede saptanan lezyonlara yönelik biyopsi için başvuran tedavi 110 hastaya ait veriler toplanmıştır. Hastalarda bulunan lezyonların bazı özellikleri kaydedilmiştir.

Veri kümesinde lezyona ait değişkenler lezyonun boyutu, şekli, yerleşim yapısı, kenar özelliği, sınırları, posterior özellik, kalsifikasyon, vaskülarizasyon ve patolojik özellikleri kaydedilmiştir. Patolojik özelliğine göre lezyon kötü huylu ise üç sınıfa

ayrılmıştır. Lezyonun boyutu dışındaki tüm değişkenler kategorik yapıdadır ve aşağıdaki şekilde kodlanmıştır:

Lezyonun Şekli: **0 -** Oval
1 -Yuvarlak
2 - Düzensiz

Yerleşim: **0 -** Paralel
1 - Nonparalel

Kenar özelliği: **0** - Keskin
1 - Keskin Olmayan (Belirsiz, Köşeli, Mikrobüle, Spiküle)

Lezyon Sınırları: **0** - Keskin ara yüzey
1 - Ekojenik Halo

Posterior Özellik: **0** -Yok
1 - Enhansman
2 - Gölge
3 - Kombine

Kalsifikasyon: **0** - Yok
1 - Makrokalsifikasyon
2 - Mikrokalsifikasyon

Vaskülarizasyon: **0** - Vaskülarite yok
1 - Lezyon içerisinde vaskülarite var.

Lezyon Patolojik Özelliği: **0 -** Benign (İyi huylu) Lezyonlar
1 - Malign (Kötü huylu) Lezyonlar

Kötü huylu Lezyonların Sınıflandırılışı: **0** - Fibrostik
1 - Fibroadenon
2 – Papillon

Yukarıda belirtilen değişkenlerden olası iki yönlü ve üç yönlü olumsallık çizelgeleri oluşturulmaya çalışılmıştır. Ancak birçok çizelgede gözlenen hücre değerlerinin birçoğunun 5'in altında ya da sıfır olduğu gözlenmiştir. Elde edilecek sonuçların güvenilirliği açısından çapraz sınıflandırılacak değişkenlerin seçiminde oluşan tabloya ait hücre değerlerinin %80'inin 5'e eşit veya 5'ten büyük olmasına dikkat

edilmiştir. Bundan dolayı, değişkenlerin bir kısmının bazı kategorileri birleştirilerek analize sokulmuş ya da bazı kategorilerinde çok az gözlem bulunan değişkenler analize alınmamıştır.

Lezyon sınırları ve kalsifikasyon değişkenlerine ait "ekojenik halo" ve "makrokalsifikasyon" kategorilerinde çok az gözlem bulunmasından dolayı bu değişkenler çizelge oluşumunda kullanılmamıştır.

Orijinal veride Keskin, Belirsiz, Köşeli, Mikrobüle, Spiküle olarak kodlanan kenar özelliği değişkeni keskinlik baz alınarak "keskin" ve "keskin olmayan" olarak yeniden kodlanmıştır. Ancak keskin olmayan lezyon sayısı çok düşük olduğundan bu değişken de analizlerde kullanılmamıştır.

Kenar özelliği değişkenine benzer şekilde orijinal veride 4 kategori olarak verilen lezyonların patolojik özelliği de kötü huylu lezyonlar bir grupta toplanacak şekilde "benign (iyi huylu)" ve "malign (kötü huylu)" şeklinde iki kategori olarak kodlanmış ve bu şekilde analize alınmıştır.

4.2.1. Şekil - Vaskülarite İlişkisi

Lezyonun şekil (3 kategori) ve vaskülarizasyon özellikleri (2 kategori) arasındaki ilişkiyi incelemek için 110 olan gözlem sayısı eksik değer içeren bir gözlem çıkarılıp 109'a indirildi ve Çizelge 4.11 ile verilen 6 hücreli iki yönlü bir olumsallık çizelgesi oluşturuldu.

Çizelge 4.11. Şekil – Vaskülarite olumsallık çizelgesi

Vaskülarite			
Şekil	Yok	Var	**Toplam**
Oval	31	25	56
Yuvarlak	8	6	14
Düzensiz	15	24	39
Toplam	54	55	109

İki-yönlü bu çizelgeye logaritmik doğrusal modeller uygulanarak ana terim ve etkileşim terimlerinin anlamlılığı incelenmiştir. Vaskülarite ve Şekil değişkenleri arasında etkileşim olup olmadığını incelemek için,

$$H_0 : (\alpha\beta)_{ij} = 0$$

sıfır hipotezi,

$$H_A : (\alpha\beta)_{ij} \neq 0$$

hipotezine karşı test edilmelidir. Burada, i=1,2,3 ve j=1,2'dir. Dolayısıyla, modelin serbestlik derecesi (3-1)*(2-1) = 2 olacaktır.

H_0 hipotezinin kabulü durumunda geçerli olacak log-doğrusal model, toplamsal model; reddi durumunda ise etkileşim terimli model olacaktır. Analiz sonucunda toplamsal modele ait sapma, ki-kare ve log-olabilirlik uyum istatistiği değerleri sırasıyla D = 26.294, χ^2 = 25.470 ve $\log L$ = 225.201 olarak elde edilmiştir. Etkileşim teriminin dahil olduğu model doygun model olduğundan bu modele ait sapma ve ki-kare değerleri 0'dır. Log-olabilirlik değeri ise $\log L$ = 238.348'dir. İki modelin log-olabilirlik istatistiği değerleri arasındaki farkın 2 katı, sapma istatistiği değerine eşittir: $\Delta \log L = 2(238.348 - 225.201) = 26.294$. Bundan dolayı, toplamsal modele ait D değeri toplamsal ve doygun model arasında karar vermek için kullanılabilir. Bu değer, %95 güven düzeyinde 2 serbestlik dereceli ki-kare dağılımı değeri ile

karşılaştırıldığında çizelge değerinden ($\chi^2_{0.05,2}$ = 5.99) daha büyük kaldığından H_0 yani değişkenlerin bağımsızlığı hipotezi reddedilir (p<0.0001). Bu durumda geçerli olan model, ana etkiler ve etkileşim terimlerini içeren,

$$\log(\hat{m}_{ijk}) = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij}$$

doygun modeli olacaktır.

Çizelge 4.12'den de görüldüğü üzere %95 güven düzeyinde şekil değişkeninin model üzerinde anlamlı bir etkisi varken (p<0.0001), vaskülaritenin model üzerinde anlamlı bir etkisi yoktur (p=0.3112>0.05). Ancak şekil ve vaskülarite etkileşimi anlamlı (p=0.0001<0.05) olduğundan log-doğrusal modellerde hiyerarşi prensibi gereği vaskülarite ana etkisi de modele dahil edilmelidir.

Çizelge 4.12. Şekil - Vaskülarite ana etkilerinin anlamlılığı

Değişken	**sd**	**p**
Şekil	2	<.0001
Vaskülarite	1	0.3112
Şekil*Vaskülarite	2	<.0001

Tam modele ilişkin parametre tahminleri bir sonraki sayfada bulunan Çizelge 4.13'de verildiği gibidir.

Çizelge 4.13. Şekil*Vaskülarite doygun modeli için parametre tahminleri

Terim	Parametre değeri	Standart hata	p
Sabit	2.1972	0.3333	<.0001
Şekil			
Oval	1.5870	0.3658	<.0001
Düzensiz	0.1054	0.4595	0.8186
Yuvarlak	-	-	-
Vaskülarite			
Var	-0.5878	0.5578	0.2920
Yok	-	-	-
Şekil*Vaskülarite			
Oval*var	-0.6315	0.6409	0.3245
Oval*yok	-	-	-
Düzensiz*var	1.6525	0.6675	0.0133
Düzensiz*yok	-	-	-
Yuvarlak*yok	-	-	-
Yuvarlak*var	-	-	-

Log-doğrusal modellerde etkileşim terimlerine ait parametre toplamları sıfır olduğu bilindiğinden, değeri belli olmayan diğer etkileşim parametreleri de hesaplanabilir. Modele ilişkin tüm etkileşim parametre değerleri Çizelge 4.14'deki gibidir.

