LOJİSTİK REGRESYON ANALİZİ İLE SİGARA İÇME ALIŞKANLIĞININ BELİRLENMESİ: R'DE UYGULAMA

Duygu GÜR

İnönü üniversitesi SOSYAL BİLİMLER ENSTİTÜSÜ Lisansüstü Eğitim-Öğretim Yönetmeliğinin EKONOMETRİ ANABİLİMDALI İçin Öngördüğü YÜKSEK LİSANS TEZİ olarak hazırlanmıştır.

MALATYA

(EYLÜL, 2010)

LOJİSTİK REGRESYON ANALİZİ İLE SİGARA İÇME ALIŞKANLIĞININ BELİRLENMESİ: R'DE UYGULAMA

Duygu GÜR

Danışman : Yrd. Doç. Dr. Ece OMAY

İnönü üniversitesi SOSYAL BİLİMLER ENSTİTÜSÜ Lisansüstü Eğitim-Öğretim Yönetmeliğinin EKONOMETRİ ANABİLİMDALI İçin Öngördüğü YÜKSEK LİSANS TEZİ olarak hazırlanmıştır.

MALATYA

(EYLÜL, 2010)

DUYGU GÜR tarafından hazırlanan "Lojistik Regresyon Analizi İle Sigara İçme Alışkanlığının Belirlenmesi: R'de uygulama" başlıklı bu çalışma, 08.09.2010 tarihinde yapılan savunma sınavı sonucunda başarılı bulunarak jürimiz tarafından Yüksek Lisans tezi olarak kabul edilmiştir.

Prof. Dr. Murat KARAGÖZ ……………………………….

Yrd. Doç. Dr. Rabia Ece OMAY ………………………………..

Yrd. Doç. Dr. Arif KUBAT ………………………………..

Yukarıdaki imzaların adı geçen öğretim üyelerine ait olduğunu onaylarım.

Prof. Dr. Mehmet TİKİCİ

Enstitü Müdürü

ONUR SÖZÜ

Yüksek Lisans Tezi olarak sunduğum “Genelleştirilmiş Doğrusal Modeller Üzerine Bir İnceleme” başlıklı bu çalışmanın, bilimsel ahlak ve geleneklere aykırı düşecek bir yardıma başvurulmaksızın tarafımdan yazıldığını ve yararlandığım bütün yapıtların hem metin içinde hem de kaynakçada yöntemine uygun biçimde gösterilenlerden oluştuğu belirtilir, bunu onurumla doğrularım.

Duygu GÜR

TEŞEKKÜR SAYFASI

Tez çalışmam boyunca değerli yardım ve katkılarıyla beni yönlendiren hocam Yrd. Doç. Dr. Rabia Ece OMAY'a,

Yine kıymetli tecrübelerinden faydalandığım ve tezin uygulama aşamasında gösterdiği yardımlardan dolayı hocam Prof. Dr. Murat KARAGÖZ'e,

Uludağ Üniversitesi kütüphanesinden yararlanmama yardımcı olan ve kaynakların bana ulaşmasında emeği geçen Uğur KOPARAN'a,

Anketlerin uygulanması sırasında büyük yardımları olan İnönü Üniversitesi 2009-2010 öğretim yılı Ekonometri Bölümü son sınıf öğrencilerine,

Manevi destekleriyle beni hiçbir zaman yalnız bırakmayan annem, babam, amcam ve kardeşime teşekkürü bir borç bilirim.

ÖZET

GÜR Duygu. Lojistik Regresyon Analizi İle Sigara İçme Alışkanlığının Belirlenmesi: R'de Uygulama

Yüksek Lisans Tezi, Malatya, 2010.

Doğrusal (linear) ve doğrusal olmayan (nonlinear) regresyon modellerinde yanıt değişkeninin normal dağılım gösterdiği varsayılır. Fakat uygulamada bu varsayım çoğu zaman gerçekleşmez. Bu durumda kullanılacak uygun modellerden biri *genelleştirilmiş doğrusal model*dir (Generalized Linear Model-GLM). GLM iki varsayımı dikkate alır. Bunlardan ilki yanıt değişkeni bağımsızdır, ikincisi ise yanıt değişkeni üstel aileden gelen bir dağılıma sahiptir.

Bağımlı değişkeninin Binom dağılım gösterdiği durumlarda Lojistik Regresyon Analizi, Poisson dağılım gösterdiği durumlarda ise Poisson Regresyon Analizi uygulanır. Temelde ise amaç, doğrusal regresyon analizinde olduğu gibi bağımlı değişken ile bağımsız değişkenler arasındaki neden sonuç ilişkisini en iyi şekilde açıklayan en sade modeli elde etmektir.

