

**T.C.
DOKUZ EYLÜL ÜNİVERSİTESİ
SOSYAL BİLİMLER ENSTİTÜSÜ
EKONOMETRİ ANABİLİM DALI
EKONOMETRİ PROGRAMI
YÜKSEK LİSANS TEZİ**

**GENELLEŞTİRİLMİŞ DOĞRUSAL MODELLERDE
ETKİLİ GÖZLEMLERİN İNCELENMESİ VE GAMMA
DAĞILIMI UYGULAMASI**

Derya YELER

**Danışman
Prof. Dr. Ali Kemal ŞEHİRLİOĞLU**

İZMİR-2019

TEZ ONAY SAYFASI



YEMİN METNİ

Yüksek lisans tezi olarak sunduğum “genelleştirilmiş doğrusal modellerde etkili gözlemlerin incelenmesi ve gamma dağılımı uygulaması” adlı çalışmanın, tarafımdan, akademik kurallara ve etik değerlere uygun olarak yazıldığını ve yararlandığım eserlerin kaynakçada gösterilenlerden oluştuğunu, bunlara atıf yapılarak yararlanılmış olduğunu belirtir ve bunu onurumla doğrularım.

Tarih

...../...../2019

Derya YELER

İmza

ÖZET
Yüksek Lisans Tezi
Genelleştirilmiş Doğrusal Modellerde Etkili Gözlemlerin İncelenmesi Ve
Gamma Dağılımı Uygulaması
Derya YELER

Dokuz Eylül Üniversitesi
Sosyal Bilimler Enstitüsü
Ekonometri Anabilim Dalı
Ekonometri Programı

Bu tezin amacı; doğrusal veya doğrusal olmayan modellerin normal dağılım ve sabit varyans özelliğini göstermediği durumlar da ilgili üstel dağılış ailesini kapsayan dağılımları kullanarak genelleştirilmiş doğrusal modellerden faydalanıp verilerin dönüşümlerini sonucunda gerekli olan incelemenin sağlanmasıdır.

Tezin genel içeriği ile ilgili bilgi vermek gerekirse; birinci bölüm de temel bir modellemeden istenilen durumlar hakkında bilgi, üstel dağılış ailesinin üyesi olan kesikli ve sürekli dağılımlar hakkında bilgi verilmiştir. İkinci bölüm de genelleştirilmiş doğrusal modellere giriş ve GLM'nin tamamlanması için link fonksiyonu, parametre tahimin yöntemi, model seçimi hakkında bilgi verilmiştir ve son olarak üçüncü bölüm de tıp fakültesinden alınmış verilere ait stata programı uygulaması yapılmış ve uygulama ya ait gerekli bilgilendirmeler yapılmıştır.

Anahtar Kelimeler: Genelleştirilmiş Doğrusal Model, Link Fonksiyonu, Gamma Dağılımı.

ABSTRACT

Master's Thesis

Generalized Linear Models Examination Of Detailed Observations And Application Of Gamma Distribution

Derya YELER

Dokuz Eylül University

Graduate School of Social Sciences

Department of Econometrics

Econometrics Program

The aim of this thesis is; In cases where linear or nonlinear models do not show normal distribution and fixed variance options, they can be filtered and filtered by using generalized linearized models by distributing their families in their respective exponential distribution.

To give information about the general content of the thesis; In the first chapter, you will learn about state status from a basic modeling, learn about discrete and continuous distributions that are members of the exponential etching family. The second part is the introduction to generalized linear models and the connections with GLM, the method of connection analysis, the model design, and the third part is the stata program implementation of the data from the medical faculty and the requirements for the application are planned.

Keywords: Generalized Linear model, link Function, Gamma Distribution.

**GENELLEŐTİRİLMİŐ DOĐRUSAL MODELLERDE ETKİLİ GÖZLEMLERİN
İNCELENMESİ VE GAMMA DAĐILIMI UYGULAMASI**

İÇİNDEKİLER

TEZ ONAY SAYFASI	ii
TEZ ONAY SAYFASI	ii
YEMİN METNİ	iii
ÖZET	iv
ABSTRACT	v
İÇİNDEKİLER	vi
KISALTMALAR	viii
TABLolar LİSTESİ	ix
ŐEKİLLER LİSTESİ	x
 GİRİŐ	 1

**BİRİNCİ BÖLÜM
DOĐRUSAL REGRESYON MODELİ**

1.1. DOĐRUSAL REGRESYON MODELİ	4
1.2. OLASILIK DAĐILIMI	4
1.2.1. Kesilki Dađılımlar	4
1.2.1.1. Bernoulli Dađılımı	4
1.2.1.2. Binom (İki Terimli) Dađılım	5
1.2.1.3. Geometik Dađılım	5
1.2.1.4. Negatif Binom (Eksi İkiterimli) Dađılım	6
1.2.1.5. Poisson Dađılım	6
1.2.2. Sürekli Dađılımlar	7
1.2.2.1. Normal Dađılım	7
1.2.2.2. Beta Dađılımı	8
1.2.2.3. Üstel Dađılım	8
1.2.2.4. GAMMA Dađılım	8
1.2.2.5. Pearson Tip 5 Dađılım	9

İKİNCİ BÖLÜM

GENELLEŞTİRİLMİŞ DOĞRUSAL MODEL

2.1. GENELLEŞTİRİLMİŞ DOĞRUSAL MODEL	10
2.2. GENELLEŞTİRİLMİŞ DOĞRUSAL MODEL DE PARAMETRE TAHMİNİ	12
2.2.1. En Çok Olabilirlik Yöntemi	12
2.3. GENELLEŞTİRİLMİŞ DOĞRUSAL MODELDE ETKİLİ GÖZLEMLERİN BELİRLENMESİ	13
2.4. DAĞILIMLARIN LİNK FONKSİYONLARI	18
2.4.1. Kesikli Dağılımların Link Fonksiyonu	18
2.4.1.1. Bernoulli Dağılım	18
2.4.1.2. BİNOM Dağılımı Link Fonksiyonu	18
2.4.1.3. Negatif Binom Dağılımı Link Fonksiyonu	19
2.4.1.4. Poisson Dağılımı Link Fonksiyonu	20
2.4.1.5. Geometrik Dağılımın Link Fonksiyonu	20
2.4.2. Sürekli Dağılım Fonksiyonları İçin Link Fonksiyonu	21
2.4.2.1. Üstel Dağılımın Link Fonksiyonu	21
2.4.2.2. Normal Dağılımın Link Fonksiyonu	21
2.4.2.3. BETA Dağılımı Link Fonksiyonu	22
2.4.2.4. GAMMA Dağılımı Link Fonksiyonu	22
2.5. GENELLEŞTİRİLMİŞ DOĞRUSAL MODELLERDE MODEL SEÇİMİ	24
2.5.1. Akaike Bilgi Kriteri	25
2.5.2. Bayes Bilgi Kriteri	26

ÜÇÜNCÜ BÖLÜM

GENELLEŞTİRİLMİŞ DOĞRUSAL MODEL İLE İLGİLİ STATA PROGRAMI ÜZERİNDE GAMMA DAĞILIMI UYGULAMASI

3.1. GENELLEŞTİRİLMİŞ DOĞRUSAL MODEL İLE İLGİLİ STATA PROGRAMI ÜZERİNDE GAMMA DAĞILIMI UYGULAMASI	27
SONUÇ	37
KAYNAKÇA	39

KISALTMALAR

GLM	genelleştirilmiş doğrusal model
AIC	akaike bilgi kriteri
BIC	bayes bilgi kriteri
OYF	olasılık yoğunluk fonksiyonu
VB	ve benzeri
BR	birim



TABLÖLAR LİSTESİ

Tablo 1: Üstel Dağılış Ailesi Üyesi Olan Dağılımların Link Fonksiyonları Ve Modelleri	s. 24
Tablo 2: Bayes Bilgi Kriteri Model Seçim Ölçütü	s. 26
Tablo 3: Tıp Fakültesinden Alınan Uygulama Ait Veriler	s. 27



ŞEKİLLER LİSTESİ

Şekil 1: Stata Programına Ait GLM İşlemi Adım1	s. 32
Şekil 2: STATA Programına Ait GLM GAMMA Dağılımı Seçimi Adım 2	s. 33
Şekil 3: STATA Programına GAMMA Dağılımı ait Model1 Çıktısı	s. 34
Şekil 4: STATA Programında GAMMA Dağılımı Model 2 Sonucu	s. 35



GİRİŞ

İstatistik ve ekonometri de çeşitli veriler toplanılarak model tahminlemesi yapısı için matematiksel modeller oluşturulmaktadır. Tahmin edilen modellerin sonucunda katsayı yorulması, modelin anlamlılığı gibi birçok bilgiye verinin boyutuna göre elle ya da programlar aracılığıyla bir model tahmin sonucundan birçok bilgiye ulaşabiliyoruz. İstatistiksel modeller genel olarak, deterministik ve stokastik olmak üzere iki bölümden oluşur. Deterministik modellerde statik ve dinamik olmak üzere ikiye ayrılır. Statik model de değişkenler dönemler içerisinde farklılık göstermezken, dinamik zamana bağlı olarak değişkenlerde değişme olmaktadır. Stokastik modelde ise değişkenlerin olasılık değerleri zamana bağlı olarak değişmiyorsa durağan, zamana bağlı değişkenlerin olasılık değerinde değişiklik oluyorsa durağan olmayan olarak adlandırılır. Stokastik modellerde, sistemin sonucu veya yanıtları değişkenlik gösterebilir.

En temel modelleme yöntemlerinden birisi doğrusal regresyon modelidir. Modelleme de bir bağımlı değişken(açıklanan değişken olarak da bilinir) ve en az bir bağımsız değişken (açıklayıcı değişken) arasındaki ilişkinin matematiksel modelleme ifade edilmesidir, modelin fonksiyonuna göre doğrusal ve doğrusal olmayan modeller iki de ayrılır ve genel olarak modelleme yapılırken modelin normal dağılım göstermesi istenir.

Genelleştirilmiş doğrusal modeller için modelin doğrusal veya doğrusal olmadan bir model olması önemli değildir, genelleştirilmiş doğrusal modeller tahmin edilen modelin normal dağılım göstermediği ve sabit varyans olmadığı durumlar da devreye girer. Uygulama yapılacak veriler toplanarak genelleştirilmiş doğrusal için veri dönüşümleri işlemleri sağlanır, GLM'nin içine kesikli ve sürekli dağılımlardan dağılımlar girmektedir. Bu dağılımların oluşturduğu toplam gruba üstel dağılış ailesi olarak biliriz. Üstel dağılış ailesine kesikli dağılımdan; Bernoulli, binom, negatif binom, poisson geometrik dağılım ve sürekli dağılımdan normal, beta, gamma ve üstel dağılım oluşturmaktadır. Günümüzde genelleştirilmiş doğrusal modeller çoğunlukta sağlık, sigorta poliçeleri, yağışla ilgili v.b alanlarda kullanılmaktadır.

Bu tezin amacı, genelleştirilmiş doğrusal model hakkında genel bilgi edinmek, genelleştirilmiş doğrusal modeller normal dağılım göstermeyen ve sabit varyans varsayımı sahip olmadığında veriler de link fonksiyonun oluşturulması, parametre

tahmini yapılması, son olarak tıp fakültesine alınmış veri olarak stata programı ait sonuçların elde edilmesi.

Nelder ve Wedderburn (1972), Nelder (1977) ve McCullagh ve Nelder (1989) genelleştirilmiş doğrusal modelleri kullanarak istatistiksel modellemeye geniş bir araştırma yapmıştır. Ayrıca, genelleştirilmiş doğrusal modelleme ile değişken dönüştürmeyle ortaya çıkan sorunların büyük ölçüde azaldığını belirtmişlerdir. Buna bağlı olarak da doğrusal regresyon modellerindeki normallik ve varyansın sabitliği varsayımlarının artık gerekli olmadığını fakat varyans değişikliğinin dayandığı araçların bilinmesi gerektiğini belirtmişlerdir.

Robert Schall (1991) genelleştirilmiş doğrusal modellerde sabit etkilerin tahmini, rassaletkiler ve dağılımın bileşenlerini ele almıştır. Rassaletkilerle genelleştirilmiş doğrusal modellerde dağılım bileşenlerini araştırmıştır.

Y. Lee ve J. A. Nelder (1996), genelleştirilmiş doğrusal modellerin doğrusal kestiricilerinde ekstra hata bileşenine olanak tanıyan hiyerarşik genelleştirilmiş lineer modelleri ele almışlardır.

Gareth M. James (2002) genelleştirilmiş doğrusal modelleri, tahmini değişkenlerin eğri veya fonksiyondan gözlemleri ile genişletmek için bir teknik ele almıştır.