Çizelge 4.14. Şekil*Vaskülarite etkileşimine ait parametre kestirimleri

	Şekil		
Vaskülarite	**0**	**1**	**2**
0 (Yok)	***0.6315***	***1.021***	***-1.6525***
1 (Var)	-0.6315	**-1.021**	1.6525

Bu çizelgeden 1. değişkende 2 olarak kodlanan lezyonun düzensiz olma özelliği ile 2. değişkende 1 olarak kodlanan lezyonda vaskülarite bulunma durumu arasında aynı yönlü bir ilişki bulunmaktadır. Bu durumda düzensiz yapıya sahip lezyonlarda vaskülarite görülme durumunun daha yüksek olduğu sonucu çıkarılabilir. Etkileşim terimlerinin anlamlılık değerleri incelendiğinde bu parametreye ait anlamlılık değeri $p = 0.0133 < 0.05$ olduğundan, bu parametre değeri anlamlı olarak bulunmuştur.

4.2.2. Patoloji - Yerleşim -Vaskülarite İlişkisi

Lezyonların patolojik, yerleşim ve vaskülarizasyon özellikleri arasındaki ilişki örüntülerinin araştırılması için Çizelge 4.15'de verilen 108 gözlemden oluşan 2 x 2 x 2'lik bir üç yönlü olumsallık çizelgesi oluşturulmuştur.

Çizelge 4.15. Patoloji- Yerleşim - Vaskülarite Çizelgesi

		Vaskülarite	
Patoloji	**Yerleşim**	Yok	Var
0 (İyi)	Paralel	31	21
	Nonparalel	14	5
1 (Kötü)	Paralel	4	9
	Nonparalel	5	19

Oluşturulan bu çizelgeye log-doğrusal modeller uygulanarak doygun modele ait etkileşim terimlerinin sıfıra eşit olduğunu belirten,

$$H_0 : (\alpha\beta)_{ij} = (\beta\gamma)_{jk} = (\alpha\gamma)_{ik} = (\alpha\beta\gamma)_{ijk} = 0$$

hipotezi, bu etkileşim terimlerinden en az bir tanesinin sıfırdan farklı olduğunu belirten alternatif hipoteze karşı test edilecektir.

Analiz sonucunda minimal modele ilişkin sapma, ki-kare ve log-olabilirlik oran istatistiği değerleri sırasıyla $D = 31.806$ ve $\chi^2 = 34.371$ olarak bulunmuştur. D ve χ^2 değerleri çizelge değerinden ($\chi^2_{0.05,4}$=9.49) büyük olduğundan etkileşim terimlerinin sıfıra eşit olduğunu ileri süren H_0 hipotezi reddedilir (p<0.0001). Bu veri kümesi için toplamsal model geçerli değildir. Yani, sıfır hipotezinden belirtilen etkileşim terimlerinden en az bir tanesinin sıfırdan farklı olduğunu belirten H_A kabul edilir; doygun model veriye daha iyi uyum sağlamaktadır. Başka bir deyişle, patoloji, yerleşim ve vaskülarite değişkenlerinden en az ikisi birbirinden bağımsız değildir.

Çizelge 4.16'da tam modele ait parametre kestirim değerleri ve bu parametrelerin anlamlılıkları verilmiştir.

Çizelge 4.16. Vaskülarite -Yerleşim-Patoloji parametre anlamlılıkları

Parametre	**sd**	**p**
Vaskülarite	1	0.457
Yerleşim	1	0.2011
Vaskülarite*Yerleşim	1	0.9058
Patoloji	1	0.0076
Vaskülarite*Patoloji	1	0.0001
Yerleşim*Patoloji	1	0.001
Vaskülarite*Yerleşim*Patoloji	1	0.236

Çizelgede görüldüğü üzere vaskülarite ve yerleşim değişkenlerine ait parametreler anlamlı değil iken patoloji değişkenine ait parametre anlamlı bulunmuştur. İkili etkileşim terimleri arasında ise Vaskülarite*Patoloji ve Yerleşim*Patoloji etkileşimleri anlamlı iken Vaskülarite*Yerleşim etkileşimi anlamlı değildir. Üçlü etkileşim terimi Vaskülarite*Yerleşim*Patoloji de p= 0.2360 > 0.05 bulunduğundan anlamlı çıkmamıştır.

Bu değerler göz önünde bulundurulduğunda Çizelge 4.11.'e ait log-doğrusal model vaskülarite, patoloji ve yerleşim ana etkileri ile vaskülarite*patoloji, yerleşim*patoloji etkileşim terimlerini içerecek biçimde yeniden uygulanabilir. Yerleşim ve vaskülarite değişkenleri anlamlı çıkmamasına rağmen, bu değişkenlerin patoloji değişkeni ile etkileşimleri anlamlı olduğundan hiyerarşi prensibi gereği modele dahil edilmelidir.

İki etkileşim terimi içeren bu modelin doygun modele göre daha iyi bir model olup olmadığını belirlemek için yeni model ile tam model arasında fark olmadığını belirten

$$H_0 : \beta_{\text{yeni model}} = \beta_{\text{tam model}}$$

hipotezi,

$$H_A : \beta_{\text{yeni model}} \neq \beta_{\text{tam model}}$$

alternatif hipotezine karşı test edilmelidir.

Yeni modele ilişkin sapma istatistiği değeri $D = 1.6696$ olarak bulunmuştur. Bu değer, %95 güven düzeyinde modele ilişkin ki-kare tablo değerinden küçük bulunduğundan ($\chi_{0.05,1}$=3.84) H_0 hipotezi kabul edilir (p=0.196).

Daha az parametreli olan yeni model ile doygun model arasında fark olmadığından prsimoni ilkesi gereğince daha az parametreli olan model tercih edilir.

Bu durumda,

$$\log(\hat{m}_{ijk}) = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\beta\gamma)_{ik}$$

koşullu bağımsızlık modeli geçerlidir. Modeldeki parametreleri incelersek, patoloji değişkeni sabit iken vaskülarite ve yerleşim değişkenlerinin birbirinden bağımsız olduğu söylenebilir.

Çizelge 4.17. Koşullu bağımsızlık modeli için parametre tahminleri ve anlamlılıkları

Terim	**Parametre değeri**	**Standart hata**	**p**
Sabit	2.8993	0.2244	<.0001
Yerleşim			
Paralel	-0.6131	0.3444	0.0750
Nonparalel	-	-	-
Vaskülarite			
Var	-1.1350	0.3832	0.0031
Yok	-	-	-
Patoloji			
İyi	-0.9595	0.3569	0.0072
Kötü	-	-	-
Yerleşim*Patoloji			
Paralel*iyi	1.6199	0.4364	0.0002
Nonparalel*iyi	-	-	-
Paralel*kötü	-	-	-
Nonparalel*kötü	-	-	-

Vaskülarite*Patoloji			
Var*iyi	1.6835	0.4555	0.0002
Var*iyi	-	-	-
Yok*kötü	-	-	-
Yok*kötü	-	-	-

4.3. Kredi Başvurusu için Gerekli Bazı Özellikler Üzerine Log-Doğrusal Modellerin Uygulanması

Bankalar kredi başvurusu yapan kişilerden, kişilerin iş ve maddiyat durumları ile ilgili bazı bilgiler istemektedir. Bu bilgilerin birçoğu kategorik yapıdadır.