GLM, doğrusal ve doğrusal olmayan regresyon modellerinin bir *birleşimi* olarak görülebilir. Model uyumu ve model çıkarsaması aynı sistem (framework) altında gerçekleştirilebilir. Hatta bu birleştirilmiş yaklaşım, yaygın olarak kullanılan ve kullanımı kolay olan bilgisayar yazılımları tarafından da desteklenmektedir. Bu yazılımlardan birisi R'dir. GLM'nin ilk kullanımı sağlık bilimleri ile sınırlı iken hızla diğer bilim alanlarındaki uygulamaları da çoğalmıştır. Bu noktada, çalışmamızda, GLM'nin özellikle sosyal bilimlerde uygulanması ve bu uygulamaların özellikle R programında yapılmasına odaklanılmıştır.

Çalışmada GLM'nin ve üstel ailenin özelliklerine değinilmiştir, lojistik ve poisson regresyon analizi incelenmiştir. Bu tez çalışmasının uygulama kısmında İnönü Üniversitesi İktisadi ve İdari Bilimler Fakültesi öğrencilerinin sigara içme alışkanlığını etkileyen faktörler ortaya konmuş ve gençlerin son zamanlarda ülkemizde uygulanmaya başlanan sigara yasağına olan bakış açılarını ortaya çıkarmak ve bu yasağın sigara kullanımına olan etkisini değerlendirmek hedeflenmiştir. Bu amaçla lojistik regresyon kullanılmış ve ilgili analizler için R programı etkin bir şekilde kullanılmıştır..

Anahtar Kelimeler: Genelleştirilmiş Doğusal Model, Lojistik Regresyon, Poisson Regresyonu, Sigara İçme Alışkanlığı

GÜR Duygu. Determination Smoking Habits by Logistic Regression Analysis : R Application

Post Graduate Thesis, Malatya, 2010.

The response variable is assumed normally distributed in Linear and nonlinear regression models. But in practice, this assumption often does not occur. In this case, one of the appropriate model to be used is generalized linear model (GLM). GLM takes into account two assumption: Response variables are independent and has a distribution from the exponential family.

If the response variable has a binomial distribution, logistic regression analysis is used. Similarly, if the response variable has a Poisson distribution, Poisson regression analysis is used. The main purpose is to obtain, the simplest model that is best to explain the causation between dependent variable and independent variables of causation is obtained.

Generalized linear models can be seen as a combination of linear and nonlinear regression models. GLM, doğrusal ve doğrusal olmayan regresyon modellerinin bir *birleşimi* olarak görülebilir. Inference models and model adaptation can be performed under the same framework. In fact, this combined approach, widely used and easy to use computer software that is supported by. One of these software is R. The first use of the generalized linear model was limited to health sciences applications rapidly increased in other scientific fields. At this point, our study, focuses on the use of generalized linear models in the social sciences and these applications to be made in R.

In this study, basically, the properties of generalized linear models and exponential families are touched upon and logistic and Poisson regression analysis was investigated. In the application part of this thesis, in the Inonu University, Faculty of Economic and Administrative Sciences, the factors affecting the smoking habits of students is revealed. In addition to smoking ban applied in our country of young people to look out and the effect of the ban on smoking to evaluate intended. Logistic regression analysis was used for this purpose and for the R program is used effectively.

Anahtar Kelimeler: Generalized Linear Model, Logistic Regression, Poisson Regression, Smoking Habit

GENELLEŞTİRİLMİŞ DOĞRUSAL MODELLER ÜZERİNE BİR İNCELEME

Duygu GÜR

İÇİNDEKİLER

Sayfa No

ÜÇÜNCÜ BÖLÜM

LOJİSTİK REGRESYON ANALİZİ

DÖRDÜNCÜ BÖLÜM

POİSSON REGRESYON ANALİZİ

BEŞİNCİ BÖLÜM

UYGULAMA

TABLOLAR VE ŞEKİLLER LİSTESİ **Sayfa No**

GİRİŞ

Doğrusal (linear) ve doğrusal olmayan (nonlinear) regresyon modellerinde normal dağılım önemli bir rol oynar. Bu modellerde yanıt değişkeninin normal dağılım gösterdiği varsayılır. Fakat uygulamada bu varsayım çoğu zaman gerçekleşmez. Örneğin, yanıt değişkeni *sayma sayılar* gibi kesikli bir değişken olabilir; yanıt değişkeni *iki değerli değişken* (binary: 0 ve 1) olabilir; ya da yanıt değişkeni sürekli değişken olmasına karşın normallik varsayımını sağlamamaktadır. Bu durumlarda kullanılacak uygun modellerden biri *genelleştirilmiş doğrusal model*dir (Generalized Linear Model-GLM).