Stuart Lipsitz, Michael Parzen, Sundar Natarajan, Joseph Ibrahim, Garrett Fitzmaurice (2004) parçalar halinde ortak değişkenlere sahip olan genelleştirilmiş doğrusal regresyon parametrelerinin tahmini için bir olasılık tabanlı bir yöntem önermişlerdir.

Hans Georg Müller, Ulrich Stadtmüller (2005), çalışmalarında bağımlı değişkenin ölçülebilir ve tahmincisinin rasgele fonksiyonu olan durumlar için genelleştirilmiş fonksiyonel lineer modelleri önermişlerdir.

MBo Hu, JunShao (2008) determinasyon katsayısını kullanarak genelleştirilmiş doğrusal model seçimini incelemiştir

Damla Şentürk, HansGeorgMüller (2009) düzenlenebilir eş değişkenli genelleştirilmiş lineer modelleri incelemiştir.

Naveen Kumar, Christina Mastrangelo, Douglas Montgomery (2011) genelleştirilmiş doğrusal modelleri ele alarak hiyerarşik modelleri araştırmıştır.

John Neuhaus, Charles McCulloch (2011) genelleştirilmiş doğrusal modelleri incelemiş, bağımlı değişkenin çeşitli dağılımları için parametre tahminleri, olabilirlik denklemleri hakkında bilgi vermişlerdir.

ShulingWang, Man Liu, QianqianZhu, Ting Wang (2012) yerel etki analizi açıklayarak, genelleştirilmiş doğrusal modellerde yerel etki analizini ele almışlardır.

Xiaoguang Wang, Junhui Fan (2013) çok değişkenli genelleştirilmiş doğrusal modellerde model seçiminde tutarlılığı için yeni bir ölçüt ortaya koymuşlardır.

Cristian Villegas, Gilberto A. Paula, Francisco José A. Cysneiros ve Manuel Galea (2013) simetrik dağılımları açıklayarak genelleştirilmiş simetrik doğrusal modellerin tahminleri ile artık analizi konularını ele almışlardır.

Gilberto A. Paula (2013) büyük örneklem hacimleri için çift genelleştirilmiş lineer modellerinde model teşhis yöntemlerini açıklamıştır.

Alexandro D. Ramirez, Liam Paninski (2014) beklenen log olabilirlik yoluyla genelleştirilmiş lineer modellerde hızlı çıkarımlar yapmayı ele almışlardır.

Marina Valdora, Víctor J. Yohai (2014) genelleştirilmiş lineer modeller için doğru tahmin veren bir dağılım ailesi önermişlerdir. Burada temel fikir bağımlıdeğişken dönüşümünü stabilize eden bir varyans uyguladıktan sonra bir M-tahmincisi kullanmaktır.

Gauss M. Cordeiro, Edwin Ortega ve ArturLemonte (2015) genelleştirilmiş poisson Hata Oranı Modelini elde etmiştir.

Valentino Dardanoni, Giuseppe De Luca, SalvatoreModica, FrancoPeracchi (2015) ortak değişkenler ile genelleştirilmiş lineer modellerde, model ortalama tahminini ele almışlardır.

BİRİNCİ BÖLÜM

DOĞRUSAL REGRESYON MODELİ

1.1. DOĞRUSAL REGRESYON MODELİ

Doğrusal regresyon modelinin parametreleri doğrusaldır, regresyon modelin de bağımsız değişkenlerin değerleri yinelenen örneklerde değişmez dolayısıyla regresyon analizinde bağımsız değişkene koşullu regresyon analizi olduğu da söylenebilir.

Doğrusal regresyonun tahminlerinin yapılabilmesi için gerekli şartlar;

- Hata terimleri ortalaması sıfır ($\mu=0$) ve varyansı (σ^2) sabit olmalıdır.
- Hata terimleri arasında otokorelasyon yoktur.

$$\text{cov}(\varepsilon_i, \varepsilon_j) = 0 \quad i \neq j$$

- Hata terimlerinin yani ε_i 'lerin varyansları aynıdır, buna homojen varsayımı denir

$$\varepsilon_i \sim N(0, \sigma^2)$$

- Bağımsız değişken ile hata terimi birbirinden bağımsızdır.
- Hata terimleri normal dağılıma sahiptir.

Basit regresyon modeli;

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon_i \quad (1.1)$$

Y_i açıklanan değişken $i=1,2,\dots,n$ doğrusal model için açıklayıcı değişken $X_j \quad j = 1,2, \dots, n$

1.2. OLASILIK DAĞILIMI

Olasılık dağılımı kesikli ve sürekli dağılım olmak üzere ikiye ayrılır.

1.2.1. Kesikli Dağılımlar

1.2.1.1. Bernoulli Dağılımı

Bernoulli dağılımı bir denemenin başarılı veya başarısız, iyi veya kötü v.biki sonuç arasında olması durumunda kullanılan kesikli bir dağılımdır. İki sonuçlu bir

Bernoulli deneyinde ilgilenilen sonuca başarı, diğer sonuca başarısızlık dersek bunu X değişkeni ile olarak aşağıdaki gibi gösterilir.¹

Bernoulli dağılımının olasılık dağılımı;

$$P(X=x)=\begin{cases} p^x(1-p)^{1-x} & x = 1,0 \\ 0 & \text{diğer durum} \end{cases} \quad (1.2)$$

Bernoulli dağılımının beklenen değer ve varyansı;

$$\mu = P\sigma^2 = P(1-P)$$

1.2.1.2. Binom(İki Terimli) Dağılım

Binomrassal değişkeni birbirinden bağımsız n Bernoulli denemesinden başarılı olanların toplam sayısı X rasgele değişkeni olduğu kabul edilirse, bir tek deneme için başarılı olma olasılığı p, başarısız olma olasılığı 1-p ise X'ebinomrassal değişkeni denir. Binom dağılım da p (başarı olasılığı) her deneme de sabittir, denemeler birbirinden bağımsızdır.²

Binom dağılımı olasılık dağılımı;

$$P(X=x)=\begin{cases} \binom{n}{x}p^x(1-p)^{(n-x)} & x = 1,2, \dots, n \\ 0 & \text{diğer durum} \end{cases} \quad (1.3)$$

Binom dağılımının beklenen değer ve varyansı;

$$\mu = np$$

$$\sigma^2 = np(1-p)$$

1.2.1.3. Geometrik Dağılım

Bağımsız Bernoulli denemelerinin her dizisinde her bir deneme için başarı olasılığı p ve ilk başarının elde edilmesi için gerekli denemelerin sayısı X rasgele değişkeni olarak alındığı taktirde X'e geometrik rasgele değişkendir denir. ³

Geometrik dağılım olasılık yoğunluk fonksiyonu;

¹ Irwin Miller, Maryless Miller, Freund, John E Olasılık Dağılımlar matematiksel İstatistik , Litaretür yayıncılık, Sayfa Sayısı 168.

² Irwin Miller, Maryless Miller, Freund, John E Olasılık Dağılımlar Matematiksel İstatistik, Litaretür yayıncılık, Sayfa Sayısı 180.

³ Irwin Miller, Maryless Miller, Freund, John E Olasılık Dağılımlar matematiksel İstatistik, Litaretür yayıncılık, Sayfa Sayısı 180.

$$p(x; p) = \begin{cases} (1-p)^{x-1}p & x = 1, 2, \dots, n \\ 0 & \text{diğer durum} \end{cases} \quad (1.4)$$

geometri dağılımının beklenen değer ve varyansı;

$$\mu = \frac{1}{p} \sigma^2 = \frac{1-p}{p^2}$$

1.2.1.4. Negatif Binom (Eksi İkiterimli) Dağılım

X'inci denemede başarının sağlanma olasılığının p, k'inci başarının x'inci denemede gelme olasılığını belirlemek istediğimiz durumlar da negatif binom dağılımından faydalanabiliriz. Negatif binom dağılımı pascal dağılımı adını da almaktadır. Binom'da n deneyde x başarı elde edilme olasılığı aranırken dağılımında k. başarının x. denemede elde edilme olasılığı aranır.

k = elverişli sonuç

p = bir denemede başarı olasılığı

x = k. başarı elde edilene kadar ki deneme sayısı

Negatif binom olasılık dağılımı;

$$P(X=x) = \begin{cases} \binom{x-1}{k-1} p^k (1-p)^{x-k} & X = k, k+1, k+2, \dots \\ 0 & \text{diğer durum} \end{cases} \quad (1.5)$$

Negatif binom dağılımının beklenen değer ve varyansı;

$$\mu = \frac{k}{p} \sigma^2 = \frac{k}{p} \left(\frac{1}{p} - 1 \right)$$

1.2.1.5. Poisson Dağılım

Poisson dağılımı belirli bir zaman aralığında gerçekleşen olayların kesikli sonuçlarını veren dağılımdır. Örneğin; belirli bir zaman arasında okula gelen öğrenci sayısı gibi.

Poisson dağılımı olasılık fonksiyonu;

$$P(X=x) = \begin{cases} \frac{e^{-\lambda} \lambda^x}{x!} & x = 1, 2, \dots \\ 0 & \text{diğer durum} \end{cases} \quad (1.6)$$

Poisson dağılımının beklenen değeri ve varyansı;

$$\mu = \lambda \sigma^2 =$$

1.2.2. Sürekli Dağılımlar

Bir rassal değişkenin hesaplanamayacak şekil de sınırsız olarak değer olmasıdır ancak bu rassal değişken belirli bir aralık içine alınarak gerekli hesaplamalar yapılabilir.

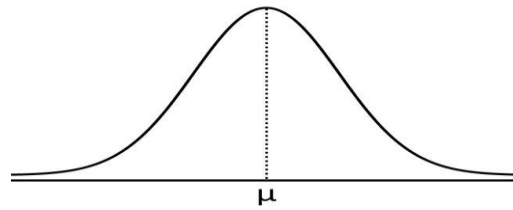
1.2.2.1. Normal Dağılım

İstatistikte en çok kullanılan ve çok geniş uygulama alanına sahip olan normal dağılım ilk olarak 1773 yılında De Moivre tarafından ortaya atılmıştır Normal dağılımlar, istatistik konusunu işlerken tekrar tekrar görülür. Normal dağılımın bazı ilginç özellikleri vardır: çan şeklindedir, ortalama ve medyan eşittir.⁴ Normal dağılımın bazı ilginç özellikleri vardır: çan şeklindedir, ortalama ve medyan eşittir. Normal dağılımlar aşağıdaki özelliklere sahiptir:

- Simetrik çan eğrisi
- Ortalama ve medyan eşittir; her ikisi de dağılım merkezindedir.
- Verinin %68'i ortalamanın 1 standart sapma içine düşer
- Verinin %95'i ortalamanın 2 standart sapma içine düşer
- Verinin %97'si ortalamanın 3 standart sapma içine düşer

Şekil 1: Dağılım Grafiği

Normal Dağılım Grafiği



Normal dağılım olasılık yoğunluk fonksiyonu;

$$F(X;\mu,\sigma)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}-\infty < x < \infty \quad (1.7)$$

⁴ <https://tr.khanacademy.org/math/statistics-probability/modelingdistributionsofdata/normaldistributions-library/a/normal-distributions>

1.2.2.2. Beta Dağılımı

1895 yılında Karl Pearson tarafından tanıtılan beta dağılımını açıklamak için bir beta fonksiyonu tanımlanır ve bu fonksiyon sayesinde beta dağılımı bulunur. Beta dağılımı Eulerian integralinin birinci tipidir.⁵

(a,b) aralığında tanımlanan $f(x) = c(x-a)^{\alpha}(b-x)^{\beta}$ burada $\alpha > -1$ ve $\beta > -1$ bu fonksiyondaki c sabit ve α, β tamsayı olarak ifade ederiz. Olmak üzere $\alpha, \beta > 0$ ve (a,b) tanım aralığı beta dağılımının olasılık yoğunluk fonksiyonu olarak tanımlayabileceğimiz (0,1) aralığında

$B(\alpha, \beta) \int_0^1 x^{\alpha-1} (1-x)^{\beta-1}$ beta fonksiyonu elde ederiz.⁶

beta dağılım olasılık yoğunluk fonksiyonu;

$$f(x; \alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1} (1-x)^{\beta-1} \quad 0 < x < 1 \quad (1.8)$$

Burada $\alpha, \beta > 1$ olup şekil parametreleridir

1.2.2.3. Üstel Dağılım

(Üstel şans değişkeni): Bir X şans değişkeninin olasılık yoğunluğu aşağıdaki tanıma uyuyorsa üstel dağılım gösterir ve üstel şans değişkeni adını alır.

Üstel dağılım olasılık yoğunluk fonksiyonu;

$$f(x; \lambda) = \lambda e^{-\lambda x} \quad x > 0$$

Burada $\lambda > 0$ olup ölçek parametresidir.