Prof. Dr. Hans Hofman tarafından düzenlenen ve Alman Kredi Verisi olarak bilinen bu veri seti http://archive.ics.uci.edu/ml/datasets/Statlog+(German+Credit+Data) adresinden alınmıştır. Veri madenciliğinde kullanılan bu veri setinde bulunan değişkenler arası ilişki örüntülerinin araştırılması için log-doğrusal modeller uygulanmıştır. Veri 1000 gözlemden oluşmaktadır. Veri setindeki değişkenler, kişilerin demografik özellikleri, ne iş yaptıkları, kaç yıldır çalıştıkları, şu ana kadar yaptıkları maddi birikim miktarları, bankadan talep edilen kredi miktarı, oturdukları konutun kendilerine ait olup olmadığı gibi çoğunlukla nitel değişkenlerden oluşmaktadır. Bu veri setinde ilişki örüntülerinin yanı sıra 3.bölümde bahsedilen üç yönlü olumsallık çizelgelerindeki kestirim yöntemleri aracılığıyla beklenen hücre frekanslarının elde edilmesi ve yöntemler arası ne gibi farklılıklar olduğunun incelenmesi amaçlanmıştır.

4.3.1. Birikim – Çalışma süresi – Konut değişkenleri için log-doğrusal modeller

Veride birikim (B), çalışma süresi (C) ve konut (K) değişkenleri aşağıdaki şekilde kodludur:

1- *Birikim:* **1** – 100 Mark'dan az

2 – 100 - 500 Mark arası

3 – 500 - 1000 Mark arası

4 – 1000 Mark'dan fazla

5 – Bilinmiyor

Çalışma süresi (yıl): **1 –** 1 yıldan az

2 – 1 – 4 yıl arası

3 – 4 – 7 yıl arası

4 – 7 yıldan fazla

Konut: **1 –** Kiralık

2 – Kendine ait

3 – Ücret ödenmiyor

Bu değişkenler çapraz sınıflandırıldığında bir çok hücrede boş değer bulunmuş ve oluşan tabloda %20'den fazla hücrenin frekansı 5'ten düşük çıkmıştır. Bundan dolayı birikim ve konut değişkenlerinin bazı kategorileri çıkarılıp, bazıları birleştirilerek aşağıdaki şekilde yeniden kodlanmış ve 3 x 4 x 2 boyutlu bir olumsallık çizelgesi oluşturulmuştur. Bu şekilde gözlem sayısı 671'e düşmüştür.

Birikim: **1** – 100 Mark'dan az

2 – 100 - 500 Mark arası

3 – 500 Mark'dan fazla

Konut: **1 –** Kiralık

2 – Kendine ait

Çizelge 4.18. Birikim – Çalışma Süresi – Konut üç yönlü çizelgesi

		C1	**C2**	**C3**	**C4**
B1	**K1**	31	35	17	16
	K2	82	154	71	84
B2	**K1**	4	10	5	2
	K2	13	18	18	13
B3	**K1**	4	7	3	5
	K2	7	34	15	23

Bu tabloya log-doğrusal modeller uygulandığında Çizelge 4.19'da görüldüğü üzere tüm değişkenlerin ana etkileri anlamlı çıkmıştır.

Çizelge 4.19. C – B – K ana etkilerinin anlamlılığı

Değişken	sd	p
C	3	<.0001
B	2	<.0001
K	1	<.0001

Modelin sapma istatistiği $D = 24.2395$ olarak bulunmuştur. Modelin serbestlik derecesi 17'dir ve 17 serbestlik dereceli ki-kare dağılımına ait istatistik değeri 27.59, modele ait sapma istatistiği değerinden daha büyük olduğundan

$$H_0 : (\alpha\beta)_{ij} = (\beta\gamma)_{jk} = (\alpha\gamma)_{ik} = (\alpha\beta\gamma)_{ijk} = 0$$

hipotezi kabul edilir (p=0.113). Yani, veri kümemize uygun model hiçbir etkileşim terimi içermeyen toplamsal (bağımsız) model olacaktır. Bu modele ilişkin parametre tahminleri Çizelge 4.20'de görüldüğü gibidir.

Çizelge 4.20. C – B – K değişkenlerine ait parametre tahminleri ve model iyiliği

Terim	Parametre	Standart hata	p
Sabit	2.069	0.1269	<.0001
Çalışma süresi			
C1	-0.0141	0.1187	0.9055
C2	0.5901	0.1043	<.0001
C3	-0.103	0.1214	0.3962
C4	-	-	-
Birikim			
B1	1.6094	0.1107	<.0001
B2	-0.1661	0.1492	0.2654
B3	-	-	-
Konut			
K1	-1.3422	0.0953	<.0001
K2	-	-	-

Modele ait beklenen hücre frekansları, EÇO yöntemiyle Çizelge 4.21'deki gibi bulunmuştur.

Çizelge 4.21. EÇO yöntemi ile elde edilen beklenen hücre frekansları

		C1	C2	C3	C4
B1	**K1**	21.3	39	19.5	21.6
	K2	81.6	149.3	74.6	82.8
B2	**K1**	3.6	6.6	3.3	3.6
	K2	13.8	25.3	12.6	14
B3	**K1**	4.2	7.8	3.9	4.3
	K2	16.3	29.8	14.9	16.5

Aynı veri kümesine toplamsal model için Medyan Polish yöntemi uygulanırsa 1. döngünün 1. aşamasında çalışma süresi (C) değişkenine ait sütunların ortancası hesaplanıp, ilgili hücre değerlerinden çıkarılmalıdır. Bu durumda elde edilen yeni hücre değerleri ve C-etki değerleri Çizelge 4.22'deki gibidir.

Çizelge 4.22. Medyan Polish Yöntemi 1. döngü 1. aşama

		C1	C2	C3	C4
B1	**K1**	31	35	17	16
	K2	82	154	71	84
B2	**K1**	4	10	5	2
	K2	13	18	18	13
B3	**K1**	4	7	3	5
	K2	7	34	15	23
C-etkileri		**10**	**26**	**16**	**14.5**
		C1	**C2**	**C3**	**C4**
B1	**K1**	21	9	1	1.5
	K2	72	128	55	69.5
B2	**K1**	-6	-16	-11	-12.5
	K2	3	-8	2	-1.5
B3	**K1**	-6	-19	-13	-9.5
	K2	-3	8	-1	8.5

2. aşamada Birikim (B) ve Çalışma Süresi (C) değişkenlerinin yeri değiştirilerek yeni bir çizelge elde edilir ve B değişkeninin kategorilerine göre sütun ortancaları hesaplanır. 2.aşamada elde edilen hücre değerleri Çizelge 4.23'teki gibidir.

Çizelge 4.23. Medyan Polish yöntemi 1. döngü 2. aşama

		B1	**B2**	**B3**	***C-etkileri***
C1	**K1**	21	-6	-6	***10***
	K2	72	3	-3	***10***
C2	**K1**	9	-16	-19	***26***
	K2	128	-8	8	***26***
C3	**K1**	1	-11	-13	***16***
	K2	55	2	-1	***16***
C4	**K1**	1.5	-12.5	-9.5	***14.5***
	K2	69.5	-1.5	8.5	***14.5***
B-etkileri		**38**	**-7**	**-4.5**	
		B1	**B2**	**B3**	***C-etkileri***
C1	**K1**	-17	1	-1.5	***10***
	K2	34	10	1.5	***10***
C2	**K1**	-29	-9	-14.5	***26***
	K2	90	-1	12.5	***26***
C3	**K1**	-37	-4	-8.5	***16***
	K2	17	9	3.5	***16***
C4	**K1**	-36.5	-5.5	-5	***14.5***
	K2	31.5	5.5	13	***14.5***

3. aşamada elde edilen B-etkileri hücre değerlerinden çıkarılarak yeni hücre değerleri elde edilir ve Konut (K) değişkeni sütun değişkeni olacak şekilde tablo yeniden düzenlenir. Bu aşamada Çizelge 4.24 elde edilir.