GLM *üstel aile* olarak adlandırılan çok daha genel bir dağılım gösteren, tek değişkenli yanıt verileri için regresyon modelleri uyumu yapmamızı sağlamak amacıyla geliştirilmiştir. Üstel aile, normal, binom, Poisson, geometrik, negatif binom, üstel, gamma ve ters normal dağılımları içermektedir. GLM, *yanıt değikeni dağılımı*, *doğrusal kestirici* ve *link fonksiyonu* olarak ifade edilen üç bileşenden oluşur.

GLM, doğrusal ve doğrusal olmayan regresyon modellerinin bir *birleşimi* olarak görülebilir. Model uyumu ve model çıkarsaması aynı sistem (framework) altında gerçekleştirilebilir. Hatta bu birleştirilmiş yaklaşım, yaygın olarak kullanılan ve kullanımı kolay olan bilgisayar yazılımları tarafından da desteklenmektedir. Bu yazılımlardan birisi R'dir. GLM'nin ilk kullanımı sağlık bilimleri ile sınırlı iken diğer bilim alanlarındaki uygulamaları da hızla çoğalmıştır. Bu noktada, çalışmamızda, GLM'nin özellikle sosyal bilimlerde uygulanması ve bu uygulamaların özellikle R programında yapılması oldukça önemlidir.

Bu tez çalışmasındaki amaçlar, GLM için rehber olacak bir çalışma yapmak ve GLM'yi sosyal bilimler alanında uygulamak ve bu uygulamaları yaparken R programını en etkin şekilde kullanmaktır. Bu amaçlar doğrultusunda tez temel olarak beş bölümden oluşmaktadır. Birinci bölümde, diğer bölümlerde detaylı bir şekilde incelenecek olan lojistik, binom ve normal dağılımın konuya giriş olması bakımından genel özelliklerine değinilmiştir. İkinci Bölümde, genelleştirilmiş doğrusal modellere giriş yapılmış, GLM'nin ve üstel ailenin özelliklerine değinilmiştir. Üçüncü bölümde, lojistik regresyon analizi detaylı bir şekilde incelenmiş ve lojistik regresyon modellerinin aşamalarına değinilmiştir. Dördüncü bölümde, üçüncü bölümde olduğu gibi bu seferde poisson regresyon analizi detaylı bir şekilde incelenmiş ve poisson regresyon modellerinin aşamalarına değinilmiştir. Beşinci bölümde ise ampirik bir çalışma adı altında uygulamaya yer verilmiş ve lojistik regresyonun R de uygulanması incelenmiştir.

BİRİNCİ BÖLÜM

BAZI OLASILIK DAĞILIMLARI VE ÖZELLİKLERİ

Regresyon modellerinde bağımlı değişken ya sürekli değerler alan bir değişkendir ya da kesikli değerler alan kategorik bir değişkendir. Bu tez çalışmasında bağımlı değişkenin kesikli değerler aldığı Binom ve Poisson regresyonlar incelenecektir. Bu amaçla bu bölümde, ilerleyen bölümlerde detaylı olarak incelenecek olan Binom ve Poisson dağılımın özelliklerine kısaca değinilecektir. Ayrıca yine bu dağılımlara temel oluşturması sebebiyle normal dağılımdan da kısaca bahsedilecektir.

1.1. Binom Dağılımı

Binom dağılımı Abraham De Moivre tarafından bulunmuştur. 1712'de binom dağılımına Poisson yaklaşımını, 1733'de 66 yaşında iken Binom dağılımına Normal yaklaşımını ortaya koymuştur. De Moivre'in bu sonucu Laplace tarafından 1812'de (Analytical Theory of Probabilities (Olasılıklar İçin Analitik Teori)) geliştirilmiş ve bu sonuç şimdi de Moivre-Laplace teoremi olarak isimlendirilmektedir (Akdeniz, 2007:60).

Bir deneyin sadece iki sonucu varsa, böyle deneylere Bernoulli deneyi denir. Bu sonuçlar, bir sınav sonucunun başarılı veya başarısız olması ya da kalite kontrolü için alınan bir ürünün sağlam veya kusurlu olması şeklinde ortaya çıkabilir.