1.2.2.4. GAMMA Dağılım

İki parametrelili bir sürekli olasılık dağılımıdır. Bu parametreden biri ölçek parametresi: β , şekil parametresi: α alınır. Üstel dağılım gamma dağılımının özel bir durumudur, üstel şans değişkenlerinin toplamı gamma dağılımını göstermektedir. Bir X şans değişkeni aşağıdaki olasılık

⁵ Ali Kemal ŞEHİRLİOĞLU Sürekli Olasılık Dağılımları, Beta Dağılımı, s. 8, http://kisi.deu.edu.tr/kemal.sehirli/mat%20ist/B%20c3%b6l%20c3%bcm_5_S%20c3%9cREKL%20c4%b0_%20c5%9eANS_DE%20c4%9e%20c4%b0%20c5%9eKEN%20c4%b0_DA%20c4%9eLIMLARI_29122015.pdf

⁶ Irwin Miller, Maryless Miller Freund, John E. , Olasılık Dağılımlar Matematiksel İstatistik , Litaretür Yayıncılık, Sayfa Sayısı 210

yoğunluğuna uyuyorsa (gama dağılımı) gama şans değişkeni olarak adlandırılır⁷

$$F(x;\alpha,\beta)=\frac{1}{\beta^{\alpha}\Gamma(\alpha)}x^{\alpha-1}e^{-x/\beta}x>0$$

Burada $\alpha>0$, $\beta>0$ ve $\Gamma(\alpha)>0$ şeklindedir.

Diğer bir gösterim şekli;

$$g(X;\alpha,\beta)=X^{\alpha-1}\frac{\beta^{\alpha}e^{-\beta/x}}{\Gamma(\alpha)}x>0$$

eğer α bir pozitif tamsayı ise; o halde

$$\Gamma(\alpha)=(\alpha-1)!$$

1.2.2.5. Pearson Tip 5 Dağılımı

Pearson dağılım ailesi 1890 ve 1895 yılları arasında Karl Pearson tarafından tasarlanmıştır. Pearson tip dağılımı aynı zamanda gamma dağılımı veya karşılıklı gamma dağılımı olarak da bilinmektedir. Her üyesinin olasılık yoğunluğu fonksiyonunun $f(x)$ diferansiyel denklemini karşıladığı bir sistemdir.

$$\frac{1}{p} \frac{dp}{dx} = -\frac{a+x}{c_0+c_1+c_2x^2} \quad (1.9)$$

Dağılımın şekli ise a, c_0, c_1 ve c_2 parametrelerine bağlıdır. Pearson tip dağılımına ait denklem şu şekilde yazılır;

$$\frac{d \log f(x)}{dx} = -\frac{x+a}{c_2(x+c_1)^2} \quad (1.10)$$

Pearson tip 5 olasılık yoğunluk fonksiyonu;

$$f(x) = \frac{1}{\beta^{\alpha}\Gamma(\alpha)} \frac{e^{-\beta/x}}{(x/\beta)^{\alpha+1}} \quad a>0, \beta>0 \quad 0 \leq x < +\infty \quad (1.11)$$

Pearson 5 tip dağılımı, bazı minimum gecikmelerden neredeyse kesin olarak ve maksimum gecikmenin sınırlandırılmadığı zaman gecikmelerini modellemek için kullanılmıştır. Örneğin; acil servislerin varışındaki gecikme ve bazı makinelerin onarımı için zaman v.b

⁷Modern Economy, 2015, 6, 1001-1017 Published Online September 2015 in SciRes. <http://www.scirp.org/journal/me> <http://dx.doi.org/10.4236/me.2015.69095> An Application of Gamma Distribution to the Income Distribution and the Estimation of Potential Food Demand Functions

İKİNCİ BÖLÜM

GENELLEŞTİRİLMİŞ DOĞRUSAL MODEL

2.1. GENELLEŞTİRİLMİŞ DOĞRUSAL MODEL

İngiliz John Ashworth Nelder 1924-2010) ile İskoç Robert William MacLagan Wedderburn (1947–1975) bir araya gelerek 1972 yılında Genelleştirilmiş Lineer Modeli formülize etmişler. Genelleştirilmiş lineer modellerde bağımlı değişkenin dağılımı; normal, binom, poisson, üstel ve gamma dağılımlarını içeren üstel ailenin bir üyesi olmalıdır.

Genelleştirilmiş doğrusal model (GLM), genel doğrusal modelden uzanır, doğrusal modeldeki varsayımları gerçekleştirmediği (normal dağılım göstermeyen göstermediği ve sabit varyans) durumlar için veri dönüşümüne alternatif olarak genelleştirilmiş lineer modeller kullanılabilir. Hem lineer hem de lineer olmayan modellerde sonuç çıkarımı için Y bağımlı değişkeninin normal dağılıma sahip olduğu varsayılmaktadır.⁸

Genelleştirilmiş doğrusal model (GLM) üç bileşenden oluşur. Bunlar rassal bileşen, sistematik bileşen ve link fonksiyonudur. Üstel dağılış ailesin de belirli bir parametresi doğal parametre olarak hizmet vermektedir. Bu parametre normal dağılım için ortalama (μ), binom dağılımı için log olasılık ve poisson dağılımı için log ortalamadır. Link fonksiyonu g dönüştüren μ_i doğal parametresine kanonik link denir.

Veri;

y_1	$x_{11}x_{21} \dots \dots x_{k1}$
y_2	$x_{12}x_{22} \dots \dots x_{k2}$
.	.
.	.
$y_1x_{1n}x_{21} \dots \dots x_{kn}$	

$y_1, y_2, y_3, y_4, \dots, y_n$ bağımlı değişken ve $x_1, x_2, x_3, \dots, x_k$ bağımsız değişkendir.

⁸Rameshwar D. Gupta And Debasis Kundu University Of New Brunswick And Institute Of Technology Generalized Exponential Distributions

$$\eta = E(Y_i/X_i) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \dots + \beta_k X_{ki}$$

$$i=1,2,\dots,n$$

$$y_i = E(Y/X_{1i}, X_{2i}, X_{3i}, \dots, X_{ki}) + \varepsilon_i = \mu_y / X_{1i}, X_{2i}, X_{3i}, \dots, X_{ki} + \varepsilon_i \quad (2.2)$$

$$i=1,2,\dots,n$$

$$E(Y) = \mu$$

Üstel ailesi bazen bu form da yazılır;

$$f(Y; \theta, \Phi) = \exp \left\{ \frac{Y\theta - b(\theta)}{\phi} + c(Y, \theta) \right\}$$

Burada b(.) ve c(.) fonksiyon olduğunu biliyoruz. Φ veya Θ bu aralık da Y değişkenin bağımlı değişidir. Bu formül de Θ kanonik parametre ve Φ dispersion parametredir.

Dağılımın y cinsinden ortalaması μ parametrelili ise $\theta = g(\mu)$ bazı g fonksiyon kanonik link olarak isimlendirilir.⁹

Φ parametresi biliyorsa üstel ailesi kanonikformu kullanabilir.

Üstel ailesi bu formda da gösterilmektedir;

$$f(Y; \theta, \Phi) = \exp\{Y_i b(\theta_i) + c(\theta_i) + d(Y_i)\}$$

- RASSAL BİLEŞEN

Bir dağılım da sonu değişkenin rassal bileşen belirler. Doğrusal regresyon da bu durum $Y/X \sim N(\mu, \sigma^2)$

GLM olasılık fonksiyonu;

$$F(X; \Theta, \Phi) = \exp \left(\frac{Y\theta - B(\theta)}{A(\Phi)} + C(Y, \Phi) \right)$$

SİSTEMATİK BİLEŞEN;

X tahminleri bir η parametresiyle ilgilidir, genelleştirilmiş doğrusal model;

$$\eta = E(Y_i/X_i) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \dots + \beta_p X_{pi} = \sum_{j=1}^p X_{ij} \beta_j$$

$i=1,2,\dots,n$ olarak gösterilir.

- LİNK FONKSİYONU;

Rassal ve sistematik bileşenin bileşiminden link fonksiyonu elde edilir.

g. İle gösterilir.

$$g(u) = \eta$$

Doğrusal model de; $E(Y_i)$ beklenen değer ve parametreler arasındaki ilişki doğrusaldır, GLM'de $E(Y_i)$ bir fonksiyon ve parametreler arasındaki ilişki doğrusaldır.

⁹Hutcheson, G. D. (2011). Generalized Linear Models. In L. Moutinho and G. D. Hutcheson, The SAGE Dictionary of Quantitative Management Research. Pages 132-134.

2.2. GENELLEŞTİRİLMİŞ DOĞUSAL MODEL DE PARAMETRE TAHMİNİ

Modellere ait parametrelerin tahmin edilmesi için farklı yöntemler vardı. Bunlar bir tanesi de en çok olabilirlik yöntemi, istatistik ve ekonometride en sık kullanılan nokta tahmin yöntemlerinden biridir. Genelleştirilmiş doğrusal modeller de parametre tahminini gerçekleştirirken en çok olabilirlik yönteminden faydalanabiliriz.

2.2.1. En Çok Olabilirlik Yöntemi

En çok olabilirlik(maximum likelihood estimation) yöntemi ilk defa Crump(1951)tarafından önerilmiş ve Hatley ve Rao tarafından açıklanmıştır. MLE uygulamasındaki amacı modellere ait parametreleri tahmin etmektedir. MLE edicileri olabilirlik fonksiyonunu en yükseğe çıkaran tahmin ediciler olarak tanımlanır. MLE üstel dağılış ailesinde; değişmezlik, yansızlık, tutarlılık ve minimum varyanslılık özelliklerini sağlar. Genelleştirilmiş doğrusal modelin tahmini aşağıdaki gibi gösterilmektedir.¹⁰

Aşağı da β parametresinin $L(Y/\beta)$ fonksiyonuna ait parametre tahmini;

$$L(Y/\beta)=\prod_{i=1}^n \exp\left\{\frac{\theta_i Y_i + c(\theta_i)}{\phi} + d(\phi, Y_i)\right\} \quad (1.15)$$

Log-olabilirlik fonksiyonu;

$$\log L(\beta)=\sum_{i=1}^n \left\{\frac{\theta_i Y_i + c(\theta_i)}{\phi} + d(\phi, Y_i)\right\} \text{ link fonksiyonu} \quad (1.16)$$

β göre türev alınır.

$$\begin{aligned} \frac{d \log L(\beta)}{d \beta_j} &= \sum_{i=1}^n \frac{d \log L}{d \theta_i} \frac{d \theta_i}{d \mu_i} \frac{d \mu_i}{d \beta_j} \\ &= \frac{Y_i + c'(\theta_i)}{\phi} = \frac{y_i - \mu_i}{\phi} \end{aligned}$$

$$\text{a. } \mu_i = -c'(\theta_i) \rightarrow \frac{d \mu_i}{d \theta_i} = -c''(\theta_i) = v(\mu_i)$$

¹⁰ Maximum likelihood estimation of misspecified modelsHWhite Econometrica: Journal of the Econometric Society, 1982 - JSTOR

$$\frac{d\theta_i}{d\mu_i} = \frac{1}{v(\mu_i)}$$

$$b. \frac{d\eta_i}{d\mu_i} = g'(\mu_i) \rightarrow \frac{d\mu_i}{d\eta_i} = \frac{1}{g'(\mu_i)}$$

$$c. \frac{d\eta_i}{d\beta_j} = \frac{d}{d\beta_j} \{ \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \dots + \beta_j X_{ji} + \dots + \beta_k X_{ki} \}$$

SONUÇ;

$$\frac{d \log(\beta)}{d\beta_j} = \sum_{i=1}^n \frac{(y_i - \mu_i) X_{ji}}{\Phi v(\mu_i) g'(\mu_i)} = \sum \frac{(y_i - \mu_i) X_{ji}}{var(y_i) g'(\mu_i)}$$