Çizelge 4.24. Medyan Polish Yöntemi 1. döngü 3.aşama

		K1	**K2**	***C-etkileri***	***B-etkileri***
C1	**B1**	-17	34	***10***	***38***
	B2	1	10	***10***	***-7***
	B3	-1.5	1.5	***10***	***-4.5***
C2	**B1**	-29	90	***26***	***38***
	B2	-9	-1	***26***	***-7***
	B3	-14.5	12.5	***26***	***-4.5***
C3	**B1**	-37	17	***16***	***38***
	B2	-4	9	***16***	***-7***
	B3	-8.5	3.5	***16***	***-4.5***
C4	**B1**	-36.5	31.5	***14.5***	***38***
	B2	-5.5	5.5	***14.5***	***-7***
	B3	-5	13	***14.5***	***-4.5***
K-etkileri		**-8.75**	**11.25**		
		K1	**K2**	***C-etkileri***	***B-etkileri***
C1	**B1**	-8.25	22.75	***10***	***-17.25***
	B2	9.75	-1.25	***10***	***9***
	B3	7.25	-9.75	***10***	***13.25***
C2	**B1**	-20.25	78.75	***26***	***-17.25***
	B2	-0.25	-12.25	***26***	***9***
	B3	-5.75	1.25	***26***	***13.25***
C3	**B1**	-28.25	5.75	***16***	***-17.25***
	B2	4.75	-2.25	***16***	***9***
	B3	0.25	-7.75	***16***	***13.25***
C4	**B1**	-27.75	20.25	***14.5***	***-17.25***
	B2	3.25	-5.75	***14.5***	***9***
	B3	3.75	1.75	***14.5***	***13.25***

Bu aşamada her bir değişkene göre ortanca alma işlemi sona erdiğinden 1. döngü bitmiş olur.

Üçüncü bölümde sırasıyla (3.44), (3.46), (3.47) denklemleri ile verilen değişkenlerin etki değerleri kümesi birinci döngü sonunda,

$$a_1^{(1)} = 10$$
$$a_2^{(1)} = 26$$
$$a_3^{(1)} = 16$$
$$a_4^{(1)} = 14.5$$
$$b_1^{(1)} = 38$$
$$b_2^{(1)} = -7$$
$$b_3^{(1)} = -4.5$$
$$c_1^{(1)} = -8.75$$
$$c_2^{(1)} = 11.25$$

bulunur. Bu etki kümelerinin ortanca değerleri (3.49), (3.50) ve (3.51) denklemlerinden sırasıyla,

$$m_a^{(1)} = 15.25$$

$$m_b^{(1)} = -4.5$$

$$m_c^{(1)} = 1.25$$

olarak elde edilir. Buradan, merkezileştirilmiş etki kestirim değerleri,

$$\hat{\alpha}_1^{(1)} = -5.25$$

$$\hat{\alpha}_2^{(1)} = 10.75$$

$$\hat{\alpha}_3^{(1)} = 0.75$$

$$\hat{\alpha}_4^{(1)} = -0.75$$

$$\hat{\beta}_1^{(1)} = -12.75$$

$$\hat{\beta}_2^{(1)} = 13.5$$

$$\hat{\beta}_3^{(1)} = 17.75$$

$$\hat{\gamma}_1^{(1)} = -10$$

$$\hat{\gamma}_2^{(1)} = 10$$

$$\hat{\mu}^{(1)} = 12$$

bulunur.

Bu değerler bulunduktan sonra ikinci aşamaya geçilir. Bu döngü, sıfıra yakınsanıncaya dek devam eder. Bu veri setinde 3. döngüde sıfıra yakınsanmıştır. 3. döngü sonunda elde edilen hücre değerleri Çizelge 4.25'deki gibidir.

Çizelge 4.25. MedPol yöntemi sonucu elde edilen hücre artık değerleri

B	**K**	**C**			
		C1	**C2**	**C3**	**C4**
B1	**K1**	-10.2	-15.34	-26.21	-29.21
	K2	23.2	86.09	10.21	19.46
B2	**K1**	4.34	1.21	3.34	1.34
	K2	-4.21	-8.34	-1.21	-6.96
B3	**K1**	1.78	-4.34	-1.21	2.53
	K2	-12.78	5.09	-6.78	1.21

Bu değerler hücre artık değerlerini temsil eder. Bu değerlerin orijinal çizelgedeki hücre değerlerine eklenmesi ile hücre kestirim değerleri bulunur. Bulunan hücre kestirim değerleri Çizelge 4.26'da verilmiştir.

Çizelge 4.26. MedPol yöntemi ile elde edilen beklenen hücre frekansları.

		C1	**C2**	**C3**	**C4**
B1	**K1**	20.7	19.6	-9.21	-13.2
	K2	105.2	240	81.2	103.5
B2	**K1**	8.34	11.2	8.3	3.3
	K2	8.78	9.6	16.7	6
B3	**K1**	5.78	2.6	1.7	7.5
	K2	-5.78	39	8.2	24.2

MedPol yönteminde her bir döngüde elde edilen etki değerleri Çizelge 4.27'de verilmiştir.

Çizelge 4.27. MedPol ana-kestirim analizinde etki değerlerindeki değişim.

	Döngü			
Etki	**1**	**2**	**3**	**Toplam**
C1	-5.25	2.25	0	-3
C2	10.75	-3.75	-0.875	6.125
C3	0.75	-1.75	0	-1
C4	-0.75	1.75	2.375	3.375
B1	-12.75	-3	0	-15.75
B2	13.5	0	0.4375	13.9375
B3	17.75	0.75	-0.25	18.25
K1	-10	0.875	0.34375	-8.78125
K2	10	-0.875	-0.34375	8.78125

EÇO ve MedPol yöntemlerinden elde edilen beklenen hücre frekansları ve gözlenen hücre frekansları arasındaki farklar karşılaştırılarak modelin artık analizi yapılabilir.

Çizelge 4.28. EÇO ve MedPol yöntemlerinden elde edilen artık değerleri

EÇO yöntemi ile elde edilen artık değerleri					
		C1	**C2**	**C3**	**C4**
B1	**K1**	9.67	-4.02	-2.51	-5.6
	K2	0.36	4.62	-3.68	1.2
B2	**K1**	0.38	3.38	1.69	-1.66
	K2	-0.82	-7.30	5.34	-1.02
B3	**K1**	-0.26	-0.80	-0.90	0.67
	K2	-9.32	4.12	0.06	6.44
MedPol yöntemiyle elde edilen artık değerleri					
		C1	**C2**	**C3**	**C4**
B1	**K1**	-10.2	-15.34	-26.21	-29.21
	K2	23.2	86.09	10.21	19.46
B2	**K1**	4.34	1.21	3.34	1.34
	K2	-4.21	-8.34	-1.21	-6.96
B3	**K1**	1.78	-4.34	-1.21	2.53
	K2	-12.78	5.09	-6.78	1.21

Bu iki çizelgeye ait değerlere bakıldığında her iki yöntem de de farklı hücrelerin artık değerlerinin yüksek olduğu ve MedPol yönteminden elde edilen artık değerlerinin EÇO yöntemine kıyasla çok daha yüksek olduğu görülür. Medyan Polish sağlam bir yöntem olduğundan bu normal bir durumdur.