Bernoulli dağılımında deney bir kez yapılır ve olumlu veya başarılı sonuçla ilgilenilir. Eğer deney n kez peşpeşe birbirinden bağımsız olarak tekrarlanırsa Bernoulli dağılımının özel bir hali olan Binom dağılımı ortaya çıkar.

Binom dağılımının kullanılabilmesi için aşağıdaki şartların sağlanması gerekir:

- Rassal deney aynı koşullar altında n kez tekrarlanmalı
- Her deneyin olumlu-olumsuz, başarılı-başarısız gibi iki mümkün sonucu olmalı
- n deney için başarı (p) ve başarısızlık $(1 - p = q)$ bir deneyden ötekine değişiklik

göstermemelidir. Yani n deney için p ve $1 - p$ sabit olmalıdır.

- Bir deneyin sonucu diğer deneylerin sonuçlarını etkilememelidir. Yani her deney

birbirinden bağımsız olmalıdır.

- Tek bir deneyde istenen sonuç p ise deney n kez tekrarlandığında x kez istenen sonucu

elde etme olasılığını

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

fonksiyonu veriyorsa, başarı sayısını gösteren X rassal değişkeni bir binom dağılımına sahiptir denir. Binom dağılımını belirleyen $n\ ve\ p$ değerleri aynı zamanda bu dağılımın parametreleridir.

X rassal değişkeni binom dağılımı gösteriyorsa aşağıdaki gibi gösterilir:

$X \sim B(n, p)$

p=q durumunda binom dağılımı simetriktir. $p \neq q$ durumun da ise simetriden uzaklaşılır. n sabit kaldığında $p \to 0.5$ için ve p sabit kaldığında $n \to \infty$ için dağılım simetriye yaklaşır.

Ortalaması:

Binom dağılmış bir rassal değişken için ortalama,

$$E(X) = \sum_{x=0}^{n} x\, P(x) = \sum_{x=0}^{n} x \binom{n}{x} p^x q^{n-x} = \sum_{x=0}^{n} x \frac{n!}{x!(n-x)!} p^x q^{n-x}$$

$$E(X) = \sum_{x=0}^{n} x \frac{n!}{(x-1)!(n-x)!} p^x q^{n-x} = np \sum_{x=1}^{n} x \frac{(n-1)!}{(x-1)!(n-x)!} p^{x-1} q^{n-x}$$

$$E(X) = np \sum_{x=1}^{n} \binom{n-1}{x-1} p^{x-1} q^{n-x}$$

k=x-1 olarak tanımlanırsa,

$$E(X) = np \sum_{k=0}^{n-1} \binom{n-1}{k} p^k q^{n-k-1}$$

ve m=n-1 olarak tanımlanırsa,

$$E(X) = np \sum_{k=0}^{m} \binom{m}{k} p^k q^{m-k}$$

olur.

$$\sum_{k=0}^{m} \binom{m}{k} p^k q^{m-k} = 1$$

olduğundan,

$$E(X) = np$$

olarak bulunur.

Varyansı:

μ'e X rassal değişkeninin beklenen değeri dersek varyans şu şekilde tanımlanır:

$$V(X) = E[(X - \mu)^2] = E[(X - E(X))^2] = E(X^2) - [E(X)]^2$$

Öncelikle $[E(X)]^2$'nin hesaplanması gerekir.

$$E(X^2) = \sum_{x=0}^{n} x^2 \binom{n}{x} p^x q^{n-x} = \sum_{x=0}^{n} x^2 \frac{n!}{x!(n-x)!} p^x q^{n-x}$$

$$E(X^2) = np \sum_{x=1}^{n} x \frac{(n-1)!}{(x-1)!(n-x)!} p^{x-1} q^{n-x}$$

$$E(X^2) = np \sum_{x=1}^{n} (x - 1 + 1) \frac{(n-1)!}{(x-1)!(n-x)!} p^{x-1} q^{n-x}$$

$$E(X^2) = np\left[\sum_{x=1}^{n}(x-1)\frac{(n-1)!p^{x-1}q^{n-x}}{(x-1)!(n-x)!} + \sum_{x=1}^{n}\frac{(n-1)!p^{x-1}q^{n-x}}{(x-1)!(n-x)!}\right]$$

olur.