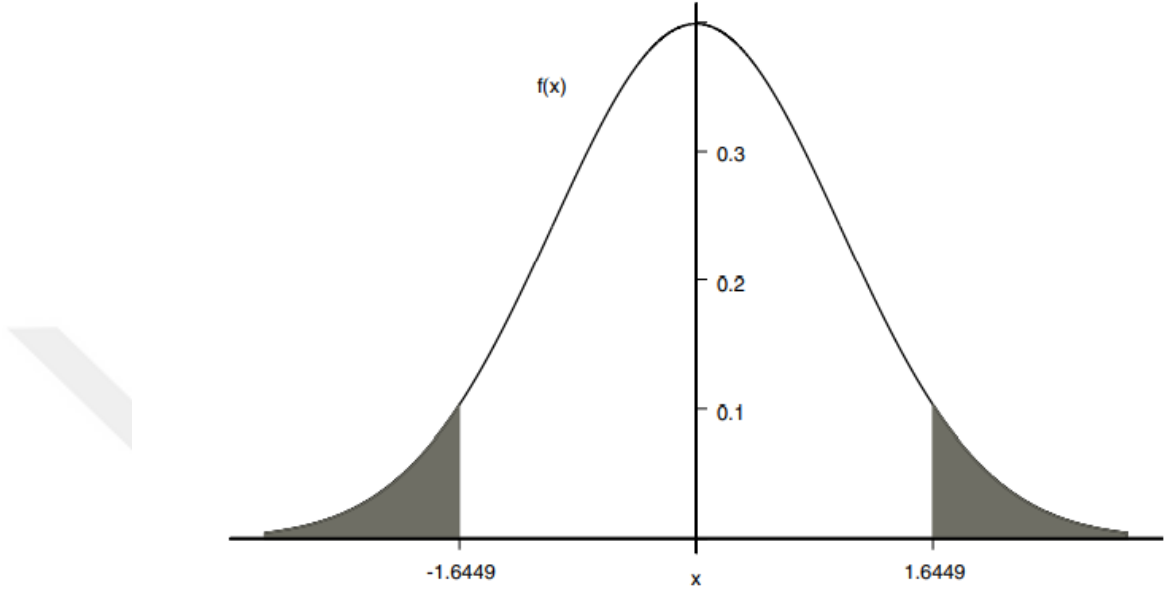
2.3. GENELLETİRİLMİŞ DOĞRUSAL MODELDE ETKİLİ GÖZLEMLERİN BELİRLENMESİ

İstatistiksel analizde veri setlerinden yararlanarak yaptığımız araştırmalarda modele ait parametre tahmini yapılırken her gözlemin tek tek incelenmesi gerekir çünkü bir veri bile modelin katsayı değerini etkileyebilir. Gözlem sonucunda elde elinden veri kümesinde diğer gözlemlere göre küçük ya da büyük olan değerlere aykırı gözlem denir. Gözlem sayısı az olan veri kümelerinde elle hesaplama yapılabilir ancak gözlem sayısı büyük olan veri setleri için daha kolay aykırı gözlemleri hesaplamak için hazır paket (stata, spss, minitab vb) programlar bulunmaktadır veya farklı bir biçimde tanımlamak istersek regresyon analizinde gözlemlerden bir veya birden fazlası önemli derece de farklılık göstermektedir. Böyle gözlemlere *aykırı(outlier)* gözlem denir, gözlem ortalamadan 3 ve 4 standart sapma gösterir, çıkartılan aykırı gözlemler tahmin de önemli derece de değişikliğe neden olduğu için *etkili gözlem* olarak isimlendirilir. Her etkili gözlem aykırı gözlem olabilir ancak her aykırı gözlem etkili gözlem değildir.

Verilerdeki aykırı değerler, özellikle regresyon modellerinde bunları uygun şekilde algılamıyor, veri aktarımında hata yapılması, yanlış yazılması bu gibi tahminleri bozabilir ve doğruluğunu etkileyebilir. Bu nedenle aykırı değer veya değerleri elde etmek zordur ancak GLM bu konu da yardımcı olmak için ilk olarak kabul edebiliriz.

Aşağıdaki grafik de aykırı gözlemin bulunması regresyon çizgisini değiştirmemektedir.

Şekil 2: 0.05 Aykırı Bölgede N(0,1) Dağılım Sahip



$X = \{1, \dots, N\}$, $i \in X$, gözleme sahip olsun. Y_i bağımlı değişken ve X_i bağımsız değişken için aykırı değer veya değerler için n gözlem olduğunu varsayalım.

V^k k'nın aykırı gözlem sayısı belirttiği dış gözlemlerin kümesi olsun. GLM dayanan tüm aykırı değere sahip olan /olmayan model gözüne alındığında;

$$p(y_i | \theta_i, \phi_i, \delta) = \begin{cases} \exp\{\phi_i(y_i, \theta_i) + c_i(\theta_i, \phi_i) + d_i(y_i)\} & i \in X - V^k \\ \exp\left\{\frac{\phi_i}{\delta}(y_i, \theta_i) + c_i(\delta, \theta_i, \phi_i) + d_i(y_i)\right\} & i \in V^k \end{cases}$$

θ_i i şekil parametresi ϕ_i ve δ ölçek parametreleridir; $c(\cdot)$, $d(\cdot)$ bilinen fonksiyonlar.

$g(\Theta_i) = \eta = \sum_{j=1}^p X_{ij} \beta_j + \varepsilon_i$ link fonksiyonunda modele ait hata bileşeni ε_i bağımsız ve normal dağılmıştır. Sonuç olarak $g(\theta_i) | \sigma^2 \sim N(x_i' \beta, \sigma^2)$ biçiminde yazılabilir.

Genellikle tüm gözlemlerin ortalamasının aynı olması beklenir ancak aykırı gözleme ait değerlerde yüksek değişiklik beklenir. ($\delta > 1$ olması) örneğin herhangi bir normal dağılım için $N(\mu, \sigma^2)$ ve α , $0 < \alpha < 1$, iken aykırı bölgeye karşılık gelen bölge;

$$\text{Out}(\alpha, \mu, \sigma^2) = \{X: |X - \mu| > \sigma z_{1-\alpha/2}\}$$

$z_{1-\alpha/2}$, $1-\alpha/2$ standart normal dağılımdır. Eğer $X \in out(\alpha, \mu, \sigma^2)$ ise herhangi bir X için α -outlier(aykırı) $N(\mu, \sigma^2)$ dir. Şekil2 de 0,05 aykırı(outlier) bölgede $N(0,1)$ olduğunu göstermektedir. Mutlak değeri 1.96'dan büyük olan herhangi bir değer aykırı olarak kabul edilir.

P üstel ailesi olsun, $P_i \in P$, f_i yoğunluğa sahip bir ölçüt ve (bilinen) bir destek $supp(P_i)$ vardır. $\alpha_i \in (0,1)$ aralığında, $P_i \in P$ aykırı bölge tanımı;

$$Out(\alpha_i, P_i) = \{X \in supp(P_i) : f_i(X) < K(\alpha_i)\}$$

Buradan;

$$K(\alpha_i) = \sup\{K > 0 : P(\{X : f_i(X) < K\}) \leq \alpha_i\}$$

ile

$Inl(\alpha_i, P_i) = supp(P_i) \setminus out(\alpha_i, P_i)$ α_i -inlier bölge olarak tanımlanır, $X \in out(\alpha_i, P_i)$ α -aykırı olarak P_i ve her $X \in inl(\alpha_i, P_i)$.

Şimdi $\{y_i/x_i \quad i = 1, 2, \dots, n\}$ GLM model olarak kabul edilen modelinde y_i bağımlı değişken olarak tanımlanır ve α_i -aykırı gibi koşullu α -aykırı bölge bulunuyorsa dağılımda, α_i seviyeleri modelin altında olarak şekilde seçilmeli, böylece tüm gözlemlerin α aşma olasılığı yoktur. $\alpha_i \quad i=1, 2, \dots, n$ örneklem büyüklüğüne bağlı olarak seçilmiş;

$$\alpha_i = 1 - (1 - \tilde{\alpha})^{1/n}$$

Bir örneklemdeki aykırı değerleri belirlemede $(y_n/x_n) = \{y_i/x_i \quad i = 1, 2, \dots, n\}$ tüm y_i bağımlı değişkenine karşılık gelen aykırı bölge $out(\alpha_i, P_i)$.

GLM de iki önemli aykırı(outlier) belirleme kuralı vardır: Birincisi bir adımda ilerleyen kural, ikincisi adım adım ilerleyendir. Birinci kurala göre devam ettiğimizde GLM de eşzamanlı ya da tek adım aykırı tanımlayıcı olarak tanımlanan temel de ampirik küme oluşturur $OI_i(\alpha_i, y_n/x_n)$, α - aykırı bölge de $c, i=1, 2, \dots, n$. $OI_i(\alpha_i, y_n/x_n)$ Bulunan her nokta α_i -aykırı(outlier) sınıflandırılır. Asıl sorun P_i ile ilgili kısmen biliniyor, çünkü farklı dağılımların cevapları değişkenlerin farklı değerlerinden kaynaklanır, P_i aynı bilinmeyen karakter parametre β ve bağılım parametresi φ paylaşır. Farklı aykırıların(outliers) standartlaştırma karşılaştırılabilir hale gelmesi için yararlıdır. Davies ve Gather(1993)de iki öneride bulunmuştur. İlk öneri H_0 null modeli altında yapılmıştır

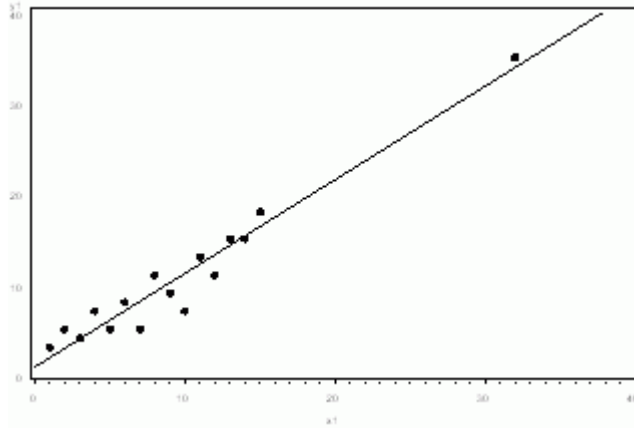
$$P_{H_0}(y_i \notin OI_i(\alpha_i, y_n/x_n) \quad i = 1, \dots, n) \geq 1 - \gamma$$

Bazı $\gamma > 0$ sık sık $\tilde{\alpha}$ seçilir. İkinci öneri null modelin altında

$$P_{H_0}(y_i \in out(\alpha_i, P_i) \quad i = 1, \dots, n) \geq 1 - \gamma$$

Her iki yaklaşım da kaçınılmaz olarak çok sayıda normalleştirme sabitinin türetilmesi (veya simüle edilmesi) zahmetli işine yol açtığı için bu tür bir standardizasyon olmadan ve $out(\alpha_i, P_i)$ tahmin etmeyi öneriyoruz. Eğer α oldukça küçük seçildiyse bu yaklaşım çok fazla tanımlamaya elverişli olmayan düzenli gözlemlerin aykırı(outlier) olarak tanımlama kurallarına yol açar. Gerçek aykırı bölge tanımlama için P_i ($i=1, \dots, n$) tahminciler ihtiyacı vardır ve bunlar tahmin edicileri β 'nın $\hat{\beta}$ tahmincisi, ϕ tahminci $\hat{\phi}$ karşılık gelen yoğunluklara f_i takarak elde edilir. Klasik GLM'deki tahmin edici, çoğu durumda sağlam olmamanın dezavantajına sahip olan Maksimum Olabilirlik (ML) tahmincisidir. Diğer veri durumları için (Davies ve Gather, 1993, Becker and Gather görebiliriz). Model parametrelerinin tahmin edicilerinin güvenilir aykırı(outlier) tanımlama kurallarının sağlamlığa dayanması gerektiği gösterilmiştir. Özellikle, outlier tanımlayıcıları sağlam olmayan tahmin ediciler ile yapılan maskeleyme(masking) ve ve yangıya-düşürme (swamping) etkilerine eğilimlidir. Bir tanımlama kuralı bazılarını tanımlayamazsa, maskeleyme gerçekleşir. Örnek, iki veya daha fazla görünüşte aykırı gözlemler içermesine rağmen (ki bunlar kendileri maskeleymektedir). Örnekteki aykırılar tanımlama kuralının düzenli bir gözlemi aykırı (outliers) olarak da sınıflandırmasına neden olur. Bu girişler bize kullanımını tavsiye etmemizi sağlar. GLM'de aykırı tanımlayıcıların yapımı için sağlam tahminciler

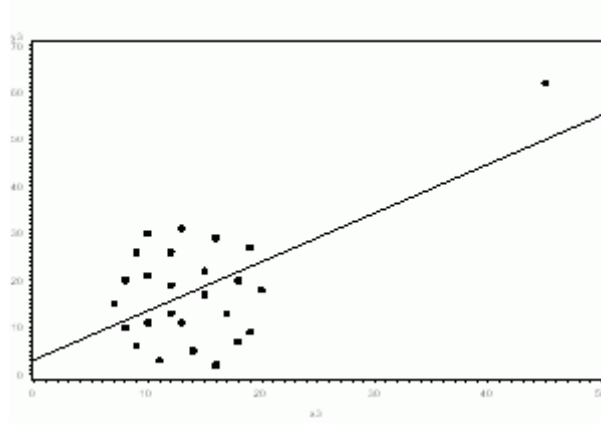
Şekil 3: Regresyon Modeli Aykırı Gözlem Dağılışı



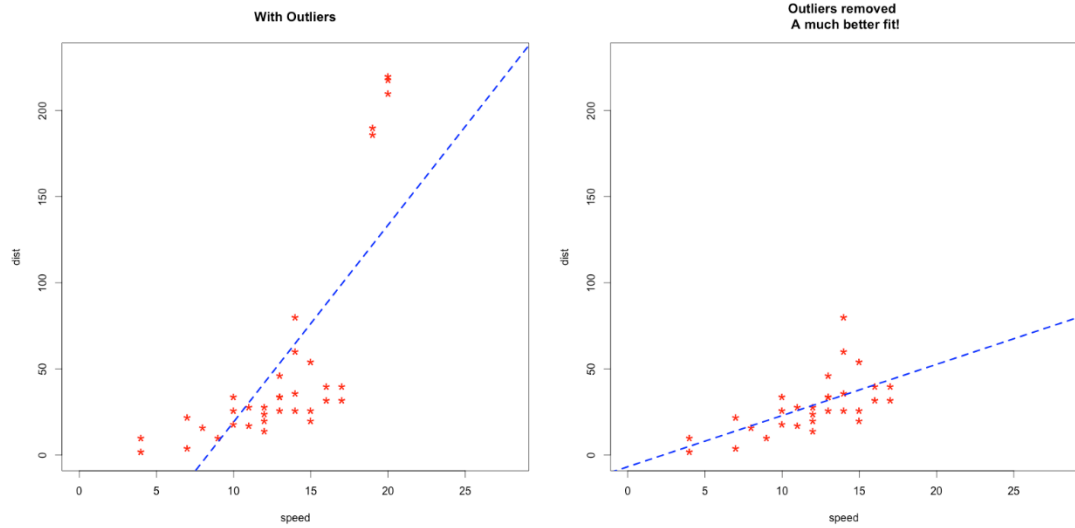
Aykırı gözlem hem sonuçları hem de varsayımları etkiler.