0.05-aykırı değer bölgesi yönteminin bu hücrelere uygulanması ile 1, 8, 9, 10, 13, 15, 17,19, 20, 22, 23 ve 24 numaralı hücreler her iki yöntem ile de aykırı değer olarak saptanmıştır.

5, 16 ve 18 numaralı hücreler sadece EÇO yöntemiyle, 2 ve 11 numaralı hücreler ise sadece MedPol yöntemiyle aykırı değer olarak bulunmuştur.

Buradan, bir hücrenin aykırı değer olup olmadığının tespitinde aykırı değer bölgesi içinde olup olmadığının artık değerinin yüksekliğinden daha önemli olduğu söylenebilir.

Güven düzeyinin %99'a çıkarılması ile EÇO yöntemi aracılığıyla 11, MedPol yöntemi aracılığıyla 12 adet aykırı hücre bulunmuştur. Bu hücrelerden 8 hücre her iki yöntem ile de aykırı değer olarak bulunmuştur. 14, 16, 18 numaralı hücreler sadece EÇO, 2, 11, 15, 17 numaralı hücreler ise sadece MedPol yöntemi ile aykırı olarak tespit edilmiştir.

Üç yönlü bu çizelgede güven düzeyinin yükselmesiyle EÇO yöntemi ile aykırı değer olarak bulunan hücre sayısında azalma olmuştur. MedPol yönteminde ise aykırı değer olarak bulunan hücre sayısında bir değişim olmamıştır. 0.01-iç bölgede, 0.05-iç bölgelerin dışında kalan 2 ve 11 numaralı hücrelerin yanı sıra 15 ve 17 numaralı hücreler de aykırı değer olarak bulunmuştur.

Bu durumda, iki yönlü olumsallık çizelgelerinde olduğu gibi, yüksek güven düzeyinde yapılan çalışmalarda MedPol yönteminin EÇO yöntemine kıyasla daha güvenilir sonuç verdiği söylenebilir.

Hücre kestirim değerleri göz önüne alındığında 3-yönlü olumsallık çizelgeleri için EÇO yönteminin MedPol yönteminden daha iyi sonuç verdiği gözlenmiştir. Bunun sebebi, son döngüde tüm değişkenlerin etkilerinin yaklaşık olarak sıfır bulunması gerekirken C değişkenin 4. kategorisinin etkisinin 2.5 çıkması olabilir. Ancak yine de diğer tüm kategorilerin etkilerinde bundan sonraki aşamalarda bir değişme olmadığından bu döngüde işleme son vermek gerekmiştir.

5. SONUÇ VE ÖNERİLER

Bu çalışmada, log-doğrusal modeller, olumsallık çizelgesi şeklinde belirtilebilen trafik, sağlık ve bankacılık sektörlerine ait üç farklı veri kümesine uygulanmıştır. Trafikte sürücülerin uyku alışkanlıkları ve son 3 yılda yaptıkları kaza sayılarını içeren veri kümesine log-doğrusal modellerin uygulanması sonucu sürücülerin yaptıkları kaza sayılarının uyku alışkanlıklarından bağımsız olduğu görülmüştür. Hücre frekansları, sağlam olan "MedPol" ve sağlam olmayan "EÇO" yöntemleri ile bulunmuş ve aralarındaki farklılıklar incelenmiştir. 2-yönlü bir olumsallık çizelgesi olarak verilen bu veriye MedPol kestirim değerleri daha iyi uyum sağlamıştır. Çizelgede aykırı değer içeren bir hücre olup olmadığının tespiti için, "aykırı değer bölgesi" olarak adlandırılan yöntem sırasıyla EÇO ve MedPol hücre kestirim değerlerine uygulanmış ve %95 güven düzeyine ait iç bölge değerleri bulunmuştur. Bu yöntem sonucunda, EÇO kestiriminde tek bir aykırı değer, MedPol kestiriminde ise iki aykırı değer saptanmıştır. Çizelgenin son hücresi iki kestirim yönteminde de aykırı değer iken, 3.hücre sadece MedPol kestirim yönteminde aykırı değer olarak bulunmuştur.

%99 güven düzeyinde ise, EÇO yöntemi ile hiç bir hücre aykırı değer olarak bulunmazken, MedPol yöntemi ile %95 güven düzeyinde aykırı değer olarak tespit edilen 3 ve 9 numaralı hücreler yine aykırı değer olarak tespit edilmiştir.

Bu durumda, iki yönlü olumsallık çizelgelerinde güven düzeyi arttıkça EÇO yönteminin yanıtlarından kuşku duyulabileceği anlaşılmıştır. Dolayısıyla, iki yönlü olumsallık çizelgelerinde ve yüksek güven düzeyinde yapılan çalışmalarda MedPol yöntemi, EÇO yöntemine tercih edilmelidir.

Çoğunlukla uykusuz araba kullanan ve son üç yıl içinde iki kez kaza yaptığını söyleyen sürücülerin sayısını gösteren son hücrenin ve son üç yıldır hiç kaza yapmadığını söyleyen sürücü sayılarının aykırı değer olarak çıkması sürücülerin kaza sayıları konusunda verdikleri bilgilerin hatalı olması şüphesini uyandırmaktadır.

Kanser lezyonu taşıyan hastaların lezyon özelliklerini belirten ikinci veri kümesine log-doğrusal modeller uygulanarak önce şekil ve vaskülarite, sonra ise patoloji-yerleşim-vaskülarite değişkenleri arasındaki ilişki örüntüleri araştırılmıştır.

Lezyonların şekil ve vaskülarite özellikleri arasındaki etkileşimin anlamlı olduğu sonucuna ulaşılmıştır. Düzensiz şekle sahip lezyonlarda vaskülarite görülme durumunun daha yüksek olduğu gözlenmiştir.

Lezyonların patolojik özelliği, yerleşim ve vaskülarite durumlarının birbirinden bağımsız olmadığı, patolojik özellik ile yerleşimve vaskülaritenin ayrı ayrı etkileşim içinde olduğu sonucuna ulaşılmıştır.

Bu veri setinde MedPol yöntemi uygulanmamıştır. Çünkü, etkileşim terimi içeren modellerde MedPol uygulaması tüm etkileşim terimlerine ait etki değerlerini içermektedir ancak burada yerleşim*vaskülarite etkileşimi anlamsız bulunmuştur.

Son uygulamanın yapıldığı kredi veri setinde kişilerin birikim miktarları, konut sahibi olma durumları ve çalışma süreleri (yıl) arasında bir etkileşim bulunmamıştır.

Bu veri setine EÇO ve MedPol kestirimleri üzerinden "α - aykırı değer bölgesi" yöntemi uygulanarak aykırı hücrelerin tespit edilmesi amaçlanmıştır. 0.05-aykırı değer bölgesi yönteminin uygulanması ile birçok hücre aykırı değer olarak bulunmuştur. Aykırı değer bulunmayan hücrelerin sayısı EÇO kestirim yönteminde dört, MedPol yönteminde üçtür.

0.01-aykırı değer bölgesi yönteminin uygulanmasıyla aykırı hücre sayısı EÇO yönteminde 5, MedPol yönteminde ise 1 adet azalmıştır. Bu durumda, güven düzeyi arttırıldığında MedPol yönteminin sonuçlarının EÇO yöntemine göre daha güvenilir olacağı düşünülebilir.

Üç yönlü olumsallık çizelgelerinde MedPol yöntemi, iki yönlü olumsallık çizelgelerinde olduğu gibi belirgin bir şekilde iyi sonuç vermemektedir. Ancak, hücre kestirim değerleri göz önüne alındığında 3-yönlü olumsallık çizelgeleri için EÇO yönteminin MedPol yönteminden daha iyi sonuç verdiği gözlenmiştir. İki yönlü olumsallık çizelgelerinde ise aykırı değer araştırılırken MedPol kestiricileri EÇO kestiricilerinden daha iyi sonuç verdiğinden, MedPol kestiricilerinin tercih edilmesi önerilir.