Burada $(p+q)^{n-1} = 1$ olduğundan ikinci toplamın değeri 1'e eşit olacaktır.

$$E(X^2) = np\left[(n-1)p\sum_{x=2}^{n}\frac{(n-2)!}{(x-2)!(n-x)!}p^{x-2}q^{n-x} + 1\right]$$

$$E(X^2) = np\left[(n-1)p\sum_{x=2}^{n}\binom{n-x}{x-2}p^{x-2}q^{n-x} + 1\right]$$

$$E(X^2) = np[(n-1)p+1] = n^2p^2 - np^2 + np = n^2p^2 + npq$$

olur.

$$V(X) = E(X^2) - [E(X)]^2 = n^2p^2 + npq - n^2p^2 = npq$$

şeklinde elde edilir (Aytaç, 1999:224).

1.2. Poisson Dağılımı

Poisson dağılımı Fransız matematikçi Siméon-Denis Poisson (1781–1840) tarafından 1837 yılında bulunmuş ve ilk uygulaması Prusya ordusundaki ölen asker sayısının at tepmesine bağlı olarak nasıl bir dağılım gösterdiği şeklinde yapılmıştır.

Binom dağılımında tanımlanan *p* veya *q* olasılıklarından birinin çok küçük olması halinde bu dağılım uygun bir matematik model olmamaktadır. Bir rassal değişkenin belli bir zaman aralığında veya belli bir mekânda çok az yinelenen olayları göstermesi durumunda ortaya çıkan olasılık dağılımı Poisson olasılık dağılımı veya sadece Poisson dağılımı olarak adlandırılır. Aslında teoride Binom dağılımı ile çözülebilecek problemleri Poisson dağılımı ile daha ekonomik ve etkili bir şekilde çözebiliriz. Örneğin herhangi bir kavşaktaki trafik kaza sayısının belirli zaman dilimindeki ihtimal dağılımı Binom ve Poisson dağılımları yardımı ile belirlenebilir. Binom dağılımına başvurursak Binom dağılımı bizden o kavşaktan geçen, kaza yapan ve yapmayan tüm arabaların sayısını isteyecektir. Bunun tesbiti çok zaman ve para alacağından bu yol oldukça masraflı olacak ve hata yapma riski de buna bağlı olarak artacaktır. Poisson dağılım ise bizden sadece istenilen zaman dilimlerinde geçmiş polis kayıtlarındaki kaza sayısını talep edecektir. Bu sayı tesbit edildiğinde istenilen zaman diliminde aritmetik ortalaması bulunur ve ihtimalin dağılımı tesbit edilir. Dolayısı ile Poisson dağılımını kullanmak çok daha ekonomiktir.

Poisson dağılımının kullanılabilmesi için aşağıdaki şartların sağlanması gerekir:

- Belirlenen periyotta meydana gelen ortalama olay sayısı sabittir.
- Herhangi bir zaman diliminde bir olayın meydana gelmesi, bir önceki zaman diliminde meydana gelen olay sayısından bağımsızdır.

- Mümkün olabilecek en küçük zaman aralığında sadece bir olay gerçekleşebilir.
- Ortaya çıkan olay sayısı ile periyodun uzunluğu doğru orantılıdır.

Poisson dağılımı çok çeşitli alanlarda uygulama alanı bulmaktadır. Örneğin, doktor ofisinde, otobüs durağında bekleme gibi durumlar için geliştirilen modellemelerde, belirli kelimenin bir kitapta geçme sayısı, kitaptaki yanlış yazılmış kelime sayısı, bu ay olacak yağmurlu gün sayısı gibi tahminleri ve saat başı kanserden ölenlerin sayısı veya herhangi bir virüsten ölenlerin sayısı gibi durumlarda uygulanabilir.

X rassal değişkeni yukarda bahsedilen özellikleri taşıyorsa, ona Poisson rassal değişkeni ve X'in fonksiyonuna da Poisson dağılımı denir. $\lambda > 0$ olmak üzere aşağıdaki gibi tanımlanır:

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

Poisson dağılımının tek parametresi λ'dır ve bu parametre kesirli değerlere de sahip olabilir.

Ortalaması:

$$E(X) = \sum_{x=0}^{\infty} x\, P(x) = \sum_{x=0}^{\infty} x \frac{e^{-\lambda}\lambda^x}{x!} = \lambda e^{-\lambda} \sum_{x=1}^{\infty} \frac{\lambda^{(x-1)}}{(x-1)!}$$

Ayrıca $1+\frac{\lambda}{1!} + \frac{\lambda^2}{2!} + \cdots += \sum_{x=0}^{\infty} \frac{\lambda^x}{x!} = e^{\lambda}$

olduğundan,

$$e^{\lambda} = \sum_{x=1}^{\infty} \frac{\lambda^{(x-1)}}{(x-1)!}$$

elde edilir.

$$E(X) = \lambda e^{-\lambda} e^{\lambda} = \lambda$$

olarak bulunur.