Aşağıdaki grafik de X ve Y arasındaki ilişki yoktur. Dolayısıyla regresyon katsayısı X'in Y üzerindeki etkisi tam olarak tamamlanamaz.

Şekil 4: Regresyon Modelinde Değişkenler Doğru Üzerindeki Dağılışı



Şekil 5: Aykırı Gözlemlerin Dağılımdan Çıkarılardan Önce Ki Ve Sonra Ki Halini Göstermektedir



Aykırı olanları çıkardıktan sonra en uygun çizginin eğimindeki değişime dikkat edildiğinde aykırı olanları modeli (soldaki grafikte) eğrinin eğilmesine neden olmuştur, grafikleri karşılaştırdığımız aykırı gözlemler çıkarıldıktan sonra tahmindeki büyük değişiklik çok net görülmektedir.

2.4. DAĞILIMLARIN LİNK FONKSİYONLARI

Üstel dağılım ailesine giren kesikli ve sürekli dağılımların link fonksiyonları ve dağılımlara ait beklenen değer ve varyansları görülmektedir. Link fonksiyonları bulunduğundan sonra dağılımların ortalama ve varyansı arasındaki doğrusal ilişkinin ortadan kalktığını ispat etmektedir.

2.4.1. Kesikli Dağılımların Link Fonksiyonu

2.4.1.1. Bernoulli Dağılım

$$F(X;P)=p^x(1-p)^{1-x}$$

$$\text{Logf}(X/P)=X\log P+(1-x)\log(1-P)$$

$$=\exp\{x\log P+(1-x)\log(1-P)\}$$

$$=\exp\left\{X\log\left(\frac{P}{1-P}\right)+\log(1-p)\right\}$$

- $f(Y; \theta, \Phi) = \exp\{Y_i b(\theta_i) + c(\theta_i) + d(Y_i)\}$

$$Y_i=X$$

$$b(\theta_i)=\log\left(\frac{p}{1-p}\right)$$

$$c(\theta_i)=\log(1-p)$$

$$d(Y_i)=\text{yok}$$

$$g(\mu_i)=\eta_i=\log\left(\frac{1-p}{p}\right) \quad \text{bernoulli dağılımı} \quad \text{log-link fonksiyona}$$

sahiptir. Bernoulli dağılımının üstel ailesi beklenen değer ve varyansı;

$$E(\mu_i) = B'(\Theta) = \frac{1-p}{p} \quad \text{ve} \quad v(\mu) = \frac{1}{p^2}$$

2.4.1.2. BİNOM Dağılımı Link Fonksiyonu

$$f(x; p) = \binom{n}{x} p^x (1-p)^{(n-x)}$$

$$= \log \left\{ \binom{n}{x} p^x (1-p)^{(n-x)} \right\}$$

$$= \exp \left\{ \log \binom{n}{x} + x \log p + (n-x) \log(1-p) \right\}$$

$$= \exp \left\{ \log \binom{n}{x} + x \log p + n \log(1-p) - x \log(1-p) \right\}$$

$$= \exp \left\{ \log \binom{n}{x} + x \log \left(\frac{p}{1-p} \right) + n \log(1-p) \right\}$$

$$f(Y; \theta, \Phi) = \exp\{Y_i b(\theta_i) + c(\theta_i) + d(Y_i)\}$$

$$Y_i = X$$

$$b(\theta_i) = \log\left(\frac{p}{1-p}\right)$$

$$c(\theta_i) = n \log(1-p)$$

$$d(Y_i) = \log\binom{n}{x}$$

$$E(\mu_i) = B'(\Theta) = \frac{1-p}{p}$$

$$g(\mu_i) = \eta_i = \log\left(\frac{1-p}{p}\right)$$

Binom dağılımı log-link fonksiyona sahiptir. Binom dağılımına ait beklenen değer ve varyansı;

$$E(\mu_i) = b'(\Theta) = \frac{1-p}{p} \text{ ve } v(\mu) = \frac{1}{p^2}$$

2.4.1.3. Negatif Binom Dağılımı Link Fonksiyonu

$$f(X; P) = \binom{x-1}{k-1} p^k (1-p)^{(x-k)} \text{ OYF}$$

$$= \log \left\{ \binom{x-1}{k-1} p^k (1-p)^{(x-k)} \right\}$$

$$= \exp \left\{ \log \binom{x-1}{k-1} + k \log(p) + (x-k) \log(1-p) \right\}$$

$$= \exp \left\{ \log \binom{x-1}{k-1} + k \log(p) + x \log(1-p) - k \log(1-p) \right\}$$

$$= \exp \left\{ \log \binom{x-1}{k-1} + k \log \left(\frac{p}{1-p} \right) + x \log(1-p) \right\}$$

- $f(Y; \theta, \Phi) = \exp\{Y_i b(\theta_i) + c(\theta_i) + d(Y_i)\}$

$$Y_i = X$$

$$b(\theta_i) = \log(1-p)$$

$$c(\theta_i) = k \log \left(\frac{p}{1-p} \right)$$

$$d(Y_i) = \log \binom{x-1}{k-1}$$

$$E(\mu_i) = B'(\Theta) = \frac{1}{1-p}$$

Negatif binom dağılımına ait link fonksiyonu;

$$g(\mu_i) = \eta_i = \log \frac{1}{1-p}$$

Negatif binom dağılımı beklenen değer ve varyansı;

$$E(\mu_i) = b'(\Theta) = \frac{1}{1-p} \text{ ve } v(\mu) = \frac{p}{(1-p)^2}$$

2.4.1.4. Poisson Dağılımı Link Fonksiyonu

$$f(X; \lambda) = \frac{e^{-\lambda} \lambda^x}{X!}$$

$$= \log\left(\frac{e^{-\lambda} \lambda^x}{X!}\right)$$

$$= \exp\{-\lambda + X \log \lambda - \log X!\}$$

- $f(Y; \theta, \Phi) = \exp\{Y_i b(\theta_i) + c(\theta_i) + d(Y_i)\}$

$$Y_i = X$$

$$b(\theta_i) = \log \lambda$$

$$c(\theta_i) = -\lambda$$

$$d(Y_i) = \log X!$$

$$E(\mu_i) = B'(\Theta) = \frac{1}{\lambda}$$

$$g(\mu_i) = \log \mu$$

Poisson dağılımı log-link fonksiyona sahiptir. Poisson dağılımının beklenen değeri ve varyansı;

$$E(\mu_i) = b'(\Theta) = \mu \text{ ve } v(\mu) = \mu$$

2.4.1.5. Geometrik Dağılımın Link Fonksiyonu

$$f(x; p) = (1-p)^{(x-1)} p$$

$$= (x-1) \log(1-p) + \log p$$

$$= \exp\{x \log(1-p) - \log(1-p) + \log p\}$$

$$= \exp\left\{x \log(1-p) + \log\left(\frac{p}{1-p}\right)\right\}$$

- $f(Y; \theta, \Phi) = \exp\{Y_i b(\theta_i) + c(\theta_i) + d(Y_i)\}$

$$Y_i = X$$

$$b(\theta_i) = \log(1-p)$$

$$c(\theta_i)=0$$

$$d(Y_i)=\log\left(\frac{p}{1-p}\right)$$

$$E(\mu_i) = B'(\Theta) = \frac{1}{1-p}$$

$$g(\mu_i) = \eta_i = \log \frac{\mu}{\mu-1} \text{ log link fonksiyona sahiptir.}$$

2.4.2. Sürekli Dağılım Fonksiyonları İçin Link Fonksiyonu

2.4.2.1. Üstel Dağılımın Link Fonksiyonu

$$f(x; \lambda) = \lambda e^{-\lambda x}$$

$$= \log(\lambda e^{-\lambda x})$$

$$= \exp\{\log \lambda - \lambda x\}$$

$$\bullet \quad f(Y; \theta, \phi) = \exp\left\{\frac{Y\theta_i - b(\theta_i)}{\phi} + c(Y, \phi)\right\}$$

$$Y_i = X$$

$$\theta_i = -\lambda$$

$$\phi = 1$$

$$b(\theta_i) = -\log \lambda$$

$$c(y, \phi) = 0$$

$$E(\mu_i) = B'(\Theta) = -\frac{1}{\lambda}$$

$$g(\mu_i) = \eta_i = \log \lambda \text{ üstel dağılım log-link fonksiyonuna sahiptir.}$$

2.4.2.2. Normal Dağılımın Link Fonksiyonu

$$F(X; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$= \log\left(\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\right)$$

$$= \log\left(\frac{1}{\sigma\sqrt{2\pi}}\right) - \frac{(x-\mu)^2}{2\sigma^2}$$

$$= \exp\left\{\log\left(\frac{1}{\sigma\sqrt{2\pi}}\right) - \frac{x^2 - 2x\mu + \mu^2}{2\sigma^2}\right\}$$

$$= \exp\left\{\log\left(\frac{1}{\sigma\sqrt{2\pi}}\right) - \frac{x^2}{2\sigma^2} + \frac{2x\mu}{2\sigma^2} - \frac{\mu^2}{2\sigma^2}\right\}$$

- $f(Y; \theta, \Phi) = \exp \left\{ \frac{y_i \theta_i - b(\theta_i)}{\Phi} + c(y_i, \Phi) \right\}$

$$Y_i = x$$

$$\theta_i = \frac{\mu}{2}$$

$$b(\theta_i) = \frac{\mu^2}{2}$$

$$c(y_i, \Phi) = \log \left(\frac{1}{\sigma \sqrt{2\pi}} \right) - \frac{x^2}{2\sigma^2}$$

$$\Phi = \sigma^2$$

$$g(\mu_i) = \mu$$

Normal dağılım kanonik link fonksiyona sahiptir. Normal dağılımın beklenen değer ve varyansı;

$$E(\mu_i) = B'(\Theta) = \mu \quad V(\mu) = 1$$

2.4.2.3. BETA Dağılımı Link Fonksiyonu

$$f(x; \alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha) \Gamma(\beta)} x^{\alpha-1} (1-x)^{\beta-1}$$

$$= \log; \left(\frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha) \Gamma(\beta)} x^{\alpha-1} (1-x)^{\beta-1} \right)$$

$$= \exp \left\{ \log \left(\frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha) \Gamma(\beta)} \right) + (\alpha - 1) \log x + (\beta - 1) \log(1 - x) \right\}$$

$$= \exp \left\{ \log \left(\frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha) \Gamma(\beta)} \right) + \alpha \log x - \log x + \beta \log(1 - x) - \log(1 - x) \right\}$$

$$= \exp \left\{ \log \left(\frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha) \Gamma(\beta)} \right) + \alpha \log x - \log \left(\frac{x}{1-x} \right) + \beta \log(1 - x) \right\}$$