KAYNAKLAR

Acar N., 2010. The Identification of Outliers in Generalized Linear Models, *International Conference on Trends and Perspectives in Linear Statistical Inference, LINSTAT 2010, Tomar, Portugal,* July 27-31.

Agresti A., 2002. Categorical Data Analysis. John Wiley & Sons Inc., New Jersey.

Agresti A., 2007. An Introduction to Categorical Data Analysis. John Wiley & Sons Inc., New Jersey.

Baker F.B., 1981. Log-Linear, Logit-Linear Models: A Didactic. *Journal of Educational Statistics*, 6(1), 75-102.

Bartlett M.S., 1935. Contingency Table Interactions. *Supplement to the Journal of Statistical Society*, 2 (2), 248-252.

Birch M.W., 1963. Maximum Likelihood in Three-Way Contingency Tables. *Journal of the Royal Statistical Society. Series B(Methodological)*, 25(1), 220-233.

Bishop Y.M.M., 1969. Full Contingency Tables, Logits and Split Contingency Tables. *Biometrics*, 25(2), 383-399.

Bishop Y.M.M. ve Fienberg S.E., 1969. Incomplete Two-Dimensional Contingency Tables. *Biometrics*, 25(1), 119-128.

Burnett J.D., 1983. Loglinear Analysis: A New Tool for Educational Researchers. *Canadian Journal of Education*, 8(2), 139-154.

Christensen R.,1997. Log-linear Models and Logistic Regression. Springer, New York.

Cook N.R., 1985. Three-Way Analyses in *Exploring Data Tables, Trends and Shapes,* pp.189-224, Eds. Hoaglin D.C., Mosteller F. ve Tukey J.W., Wiley, New York.

Davies L. ve Gather U., 1993. The Identification of Multiple Outliers. *Journal of the American Statistical Association*, 88, 782-792.

Deming W.E. ve Stephan F.F., 1940. On a Least Square Adjustment of a Sampled Frequency Table When The Expected Marginal Totals Are Known. *Annals of Mathematical Statistics*,11, 427-444.

Dobson J.A., 1990. An Introduction to Generalized Linear Models. Chapman and Hall, UK.

Dobson J.A., 2002. An Introduction to Generalized Linear Models. Chapman and Hall, USA.

Fahrmeir L. ve Kaufmann H., 1985. Consistency and Asymptotic Normality of the Maximum Likelihood Estimator in Generalized Linear Models. *The Annals of Statistic,*13, 342-368.

Fienberg S.E., 1968. The Geometry of an r x c Contingency Table. *The Annals of Mathematical Statistics*, 39(4), 1186-1190.

Fienberg S.E., 1970a. The Geometry of a Two by Two Contingency Table. *Journal of the American Statistical Association*, 65(330), 694-701.

Fienberg S.E., 1970b. The Analysis of Multidimensional Contingency Tables. *Ecology,* 51(3), 419-433.

Fienberg S.E. ve Rinaldo A., 2007. Three Centuries of Categorical Data Analysis: Log-linear Models and Maximum Likelihood Estimation. *Journal of Statistical Planning and Inference,* 137(11), 3430-3445.

Good I.J., 1963. Maximum Entropy for Hypothesis Formulation, Especially for Multidimensional Contingency Tables. *The Annals of Mathematical Statistics,* 34(33), 911-934.

Goodman L.A., 1965. On the Statistical Analysis of Mobility Tables. *The American Journal of Sociology,* 70(5), 564-585.

Goodman L.A., 1970. The Multivariate Analysis of Qualitative Data: Interactions Among Multiple Classifications. *Journal of the American Statistical Association*, 65(329), 226-256.

Haberman S.J., 1973. Log-linear Models for Frequency Data: Sufficient Statistics and Likelihood Equations. *Annals of Statistics*, 1(4), 617-632.

Hoaglin D.C., Mosteller F. ve Tukey F.W., 1983. Understanding Robust and Exploratory Data Analysis, Wiley, USA.

Jeansonne A., 2007. Log-linear Models. http://www.education.umd.edu/EDMS/fac/Hancock/Course_Materials/EDMS771/readings/LogLinearModels%20reading.pdf

Kuhnt S., 2004. Outlier Identification Procedures for Contingency Tables using Maximum Likelihood and L1 Estimates. *Scandinavian Journal of Statistics*, 31, 431-442.

Kuhnt S. ve Pawlitschko J., 2005. Outlier Identification Rules for Generalized Linear Models in *Innovations in Classification, Data Science and Information Systems,* pp.165-172, Eds. Baier D. & Wernecke D., Springer-Verlag, Heidelberg.

Kuhnt S., 2010. Breakdown Concepts for Contingency Tables. *Metrika,*71, 281-294.

Lawal H.B., 2003. Categorical Data Analysis with SAS and SPSS Applications. Lawrence Erlbaum Associates. New Jersey.

Lindsey J.K., 1997. Applying Generalized Linear Models. Springer, New York.

McCullagh P. ve Nelder J.A., 1988. Generalized Linear Models. Chapman and Hall. USA.

McCulloch C.E. ve Searle S.R., 2001. Generalized, Linear and Mixed Models. Wiley. Canada.

Mosteller F. and Parunak A., 1985. Identifying Extreme Cells in a Sizeable Contingency Table: Probabilistic and Exploratory Approaches in *Exploring Data Tables, Trends and Shapes,* pp.189-224, Eds. Hoaglin D.C., Mosteller F. ve Tukey J.W., Wiley, New York.

Myers H.R. ve diğ., 2002. Generalized Linear Models with Applications in Engineering and the Sciences. Wiley. New York.

Nelder J.A. ve Wedderburn R.W.M., 1972. Generalized Linear Models. *Journal of Statist. Soc. A.,*135, 370.

Nelder J.A., 2000. The Analysis of Contingency Tables with One Factor as the Response: Round Two. *Journal of the Royal Statistical Society,* 49(3), 383-388.

Norton H.W., 1945. Calculation of Chi-Square for Complex Contingency Tables. *Journal of Statist. Soc. A*., 40 (230), 251-258.

Powers D.A. ve Xie Y., 1999. Statistical Methods for Categorical Data Analysis. Academic Press, Inc.

Rao P. R. ve Toutenburg H., 1999**.** Linear Models: Least Squares and Alternatives. Springer-Verlag. New York.

EK 1.

Y, θ parametresine bağlı f(y; θ) olasılık yoğunluk fonksiyonuna sahip sürekli yanıt değişkeni olsun. Kesim 2.3'te belirtildiği gibi log-olabilirlik fonksiyonu,

$l(\theta; y) = \log f(y;\theta)$

dir. Log-olabilirlik fonksiyonunun türevine eşit olan skor fonksiyonu U,

$$U = \frac{dl}{d\theta} = \frac{d \log f(y;\theta)}{d\theta} = \frac{1}{f(y;\theta)} \frac{df(y;\theta)}{d\theta} \tag{E.1}$$

olarak yazılabilir. (E1) ifadesinin beklenen değerini alırsak,

$$E(U) = \int \frac{d \log f(y;\theta)}{d\theta} f(y;\theta) dy = \int \frac{df(y;\theta)}{d\theta} dy \tag{E.2}$$

elde edilir. Düzenlilik koşulları altında integral ve türev yer değiştirebildiğinden ve f(y;θ) bir olasılık yoğunluk fonksiyonu için $\int f(y;\theta) dy = 1$ olduğundan (E.2) ifadesinin sağ yanı,

$$\int \frac{df(y;\theta)}{d\theta} dy = \frac{d}{d\theta} \int f(y;\theta) dy = \frac{d}{d\theta} 1 = 0$$

olur. Yani,

$$E(U) = 0 \tag{E.3}$$

olarak bulunur.

Ayrıca (E.1) ifadesinin θ'ya göre türevlenip beklenen değeri alındığında aynı koşullar söz konusu olacağından,

$$\frac{d}{d\theta} \int \frac{d \log f(y;\theta)}{d\theta} f(y;\theta) dy = \frac{d^2}{d\theta^2} \int f(y;\theta) dy$$

olur. $\int f(y;\theta) dy = 1$ olduğundan denklemin sağ tarafı sıfıra eşittir ve sol taraf eşitliği ise,

$$\int \frac{d^2 \log f(y;\theta)}{d\theta^2} f(y;\theta)dy + \int \frac{d \log f(y;\theta)}{d\theta} \frac{df(y;\theta)}{d\theta} dy$$

şeklinde ifade edilebilir. Bu ifadedeki ikinci terime (E1) uygulanırsa,

$$\int \frac{d^2 \log f(y;\theta)}{d\theta^2} f(y;\theta)dy + \int \left[\frac{d \log f(y;\theta)}{d\theta} \right]^2 f(y;\theta)dy = 0$$

elde edilir. Bundan dolayı,

$$E\left[-\frac{d^2 \log f(y;\theta)}{d\theta^2} \right] = E\left\{ \left[\frac{d \log f(y;\theta)}{d\theta} \right]^2 \right\}$$

başka bir ifadeyle

$$E(-U') = E(U)^2 \tag{E.4}$$

olur.

(E.3) ifadesinden dolayı $E(U)$=0 olduğundan dolayı,

$$\text{var}(U) = E(U^2) - \left[E(U)\right]^2 = E(U^2) = E(-U') \tag{E.5}$$

olarak elde edilir.

U_j'lerin varyans-kovaryans matrisi olarak tanımlanan *bilgi matrisi*,

$$I_{jk} = E\left[U_j U_k\right] = E\left[\frac{\partial l}{\partial \theta_j} \frac{\partial l}{\partial \theta_k} \right] \tag{E.6}$$

şeklindedir ve (E.4) ifadesinden dolayı bu matrisin elemanları

$$I_{jk} = E\left[-\frac{\partial^2 l}{\partial \theta_j \partial \theta_k} \right] \tag{E.7}$$

olarak da tanımlanabilir.

EK 2.

Genelleştirilmiş doğrusal modeller için log-olabilirlik fonksiyonu $E(Y_i)=\mu_i$ ve g(.) bağlantı fonksiyonu $g(\mu_i)=x_i{}^T\beta=\sum_{j=1}^{p}x_{ij}\beta_j=\eta_i$ olmak üzere, (2.8)'de belirtildiği gibi,

$$l(\theta,\phi;y)=\sum_{i=1}^{N}l_i=\sum_{i=1}^{N}\left\{\frac{y_i\theta_i-b(\theta_i)}{a(\phi)}+c(y_i;\phi)\right\} \tag{E.8}$$

şeklinde yazılabilir. β_j parametresine göre U_j skor fonksiyonu,

$$U_j=\frac{\partial l}{\partial\beta_j}=\sum_{i=1}^{N}\frac{\partial l_i}{\partial\beta_j} \tag{E.9}$$

olarak yazılabilir. U_j skor fonksiyonunu elde etmek için,

$$\frac{\partial l_i(\beta)}{\partial\beta j}=\frac{\partial l_i}{\partial\theta_i}\frac{\partial\theta_i}{\partial\mu_i}\frac{\partial\mu_i}{\partial\beta_j} \tag{E.10}$$

eşitliği kullanılır. (E.10) eşitliğindeki kısmi türevler sırasıyla (2.9) ve (2.13) kullanılarak,

$$\frac{\partial l_i}{\partial\theta_i}=\frac{[y_i-b'(\theta_i)]}{a(\phi)}=\frac{[y_i-\mu_i]}{a(\phi)}, \tag{E.11}$$

$\mu_i=b'(\theta_i)$ ve (2.15) kullanılarak,

$$\frac{\partial\mu_i}{\partial\theta_i}=b''(\theta_i)=\frac{\mathrm{var}(y_i)}{a(\phi)} \tag{E.12}$$

ve $\eta_i=g(\mu_i)$ olduğundan $\frac{\partial\mu_i}{\partial\eta_i}$ türevinin bağlantı fonksiyonuna bağlı olması koşuluyla,

$$\frac{\partial\mu_i}{\partial\beta_j}=\frac{\partial\mu_i}{\partial\eta_i}\frac{\partial\eta_i}{\partial\beta_j}=\frac{\partial\mu_i}{\partial\eta_i}\frac{\partial\sum_{k=1}^{p}x_{ik}\beta_k}{\partial\beta_j}=\frac{\partial\mu_i}{\partial\eta_i}x_{ij} \tag{E.13}$$

şeklinde hesaplanır.

(E.11), (E.12) ve (E.13)'de hesaplanan kısmi türevler (E.3) denkleminde yerine konulup düzenlediğinde skor fonksiyonu,

$$U_j = \frac{\partial l_i(\beta)}{\partial \beta j} = \sum_{i=1}^{N} \frac{(y_i - \mu_i)x_{ij}}{(\operatorname{var} y_i)} \frac{\partial \mu_i}{\partial \eta_i} \qquad \text{(E.14)}$$

olarak elde edilir.

EK 3.

Türevlenebilir bir f(x) fonksiyonu, keyfi bir a noktası ele alınarak bir güç serisi olarak (E.15)'teki gibi gösterilebilir:

$$f(x) = f(a) + f'(a)(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \frac{f'''(a)}{3!}(x-a)^3 + \ldots \qquad \text{(E.15)}$$

Taylor serisi olarak adlandırılan bu açılım f(x) fonksiyonunun x'in a civarındaki değerlerine karşılık gelen değerlerin hesabında kullanılabilir.

Bu açılım, Genelleştirilmiş Doğrusal Modellerde β parametrelerinin kestirimi olan b değerlerinin elde edilmesinde kullanılır. U(β) skor fonksiyonlarının β=b'deki birinci dereceden Taylor yaklaşımı,

$$U(\beta) \cong U(b) + H(b)(\beta - b) \qquad \text{(E.16)}$$

ile verilir. Burada, H(b) log-olabilirlik fonksiyonunun ikinci dereceden türevlerinin matrisini göstermektedir.

Ek 1'de ifade edilen (E.6) ve (E.7) denklemlerinden,

$$I = E(UU^T) = E(-H) \qquad \text{(E.17)}$$

dır.

Bilgi matrisi ile ilişkili olarak H Hessian matrisi beklenen değerine eşittir ve büyük örneklemler için,

$$U(\beta) \cong U(b) - I(\beta - b) \qquad \text{(E.18)}$$

dir.

b noktası log-olabilirlik fonksiyonunun maksimum noktası olduğundan U skor fonksiyonu U(b) = 0 dır. Bundan dolayı, yaklaşık olarak,

$$(b - \beta) \cong I^{-1}U(\beta) \qquad \text{(E.19)}$$

olur.

Burada, I bilgi matrisi sabit olarak adledilir ve eşitliğin her iki tarafının beklenen değeri alınırsa, EÇO kestiricilerinin yansızlık özeliğinden E(U)=0 olacağından,

$$E(b-\beta) \cong I^{-1}E(U) = 0 \tag{E.20}$$

dir. Burada (A.6) ve I bilgi matrisinin simetriklik özellikleri kullanılarak β parametresinin b kestiricisi için varyans-kovaryans matrisi,

$$E\left[(b-\beta)(b-\beta)^T\right] \cong I^{-1}E(UU^T)I^{-1} = I^{-1} \tag{E.21}$$

olur.