- $f(Y; \theta, \Phi) = \exp \left\{ \frac{y_i \theta_i - b(\theta_i)}{\Phi} + c(y_i, \Phi) \right\}$

$$\log L(\mu, \phi | y_i) = \log \Gamma(\phi) - \log \Gamma(\mu \phi) - \log \Gamma(\phi - \mu \phi) + (\mu \phi - 1) \log(y_i) + (\phi - \mu \phi - 1) \log(1 - y_i)$$

$$g(\mu_i) = \eta = \log(\mu_i / (1 - \mu_i)) \text{ log-link fonksiyona sahiptir}$$

beta dağılımı beklenen değer ve varyansı;

$$\sigma^2 = \mu(1 - \mu) / (1 + \phi)$$

2.4.2.4. GAMMA Dağılımı Link Fonksiyonu

$$F(x; \alpha, \beta) = \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta}$$

$$\begin{aligned}
&= \log \left(\frac{1}{\beta^{\alpha_{\Gamma}(\alpha)}} x^{\alpha-1} e^{-x/\beta} \right) \\
&= \exp \left\{ \log \left(\frac{1}{\beta^{\alpha_{\Gamma}(\alpha)}} \right) + (\alpha - 1) \log x - \frac{x}{\beta} \right\} \\
&= \exp \left\{ \log \left(\frac{1}{\beta^{\alpha_{\Gamma}(\alpha)}} \right) + \alpha \log x - \log x - \frac{x}{\beta} \right\} \\
&= \exp \left\{ -\alpha \log \beta - \log \Gamma(\alpha) + \alpha \log x - \log x - \frac{x}{\beta} \right\} \\
&= \exp \left\{ -\alpha \log \beta - \log \Gamma(\alpha) + \alpha \log x - \log x - \frac{x}{\beta} \right\}
\end{aligned}$$

- $f(Y; \theta, \Phi) = \exp \left\{ \frac{y_i \theta_i - b(\theta_i)}{\Phi} + c(y_i, \Phi) \right\}$

$$Y_i = x$$

$$\theta_i = 1$$

$$b(\theta_i) = \alpha$$

$$c(y_i, \Phi) = \log \left(\frac{1}{\beta^{\alpha_{\Gamma}(\alpha)}} \right)$$

Gamma dağılımının beklenen değer ve varyansı;

$$E(\mu_i) = B'(\Theta) = \frac{1}{\mu} \text{ ve } v(\mu) = \mu^2$$

$$g(\mu_i) = \frac{1}{\mu} \text{ Gamma dağılımının link fonksiyonu.}$$

Tablo 1: Üstel Dağılış Ailesi Üyesi Olan Dağılımların Link Fonksiyonları Ve Modelleri

Üstel Ailesi Dağılımları	Kanonik Link Fonksiyonu	Varyans Fonksiyonu (μ 'ye göre)	MODEL
POISSON ($\mu=\lambda$)	$\eta=L\log\mu$	μ	$2 \sum_{i=1}^n \left[\left(y_i \log \left(\frac{y_i}{\mu^{\wedge}_i} \right) \right) - (y_i - \mu^{\wedge}_i) \right]$
BİNOM $\mu=np$	$\eta=\log \frac{\mu}{1-\mu}$	$\mu(1-\mu)$	$2 \sum_{i=1}^n \left[\left(y_i \log \left(\frac{y_i}{\mu^{\wedge}_i} \right) + (n - y_i) \log \frac{n-y_i}{n-\mu^{\wedge}_i} \right) \right]$
NORMAL (μ, σ^2)	$\eta=\mu$	1	$\sum_{i=1}^n (y_i - \check{y}_i)^2$
GAMMA	$\frac{1}{\mu}$	μ^2	$\sum_{i=1}^n \left[-\log \left(\frac{y_i}{\mu^{\wedge}_i} \right) + \frac{y_i - \mu^{\wedge}_i}{\mu^{\wedge}_i} \right]$

2.5. GENELLEŞTİRİLMİŞ DOĞRUSAL MODELLERDE MODEL SEÇİMİ

Bilgi kaybını en aza indirmeye çalışmayı içerir. Bilgi teorisi alanı, bilginin beklenen değerini ölçmek veya ölçmek için kullanıldığından model karşılaştırılması yapılarak en doğru modelin seçilmesine yardımcı olur. Modelde en yaygın kullanılan iki kriter Akaike bilgi kriteri ve Bayesian bilgisi kriteridir. Daha düşük bir AIC veya BIC değeri daha iyi bir uyum olduğunu gösterir. Bir uygulamada kullanılan istatistiksel modelin eldeki verilerin önemli özelliklerini yeterince açıklayıp açıklamadığını incelemeyi ifade eder. Model seçimi, birçok rakip model arasında en iyi verileri tanımlayan istatistiksel modelin seçimini ifade eder.

Parametreleri tahmin etmek için maksimum olasılık kullanılıyorsa ve modeller iç içe geçmemişse, model karşılaştırmaları yapmak için Akaike bilgi kriteri (AIC) veya Bayes bilgi kriteri (BIC) kullanılabilir.

2.5.1. Akaike Bilgi Kriteri

Belirli bir veri kümesi için birbirleriyle ilgili olduklarından, mevcut ekonometrik modellerin her birinin kalitesinin tahmini bir ölçüsüdür ve model seçimi için faydalı bir yöntemdir. AIC bilgi teorisindeki temeller ile geliştirilmiştir. Bilgi teorisi; sayıma ve ölçüm süreciyle ilgili uygulamalı matematiğin bir dalıdır. Ekonometrik modellerin göreceli kalitesini ölçmek için AIC kullanılırken, belirli bir model kullanılırsa kaybedilecek bilginin tahminini sağlar. Fakat AIC'nin model belirleme de sınırlaması bulunmaktadır, AIC göreceli modelin kalitesini test etmektedir¹¹. AIC'nin işareti size koşullu parametreler veya modelin uygun olup olmadığı hakkında hiçbir şey söylemez. Örneğin, doğrusal bir regresyon durumunda, AIC pozitif ise, sadece bağımlı değişken üzerindeki her gözlemi aynı sayı ile çarparak negatif yapabilirsiniz (ya da tam tersi). Bu açıkça modelin özelliklerini değiştirmiyor. Bu gibi diğer önlemler gibi, AIC de yalnızca bir modeli aynı verileri açıklamak isteyen farklı bir modelle karşılaştırmak için kullanıldığında yararlıdır (değişken). Kendi başına anlamsız.

İyi bir model, diğer tüm modeller arasında minimum AIC olan modeldir, yani daha düşük bir AIC, bir modelin gerçeğe daha yakın olduğu düşünülür. AIC de parametre sayısı ile birlikte cezai birim $2 \cdot k$ getirerek aşırı uyum riskini azaltmaktadır, N büyüklüğündeki belirli veriye sahip çok fazla parametreye sahip karmaşık modellerin sadeleşmesin yardımcı olur. AIC de yeni açıklayıcı değişken eklendiğin de AIC değeri düşmüyorsa yeni açıklayıcı eklenenin bir anlamı yoktur, çünkü yeni açıklayıcı değişken parametre artışına neden olur.

$$AIC = (-2 \ln L + 2k) / N$$

L: olasılık değeri

N: gözlem sayısı

¹¹j.deleeuw akaike(1993)information theory and an extension of maximum likelihood principle, university of california at los angeles sayfa 600-6003

k: parametre sayısı

2.5.2. Bayes Bilgi Kriteri

Bayesian bilgi kriteri (BIC) veya Schwarz'ın bilgi kriteri Schwarz (1978) tarafından önerilen (SIC), tanımlanan modeller için bir değerlendirme kriteridir. En düşük BIC değerini elde eden model en iyi model olarak seçilmiştir. BIC model seçimin de aşağıdaki tablodan faydalanılabilir;

Tablo 2: Bayes Bilgi Kriteri Model Seçim Ölçütü

FARK	DERECELENDİRME
0-2	ZAYIF
2-5	POZİTİF
5-10	GÜÇLÜ
10 ve fazlası	ÇOK GÜÇLÜ

BIC model seçim ölçütü tablosu

$$BIC = -2\ln L + k\ln(N)$$

L: olasılık değeri

N: gözlem sayısı

k: parametre sayısı

BIC farkının ilk kısmı iki modelin doğruluğunu, ikinci kısmı ise parametre sayısı ile ölçülen karmaşıklığı karşılaştırır

ÜÇÜNCÜ BÖLÜM
GENELLEŞTİRİLMİŞ DOĞUSAL MODEL İLE İLGİLİ STATA PROGRAMI
ÜZERİNDE GAMMA DAĞILIMI UYGULAMASI

3.1. GENELLEŞTİRİLMİŞ DOĞUSAL MODEL İLE İLGİLİ STATA PROGRAMI
ÜZERİNDE GAMMA DAĞILIMI UYGULAMASI

Genelleştirilmiş doğrusal model için yapılan uygulama da tıp fakültesinde yapılan çalışmaya ait veriler kullanılmıştır, verilerin gizlilik hakkından dolayı bağımlı değişken ve açıklayıcı değişkenler kodlanarak gerekli modellemeler yapılmıştır. Veriler gamma dağılışına uygunluk göstermektedir. Stata programında gamma dağılışı gösteren verilere ait GLM tahmini ve tahmin sonucun da bir GLM' ye ait bütün sonuçlar stata çıktısı vermektedir. Şuana kadar ana kütlenin hangi olasılık yoğunluk fonksiyonuna(OYF) sahip olduğunu bilmediğimiz ve bu dağılımın üstel ailesine üyesi ise giriyorsa gerekli dönüşlerin yapılarak sonuçları elde edildi, verilere ait bilgiler öncelikle excell üzerinde gerekli düzenlemeler yapılarak stata programına çekilmiştir, daha sonra

Aşağıdaki tablo da verilere ait bilgiler bulunmaktadır;

Tablo 3: Tıp Fakültesinden Alınan Uygulama Ait Veriler

Y	X1	X2	X3	X4	X5	X6	X7	X8
3,05	2,03	21	2,81	81	180	1725	33	0
1,9	1,48	127	1,78	80	250	1900	20	0
4,06	0,75	118	3,24	55	400	2000	42	0
1,55	1,19	479	2,27	41	300	1100	15	0
1,68	1,19	210	2,43	80	200	4100	18	0
1,8	1,45	226	2,11	81	360	1660	32	0
2	1,45	83	2,38	77	300	3200	21	0
1,5	0,46	34	2,17	76	400	3040	24	0
0,79	0,42	350	2,13	86	350	1750	19	0
2,68	0,95	258	2,08	70	160	1785	27	0
1,02	0,61	225	2,36	83	230	1610	22	0
0,51	0,37	275	2,44	70	300	2650	28	0

2,81	0,63	472	2,74	76	300	700	19	0
0,79	0,48	365	2,35	83	450	7200	17	0
1,67	1,03	242	2,31	79	260	2360	24	0
2,06	1,1	124	2,06	68	300	1500	23	0
1,28	0,56	328	2,4	87	300	900	8	0
1,77	1,13	397	2,29	77	400	4200	23	0
1,28	0,8	114	2,77	85	300	3000	27	0
1,11	0,57	290	2,32	85	300	3000	21	0
0,97	0,61	241	2,05	59	300	1050	32	0
2,93	2,5	345	1,5	71	350	3500	16	0
0,69	0,34	161	1,9	56	300	1500	26	0
1,6	0,81	30	3,04	83	300	900	30	0
2,81	1,15	159	2,59	71	150	2180	30	1
0,66	0,14	417	2,06	37	200	1300	12	1
1	0,57	315	2,22	86	300	3300	23	1
2,44	1,22	236	2,91	84	300	3200	12	1
1,76	0,99	275	2,89	87	300	3000	23	1
1,14	0,59	309	2,56	57	300	2100	19	0
0,75	0,45	7	2,44	66	250	3800	30	0
0,87	0,56	192	2,32	82	300	3000	12	0
1,01	0,54	261	2,66	63	400	2450	24	0
0,68	0,33	443	2,27	57	300	1500	28	1
4	2,67	166	1,82	86	400	1200	21	0
1,61	0,59	274	2,16	64	300	3300	22	0
1,54	0,71	362	2,5	90	300	1350	20	0
2,64	1,6	336	1,99	85	125	625	26	1
1,5	1,04	414	2,02	93	350	2450	17	0
1,96	1,51	37	2,08	79	300	2200	14	0
1,81	1,42	132	2,49	71	150	450	28	0
1,12	0,56	137	1,34	77	350	1850	26	1
0,94	0,58	280	2,49	41	350	6475	16	0
0,66	0,23	29	2,66	34	175	1225	18	0
0,66	0,42	171	2,4	74	300	1500	23	0
2,43	0,51	279	2,14	80	300	1200	16	1
1,35	0,81	35	1,55	72	300	1425	34	1
0,47	0,18	254	2,24	86	350	3500	23	1
0,62	0,36	212	2,15	80	400	3500	24	1