Büyük örneklemler için,

$$(b-\beta)^T I(b-\beta) \sim \chi_p^2 \tag{E.22}$$

ya da buna denk olarak,

$$b-\beta \sim N(0, I^{-1}) \tag{E.23}$$

ifadelerine *Wald istatistiği* adı verilir.

EK 4.

Olumsallık çizelgelerinde olabilirlik fonksiyonunun en büyüklenmesi, hücre frekanslarına ait olasılık değerlerine göre alınan türevin sıfıra eşitlenmesi ile mümkündür.

Çok terimli dağılımdan gelen bir örnekleme ait olasılık fonksiyonu, Bölüm 3.2.1.2'de belirtildiği üzere

$$\frac{n!}{\prod_{i=1}^{N} y_i!} \prod_{i=1}^{N} \pi_i^{y_i} \qquad \text{(E.24)}$$

şeklindedir.

Bir fonksiyonun çekirdeği (kernel) fonksiyona ait tüm bilinmeyen parametreleri içerdiğinden, olabilirlik fonksiyonunu en büyüklemek, log-olabilirlik fonksiyonunun çekirdeğini büyüklemekle eşdeğerdir.

(E.24) denklemi ile verilen fonksiyonun çekirdeği $\prod_{i=1}^{N} \pi_i^{y_i}$ olarak alındığında,

$$\text{L} = \ln(\text{kernel}) = \sum_{i=1}^{N} y_i \ln(\pi_i) \qquad \text{(E.25)}$$

elde edilir.

$\pi_i > 0$, i=1, 2,…, N , $\sum_{i=1}^{N} \pi_i = 1$ koşulları altında $\pi_N = 1 - \sum_{i=1}^{N-1} \pi_i$ olarak yazılabilir.

Buradan,

$$\frac{\partial \pi_N}{\partial \pi_i} = -1, \qquad \text{i} = 1,2,\ldots \text{N-1} \qquad \text{(E.26)}$$

$$\frac{\partial \ln \pi_N}{\partial \pi_i} = \frac{1}{\pi_N} \frac{\partial \pi_N}{\partial \pi_i} = \frac{-1}{\pi_N}, \qquad \text{i} = 1,2,\ldots,\text{N-1} \qquad \text{(E.27)}$$

$$\frac{\partial L}{\partial \pi_i} = \frac{y_i}{\pi_i} - \frac{y_N}{\pi_N} = 0, \quad \text{i} = 1,2\ldots,\text{N-1} \qquad \text{(E.28)}$$

bulunur.

(E.28)'den ,

$$\frac{\hat{\pi}_i}{\hat{\pi}_N} = \frac{y_i}{y_N}, \qquad \text{i=1,2,…, N-1} \qquad \text{(E.29)}$$

dir. Böylece,

$$\hat{\pi}_i = \hat{\pi}_N \frac{y_i}{y_N} \qquad \text{(E.30)}$$

bulunur.

$\sum_{i=1}^{N} \hat{\pi}_i = 1 = \frac{\hat{\pi}_N \sum_{i=1}^{N} y_i}{y_N}$ eşitliği kullanılarak EÇO kestiricileri, p oranları, aşağıdaki gibi bulunur:

$$\hat{\pi}_N = \frac{y_N}{n} = p_N$$
$$\hat{\pi}_i = \frac{y_i}{n} = p_i \qquad \text{(E.31)}$$

X ve Y değişkenlerinin bağımsız olması durumunda bu koşul altında EÇO kestiricileri i ve j sırasıyla satır ve sütun indislerini belirtmek üzere,

$$\hat{\pi}_{ij} = p_{i+} p_{+j} = \frac{n_{i+} n_{+j}}{n^2} \qquad \text{(E.32)}$$

ve beklenen hücre frekansları

$$\hat{m}_{ij} = n\hat{\pi}_{ij} = \frac{n_{i+} n_{+j}}{n} \qquad \text{(E.33)}$$

olarak elde edilir.

EK 5.

$Y_1, Y_2, ..., Y_J$, λ_j parametreli Poisson dağılımına sahip bağımsız rastlantı değişkenleri olsun. Bu durumda, Y_j değişkenlerine ait (j = 1,…,J) olasılık fonksiyonu $y = \begin{bmatrix} y_1 \\ \vdots \\ y_J \end{bmatrix}$ olmak üzere,

$$f(y) = \prod_{j=1}^{J} \frac{\lambda_j^{y_j} e^{-\lambda_j}}{y_j!} \tag{E.34}$$

dir.

$n = Y_1 + Y_2 + ... + Y_J$, Y_j bağımsız rastlantı değişkenlerinin toplamından oluşsun. Bu durumda $n \sim Poi(\lambda_1 + \lambda_2 + ... + \lambda_J)$ dağılımına sahip bir rastlantı değişkenidir.

Buradan, y'nin dağılımı n'e bağlı olarak,

$$f(y/n) = \prod_{j=1}^{J} \frac{\lambda_j^{y_j} e^{-\lambda_j}}{y_j!} \Bigg/ \frac{\left(\lambda_1 + \lambda_2 + ... + \lambda_J\right)^n e^{-(\lambda_1+\lambda_2+...+\lambda_J)}}{n!} \tag{E.35}$$

şeklindedir. Bu dağılım indirgenirse,

$$f(y/n) = \left(\frac{\lambda_1}{\sum \lambda_k}\right)^{y_1} ... \left(\frac{\lambda_J}{\sum \lambda_k}\right)^{y_J} \frac{n!}{y_1!...y_J!} \tag{E.36}$$

olur.

(E.36) denkleminde j =1,…,J için $\pi_{ij} = \left(\frac{\lambda_j}{\sum_{k=1}^{K} \lambda_k}\right)$ olarak yazılırsa,

$$f(y/n) = \frac{n!}{y_1! y_2! ... y_J!} \pi_1^{y_1} \pi_2^{y_2} ... \pi_J^{y_J} \tag{E.37}$$

elde edilir. Bu da çok terimli dağılımın olasılık fonksiyonuna eşittir.

Bu durumda, çok terimli dağılım, n toplamına bağlı olarak Poisson rastlantı değişkenlerinin toplamı türünden (E.35)'teki gibi ifade edilebilir (Dobson, 2002).

ÖZGEÇMİŞ

Nihan ACAR, 7 Mart 1985 tarihinde Bursa'da doğmuştur. Ortaokul ve lise eğitimini Nilüfer Milli Piyango Anadolu Lisesi'nde tamamladıktan sonra 2003 yılında Marmara Üniversitesi Fen-Edebiyat Fakültesi Matematik Bölümü'nü kazanmıştır. 2002 ve 2004 yıllarında Bursa Belediyesi ve Almanya Rathenow Belediyesi tarafından organize edilen Nilüfer Milli Piyango Anadolu Lisesi ve Rathenow Dunker Gymnasium arasındaki değişim programlarına misafir öğrenci olarak katılmış, 2003 ve 2005 yıllarında Almanya'dan gelen misafir öğrencilerle ilgilenen komitede yer almıştır. 2006 yılının bahar döneminde Erasmus değişim öğrencisi olarak Viyana Teknik Üniversitesi (TU Wien) Teknik Matematik Bölümü'nde okumuştur. Mart 2008'de Marmara Üniversitesi'nden mezun olmuş ve aynı yıl Mimar Sinan Güzel Sanatlar Üniversitesi Fen Bilimleri Enstitüsü İstatistik Anabilim Dalı'nda yüksek lisans eğitimine başlamıştır. Ocak 2010'da Mimar Sinan Güzel Sanatlar Üniversitesi İstatistik Bölümü Uygulamalı İstatistik Anabilim Dalı'na araştırma görevlisi olarak atanmıştır ve halen bu görevine devam etmektedir.