0,95	0,39	4,7	2,01	61	300	1800	31	1
1,53	0,82	61	1,74	87	250	1250	23	0
1,28	0,57	247	2,36	56	300	1800	25	1
1,54	0,88	47	1,7	70	120	360	25	0
1,07	0,65	374	2,71	79	300	3000	20	1
1,37	0,71	499	1,73	77	300	2400	24	0
0,87	0,18	68	2,67	54	300	2100	30	0
2,36	0,96	284	2,8	65	380	1860	25	0
0,99	0,42	306	2,01	75	400	5500	31	0
1,17	0,78	293	3,05	76	300	4115	28	0
1,92	1,43	136	1,76	72	150	1500	29	0
0,42	0,15	156	2,24	73	300	5200	27	0
2,18	1,13	271	2,25	61	300	1650	24	0
2,53	1,5	252	1,85	70	300	2700	24	1
2,31	1,33	374	2,18	80	300	2700	26	1
1,68	1,17	374	1,97	72	300	2150	29	0
1,41	0,93	357	1,98	65	300	2850	25	1
5,04	2,07	40	3,57	67	400	2300	41	0
1,26	0,51	10	2,01	56	300	3000	26	0
1,34	0,85	157	2,07	84	330	4620	18	0
1,67	0,79	277	1,99	72	300	3600	27	1

X9	X10	X11	X12	X13	X14	X15
1	0	1	0	0	6	0
1	1	1	1	0	7	0
1	1	0	0	0	4	0
1	0	0	0	0	0	0
1	0	0	1	0	5	0
1	0	1	0	1	7	0
0	0	0	1	0	4	0
0	0	0	1	0	3	0
1	0	0	0	0	5	0
0	0	1	1	0	9	0
1	0	0	0	0	7	0
0	0	0	1	0	3	0
0	0	0	1	0	6	0
1	0	0	0	0	7	0

0	0	0	1	1	3	0
0	1	0	1	0	6	0
1	0	0	1	0	5	0
1	0	0	1	0	6	0
0	1	0	1	0	5	0
1	0	0	1	0	6	0
0	0	0	1	0	5	0
1	0	1	1	0	5	0
1	1	0	0	0	0	0
1	0	0	0	0	10	0
1	0	0	0	1	5	0
1	0	0	1	1	6	0
1	1	0	1	1	7	0
0	0	0	1	0	9	0
1	0	0	1	0	3	0
0	0	0	1	0	4	0
1	0	0	0	0	4	0
1	0	0	0	0	12	0
0	0	0	1	1	3	0
0	0	0	1	1	3	1
1	0	1	0	0	8	0
0	1	0	0	0	3	0
1	0	0	1	0	6	0
1	0	1	1	1	6	0
1	0	0	1	0	6	0
1	0	1	1	1	7	0
1	0	0	0	0	3	0
1	0	0	0	1	5	0
1	1	0	0	1	2	1
1	1	0	1	0	6	1
1	0	0	1	0	10	0
1	0	0	0	0	9	0
1	0	1	0	0	10	1
1	0	0	0	0	6	0
1	1	0	1	0	7	1
1	0	0	1	0	8	1
1	0	1	0	0	6	0

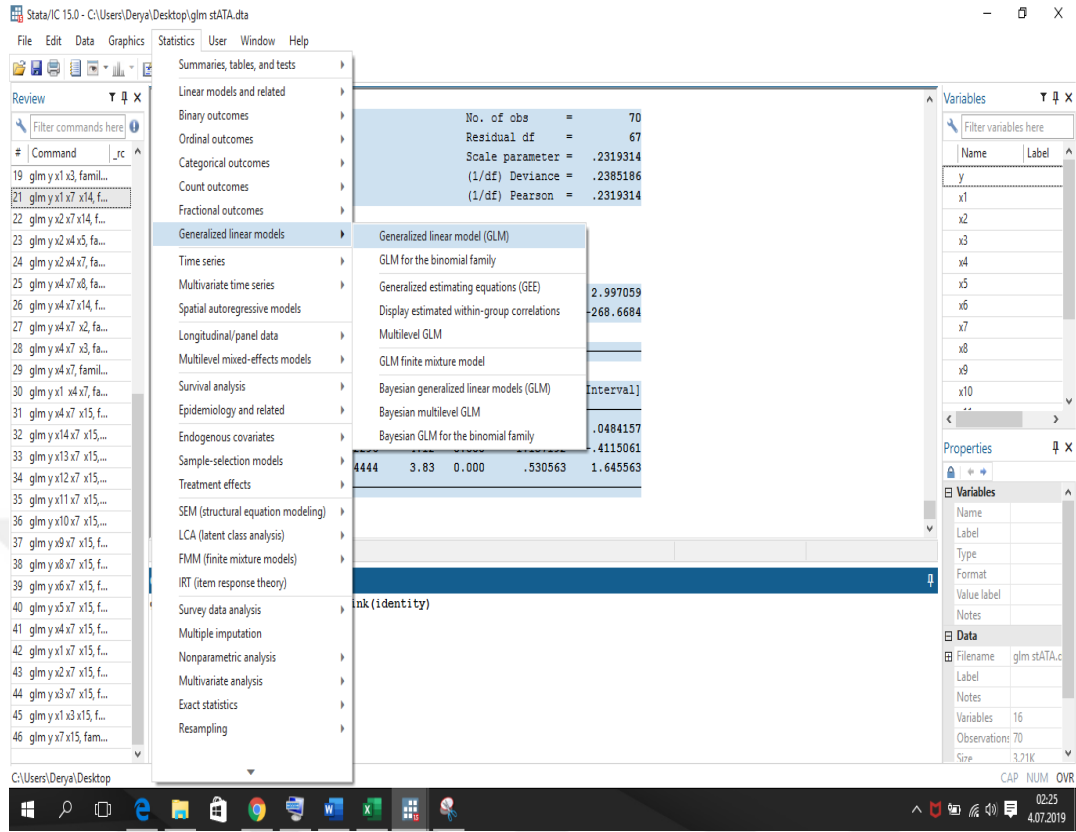
1	0	0	0	1	4	0
1	0	0	0	1	4	0
1	0	0	1	0	5	0
1	0	0	1	0	4	0
1	0	0	1	0	1	0
0	1	1	0	1	5	0
0	0	0	0	0	6	0
1	0	0	0	0	4	0
1	0	0	0	0	3	0
1	1	0	1	0	4	0
1	0	1	1	0	6	0
1	0	1	1	0	4	0
1	0	0	1	0	4	0
1	0	0	1	1	4	0
0	0	0	0	1	5	0
1	1	1	0	0	4	0
0	0	0	1	0	3	1
0	0	0	1	0	7	0
1	0	0	1	0	4	0

Gamma dağılımı verisine uygun iki tane GLM stata programı sonucu aşağı da görülmektedir. Her iki model çıktısında 70 gözlem üzerinden 66 serbestlik derecesine sahiptir. Model de parametre tahmininin en çok olabilirlik yöntemi üzerin de yapılmıştır. Model seçimin de AIC ve BIC kriterleri hesaplandı. Gamma dağılımına ait varyans fonksiyonu $v(u) = u^2$ ve link fonksiyonu $g(u) = u$ olarak bulunmuştur. Aşağı da stata programında GLM uygulamasının nasıl yapıldığı ile ilgili ekran görüntüsü bulunmaktadır. Üstel dağılış ailesine ait olan gamma dağılıma ait bütün işlemleri yaptıktan sonra hangi modeli seçmemiz konusunda bize model seçme kriterleri yardımcı olacak.

H_0 : tıp verilerine ait değişkenleri kullanarak oluşturduğumuz model1, model 2 den daha etkilidir

H_1 : tıp verilerine ait değişkenlerden oluşturduğumuz model2, model1 den daha etkilidir.

Şekil 1: Stata Programına Ait GLM İşlemi Adım1



Şekil 2: STATA Programına Ait GLM GAMMA Dağılımı Seçimi Adım 2

Stata/IC 15.0 - C:\Users\Derya\Desktop\glm stATA.dta

File Edit Data Graphics Statistics User Window Help

Review

Filter commands here

Command _rc

19 glm y x1 x3, fami...

21 glm y x1 x7 x14, f...

22 glm y x2 x7 x14, f...

23 glm y x2 x4 x5, fa...

24 glm y x2 x4 x7, fa...

25 glm y x4 x7 x8, fa...

26 glm y x4 x7 x14, f...

27 glm y x4 x7 x2, fa...

28 glm y x4 x7 x3, fa...

29 glm y x4 x7, fami...

30 glm y x1 x4 x7, fa...

31 glm y x4 x7 x15, ...

32 glm y x14 x7 x15, ...

33 glm y x13 x7 x15, ...

34 glm y x12 x7 x15, ...

35 glm y x11 x7 x15, ...

36 glm y x10 x7 x15, ...

37 glm y x9 x7 x15, f...

38 glm y x8 x7 x15, f...

39 glm y x6 x7 x15, f...

40 glm y x5 x7 x15, f...

41 glm y x4 x7 x15, f...

42 glm y x1 x7 x15, f...

43 glm y x2 x7 x15, f...

44 glm y x3 x7 x15, f...

45 glm y x1 x3 x15, f...

46 glm y x7 x15, fam...

Statistics

Summaries, tables, and tests

Linear models and related

Binary outcomes

Ordinal outcomes

Categorical outcomes

Count outcomes

Fractional outcomes

Generalized linear models

Time series

Multivariate time series

Spatial autoregressive models

Longitudinal/panel data

Multilevel mixed-effects models

Survival analysis

Epidemiology and related

Endogenous covariates

Sample-selection models

Treatment effects

SEM (structural equation modeling)

LCA (latent class analysis)

FMM (finite mixture models)

IRT (item response theory)

Survey data analysis

Multiple imputation

Nonparametric analysis

Multivariate analysis

Exact statistics

Resampling

Generalized linear model (GLM)

GLM for the binomial family

Generalized estimating equations (GEE)

Display estimated within-group correlations

Multilevel GLM

GLM finite mixture model

Bayesian generalized linear models (GLM)

Bayesian multilevel GLM

Bayesian GLM for the binomial family

2.997059

-268.6684

Interval]

.0484157

.4115061

4444 3.83 0.000 .530563 1.645563

link(identity)

Variables

Filter variables here

Name	Label
y	
x1	
x2	
x3	
x4	
x5	
x6	
x7	
x8	
x9	
x10	

Properties

Variables

Name	Label	Type	Format	Value label	Notes

Data

Filename	Label	Notes	Variables	Observations
glm stATA.d			16	70

Size 3.21K

CAP NUM OVR

02:25 4.07.2019

Şekil 3: STATA Programına GAMMA Dağılımı ait Model1 Çıktısı

```
. glm y x1 x3 x6, family(gamma) link(identity)

Iteration 0:  log likelihood = -96.763005
Iteration 1:  log likelihood = -96.59226
Iteration 2:  log likelihood = -96.592143
Iteration 3:  log likelihood = -96.592143

Generalized linear models              No. of obs   =          70
Optimization      : ML                 Residual df   =          66
                                                Scale parameter = .1018942
Deviance          = 5.370881601         (1/df) Deviance = .081377
Pearson           = 6.725015899         (1/df) Pearson  = .1018942

Variance function: V(u) = u^2          [Gamma]
Link function      : g(u) = u           [Identity]

Log likelihood    = -96.59214328        AIC           = 2.874061
                                                BIC           = -275.0298
```

y	OIM					
	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
x1	1.388168	.1415238	9.81	0.000	1.110787	1.66555
x3	.2999821	.1318586	2.28	0.023	.0415439	.5584202
x6	-.0000677	.000027	-2.50	0.012	-.0001207	-.0000147
_cons	-.1078603	.3135325	-0.34	0.731	-.7223727	.5066521

Şekilde model 1'e ait model çıktısını göstermektedir

Model1'e ait model tahmini;

$$y = 0.8143963 - 0.0006334x_2 + 0.0117968x_4 + 0.0003001x_5$$

Model1'de istenilen tahmin sonuçlarına üçüncü iterasyonda ulaşılmıştır, model 1'in log-olabilirlik fonksiyonunun değeri -96,5921 dir.

H_0 : Model1 ait parametreler anlamsızdır. ($\beta_1 = \beta_2 = \beta_3 = 0$)

H_1 : Model 1 parametreleri anlamlıdır. ($\beta_1 \neq \beta_2 \neq \beta_3 = 0$)

%5 anlamlılık düzeyinde katsayıların anlamlı olduğu görülmektedir. Modeldeki açıklayıcı değişkenlerin herhangi bir etkisi olmadığı varsayıldığı da y bağımlı değişkeni 0.8143963 birim olduğu görülmektedir, açıklayıcı x2 değişkenin y bağımlı değişkenini negatif etkilediği x2' de bir birimlik artış y bağımlı değişkenin de 0.000634 azalmaya neden olmaktadır, açıklayıcı değişken x4' de baktığımız da y bağımlı değişkenin de pozitif yön de bir artışa neden olduğunu(yani x4' deki bir br artması y bağımsız değişkenini 0.0117968 etkilemektedir), modelin son açıklayıcı değişkeni olan x5 değişkenin kontrol ettiğimiz de bağımlı değişken olan y üzerin de 0.0003001 bir etkiye sahip olduğu görülmektedir. Standart hata bize modeldeki parametre

tahmini için önemlidir. Standart hatanın sıfır olmasını veya sıfıra yakın olması örneklemelerin birbirlerini ve popülasyonu kusursuz olarak temsil ettiğini söyleyebiliriz. Modeldeki açıklayıcı değişkenlerin standart hatalarını kontrol edildiğin üç açıklayıcı değişken olan x_4, x_5, x_6 standart hataları sırasıyla 0.0079577, 0.0012852 ve 0.0000665dir, yani açıklayıcı değişkenlerin kusursuz olduğu söylenebilir. Model seçim kriteri AIC değeri 3.038617 ve BIC değeri -263.51.09 çıkmıştır.

Şekil 4: STATA Programında GAMMA Dağılımı Model 2 Sonucu

```
. glm y x7 x15, family(gamma) link(identity)
```

Iteration 0: log likelihood = -103.80517
Iteration 1: log likelihood = -101.91772
Iteration 2: log likelihood = -101.89713
Iteration 3: log likelihood = -101.89708
Iteration 4: log likelihood = -101.89708

Generalized linear models No. of obs = 70
Optimization : ML Residual df = 67
 Scale parameter = .2319314
Deviance = 15.98074648 (1/df) Deviance = .2385186
Pearson = 15.53940687 (1/df) Pearson = .2319314

Variance function: V(u) = u^2 [Gamma]
Link function : g(u) = u [Identity]

Log likelihood = -101.8970757 AIC = 2.997059
 BIC = -268.6684

y	OIM				
	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
x7	.0250856	.0119033	2.11	0.035	.0017556 .0484157
x15	-.7843492	.1902296	-4.12	0.000	-1.157192 -.4115061
_cons	1.088063	.284444	3.83	0.000	.530563 1.645563

Şekilde model 2 ait program çıktısını göstermektedir.

Model 2'ye ait tahmin;

$$y = 1.11977 + 0.008163x_4 + 0.0011488x_5 - 0.0001815$$

Model1'de istenilen tahmin sonuçlarına dördüncü iterasyonda ulaşılmıştır, model 1'in log-olabilirlik fonksiyonunun değeri -101,897 dir.

H_0 : Model1 ait parametreler anlamsızdır. ($\beta_1 = \beta_2 = 0$)

H_1 : Model 1 parametreleri anlamlıdır. ($\beta_1 \neq \beta_2 \neq 0$)

%5 anlamlılık düzeyin de açıklayıcı değişkenlere ait katsayıların anlamlılığına baktığımız da x_4, x_5, x_6 açıklayıcı değişkenlerin anlamlı olduğu görülmektedir. Model 2'nin parametrelerin modele olan katkısı; açıklayıcı değişkenlerin hiçbir etki olamadığı

varsayıldığında y açıklanan değişken 1.11977dir. x_4 açıklayıcı değişkenin y bağımlı değişkeni pozitif yön de etkilediği yani x_4 değişkenindeki bir br'lik artış y bağımlı değişkenini 0.008163 artırmaktadır, x_5 değişkenindeki bir br'lik artış y bağımlı değişkeninin de 0.0011499 arttırırken modelin son açıklayıcı değişkeni x_6 'nın y bağımlı değişkenini negatif yönde her bir br'lik artış da 0.0001815 azalmaktadır, çıktı üzerinde parametrelere ait standart hata değerlerine bakıldığında sıfır değerine yaklaşıldığı parametre tahmin değerlerinin geçerli olduğunu görülüyor. Model seçim kriteri olarak aldığımız AIC ve BIC değerleri sırasıyla 3.0383617 ve -263.5109 dur.



SONUÇ

Genelleştirilmiş doğrusal modellerin normal dağılıma ve sabit varyansa sahip olmadığına durumlarda kullanım ihtiyacı doğmaktadır. Doğrusal veya doğrusal olmayan modellerin normal dağılım gösterdiği kabul edilmektedir. Bir modelin normal dağılım sergilemesi için ortalaması ve varyansının birbirine paralel bir artış göstermemesi gerekir. Eğer ortalama arttıkça varyansda artıyorsa bu gibi durum da alternatif modelleme seçeneği olarak genelleştirilmiş doğrusal modeller devreye girmektedir. Sürekli dağılımlardan normal dağılım üstel dağılım aile ve genelleştirilmiş doğrusal modellerde istenilen sonucu çok net göstermektedir. GLM yaşanan bu tarz olayların istenilen durumlarının elde edilmesine yardımcı olur.

Üstel dağılım ailesinin üyeleri bernoulli, poisson, gamma, normal, beta v.b benzeri dağılımlar yer almaktadır. Bu tez çalışmasında gamma dağılımı dikkate alınarak uygulamalar yapılmıştır. Gamma dağılımına ait öncelikle olasılık yoğunluk fonksiyonundan faydalanarak beklenen değer beklenen değer, varyans ve link fonksiyonu oluşturulmuştur.

Bu tez çalışmasında kullanılan veriler tıp fakültesinden alınmıştır, veriler daha önce farklı bir çalışmada kullanıldığını açıklayıcı ve açıklanan değişkenler kodlanarak uygulama yapılmıştır, değişkenlerin kodlanmasından dolayı tam anlamıyla etkili bir sonuç yorumlaması yapılamamıştır, veriler gamma dağılımına uymaktadır. Üstel dağılım ailesinin üyesi olan gamma dağılımına ait veriler stata programı kullanılarak model tahmini yapılmıştır, y açıklanan değişken sabit kalacak şekilde açıklayıcı değişkenler kullanılarak aralarından en anlamlı iki model seçilmiştir, modellere ait katsayıların anlamlılığını ve parametre tahmini en çok olasılık yöntemi tercih edilmiştir. Model tahminleme işlemi gerçekleştirirken seçilen iki modelin açıklayıcı değişken sayısının eşit olmasına gerek yoktur. Bunun nedeni model karşılaştırma işlemi yapılırken akaike bilgi kriterinin özelliğidir, akaike bilgi kriterinde modele eklenen açıklayıcı değişken arttıkça modelin anlamlılığını olumsuz yönde etkilemiyorsa ve AIC değeri küçülüyorsa modeldeki açıklayıcı değişken sayısı artırılabilir. Stata programı model tahminlerinin sonucunda seçilen iki modeli karşılaştırdığımızda model1'in tahmin sonuçlarına model 2'ye göre daha önce ulaşılmıştır, her iki modelin katsayıları %5 anlamlılık düzende anlamlıdır, model seçim kriterleri olan bayes bilgi kriteri ve akaike bilgi kriteri doğru modelin seçimin de önemli etken olmuştur. Bayes bilgi kriterinin her iki modeldeki değerlerine bakıldığında değerler arasında fark azdır ancak bayes bilgi kriterinde model seçiminde Bayes bilgi

BIC model seçim ölçütü tablosu faydalanılmıştır. Tablo 2 göre bakıldığında model seçilmedi. İkinci model seçim kriteri olan akaike bilgi kriterine bakıldığında model 1'in AIC değeri model 2'ye göre daha düşük çünkü model 1'in açıklayıcı değişken sayısı model 2'ye göre fazladır bu durum da model 1'deki AIC değerini olumlu etkilemiştir ve AIC değerinin düşmesini sağlamıştır. AIC değeri azalırken modelde herhangi bir olumsuz etkiye neden olmamıştır, akaike bilgi kriterinin değerlendirme kriteri için önemli olanı çıkan değerin düşük olmasıdır. Sonuç olarak genel model karşılaştırma şartlarına baktığımızda gamma dağılımı gösteren verilerimizin incelenmesinde en iyi modelin model 1 olduğu sonucuna varılmaktadır.



KAYNAKÇA

Abramowitz, M. and Stegun, I. A. (Editors) (1964). *Handbook of Mathematical Functions With Formulas, Graphs, and Mathematical Tables*.

Akaike, H. (1973). Bilgi Teorisi ve Maksimum Olabilirlik İlkesinin Bir Uzantısı. *İkinci Uluslararası Bilgi Teorisi Sempozyumu*, s. 267-28

Anderson, T. W., Fang, K. T. and Olkin, I. (Editors) (1994). *Multivariate Analysis and Its Applications*, Institute of Mathematical Statistics Lecture Notes–Monograph Series, Vol. 24, Institute of Mathematical Statistics, Hayward, California. [271]

Artin, E. (1964). *The Gamma Function*, Holt, Rinehart and Winston, New York. [70]

Beckwith, N. B. and Dudewicz, E. J. (1996). "A bivariate Generalized Lambda Distribution (GLD–2) using Plackett's method of construction: distribution, examples and applications," *American Journal of Mathematical and Management Sciences*, 16, 333–393. [vi, 217, 219, 232, 233, 242, 252, 253, 270, 301].

Bhat, H. S. and Kumar, N., "On the derivation of the Bayesian Information Criterion", <http://nscs00.ucmerced.edu/~nkumar4/BhatKumarBIC.p>

Breslow N. E. (1996). Generalized linear models: Checking assumption and strengthening conclusions. *Statistica Applicata* 8, 23-41.

Brockwell, P. J. and Brown, B. M. (1978). Expansions for the positive stable laws. *Z. Wahrsch. Verw. Gebiete*, 45, 213-224.

Brockwell, P. J. and Brown, B. M. (1979). Estimation for the Positive Stable Laws, I. *Austral. J. Statist.*, 21, 139-148.

Cook, R. D. and Weisberg, S. (1994). ARES Plots for Generalized Linear Models. *Computational Statistics and Data Analysis*, 17, 303-315.

Dobson, A. J. (2001). *An Introduction to Generalized Linear Models* (2nd Edition). Chapman & Hall.

Dunn, K. P. and Smyth, G. K. (1996). Randomized Quintile Residuals. *J. Comput. Graph. Statist.*, 5, 236-244.

Dobson, A. J. and Barnett, A. G. (2002). Exponential family and generalized linear models. *An Introduction to Generalized Linear Models*. Boca Raton, FL: Chapman and Hall/CRC.

Feller, W. (1971). *An Introduction to Probability The organist Applications, Volume II, Second Edition*. Wiley, New York. Section XIII.6-7.

Generalized Linear Models. *Journal of the Royal Statistical Society, A*, 135, 370-384.

Hutcheson, G. D. (2011). Generalized Linear Models. In L. Moutinho and G. D. Hutcheson, *The SAGE Dictionary of Quantitative Management Research*. Pages 132-134.

Hutcheson, G. D. (2011). Generalized Linear Models. In L. Moutinho and G. D. Hutcheson, *The SAGE Dictionary of Quantitative Management Research*. Pages 132-134.

India Deviance- Based Criterion for Model Selection in glm Pages 34-48 | Received 31 Dec 2011

J. K. Ghosh, M. Decamped, and T. Samanta, *An Introduction to Bayesian Analysis: Theory and Methods*. Springer-Verlag, New York, 2006

j.deleeuw akakike(1993). Information Teory and An Extension of Maximum Likelihood Principle, University of California at Losangeles

Jorgensen, B. (1997), *The Theory of Dispersion Models*. Chapman and Hall, London. Section 4.2.1.

Lindsey, J. K. (1997). *Applying Generalized Linear Models*. Springer Texts in Statistics. Springer Verlag.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*. (2nd edition). Chapman and Hall. Nelder, J. \ and Wedderburn, R. W. M. (1972).

Miller I. and M. Miller (1999): John E. Freund's Mathematical Statistics, Sixth Edition, Prentice-Hall, Inc.: New Jersey

Modern Economy, 2015, 6, 1001-1017 Published Online September 2015 in SciRes. <http://www.scirp.org/journal/me><http://dx.doi.org/10.4236/me.2015.69095> An Application of Gamma Distribution to the Income Distribution and the Estimation of Potential Food Demand Functions Generalized linear models advanced methods for data analysis 36-402/ 36-608 2014

National Bureau of Standards Applied Mathematics Series•55, U.S. Government Printing Office, Washington, D.C. 20402. [177]

Nelder, J. A., Wedderburn, R. W. M., "Generalized Linear Models", Journal of The Royal Statistical Society Series A, 135, 03, 370–384, 197.

Newsom PSY 510/610 Categorical Data Analysis Fall 2016 Generalized Linear Models

Nolan, J. P. (1997). Numerical Calculation Of Stable Densities And Distribution Functions. *Communications In Statistics –Stochastic Models*, 13, 759-774.

Pierce, D. A. and Schafer, D. W. (1986). Residuals In Generalized Linear Models. *Journal Of The American Statistical Association*. 81

Raftery, A. (1995). Bayesian Model Selection in Social Research. In Vol. 25 of *Sociological Methodology*, ed. P. V. Marsden, 111-163. Oxford: Blackwell.