Varyansı:

$$E(X^2) = \sum_{x=0}^{\infty} x^2 p(x) = \sum_{x=0}^{\infty} \frac{[x(x-1)+x]e^{-\lambda}\lambda^x}{x!}$$

$$E(X^2) = \sum_{x=0}^{\infty} \frac{x(x-1)e^{-\lambda}\lambda^x}{x!} + \sum_{x=0}^{\infty} \frac{xe^{-\lambda}\lambda^x}{x!} = \sum_{x=2}^{\infty} \frac{e^{-\lambda}\lambda^x}{(x-2)!} + \lambda$$

$$V(X) = E(X^2) - [E(X)]^2 = \lambda^2 + \lambda - \lambda^2 = \lambda$$

olarak bulunur (Aytaç, 1999:249).

1.3. Normal Dağılım

19. yüzyılın ilk yarısında Gauss'un katkılarıyla Normal Dağılım ve Çan eğrisi istatistikte yerini almıştır. Normal dağılım, ayrıca "Laplace-Gauss Dağılımı" olarak da bilinir (Akdeniz, 2007:238).

Dağılımlar arasındaki ilişkilerde Normal dağılım, tam bir merkez konumundadır. Normal dağılım, gerek kendi özelliğinden dolayı gerekse teoremler yardımıyla uygulamada o kadar geniş alanlar yaratır ki, bazı rassal değişkenlerin dağılımlarını, ister kesikli ister sürekli olsun, Normal dağılıma yaklaştırmak isteği ağırlık kazanır.

Normal dağılım başlıca üç alanda yoğun olarak kullanılmaktadır:

- Uygulamada ele alınan birçok değişken Normale benzer bir dağılım gösterir.

Örnek olarak, ölçme hataları, bir fabrikada üretilen vidaların uzunlukları ve belli bir sürede uçakların almış olduğu yol verilebilir. Aslında bu tür rassal değişkenlerin dağılımları tam olarak bir Normal dağılıma uymasa da yaklaştıkları görülür. Fakat uygulamada çok sayıda birbirinden bağımsız olarak ortaya çıkan rassal değişkenlerin bir Normal dağılım gösterdikleri kabul edilir.

- Normal dağılımın, istatistik tümevarım ve örnekleme teorisinde önemli bir ağırlığı

vardır. Çünkü, örneklemden elde edilen aritmetik ortalama, toplam gibi bazı niteleyici değerlerin örnekleme dağılımları, anakütle normal dağılmasa bile, örneklem hacmi n yeterince büyük seçildiğinde $(n \geq 30)$ normale yaklaşır.

- Örnekleme dağılımları olan ki-kare, t ve F dağılımları Normal dağılımdan

türetilmiştir. Ayrıca örneklem hacmi n arttıkça, normal dağılım binom ve poisson dağılımlarının çok iyi bir yaklaşımını oluşturur.

X rassal değişkeni gerçel sayılar uzayında tanımlanmak üzere,

$$f(x) = \begin{cases} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} & , \quad \sigma > 0, \quad -\infty < x < +\infty \\ 0 & , \quad \quad diğer\ durumlarda \end{cases}$$

olasılık yoğunluk fonksiyonuna sahipse normal dağılmıştır. μ ve σ normal dağılımın parametreleridir.

X rassal değişkeni normal dağılım gösteriyorsa şu şekilde gösterilir:

$X \sim N(\mu, \sigma^2)$

Normal dağılımın özellikleri şu şekildedir:

- Dağılım aritmetik ortalama etrafında simetriktir. Aritmetik ortalama aynı zamanda mod ve medyana eşittir.

- Dağılım, $X = \mu$ değerinde bir maksimuma sahiptir. $X = \mu$, dağılımda yerine koyulursa maksimum noktayı veren y değeri,

$Y = f(x) = \frac{1}{\sigma\sqrt{2\pi}}$

olarak elde edilir.

- y=0 doğrusu aynı zamanda bir yatay asimptottur.
- Normal eğri ile X ekseni arasında kalan alanın değeri 1'e eşittir.
- Normal dağılımın parametreleri μ ve σ^2 olduğu için bunların aldığı değerlere göre normal dağılımın grafiği değişir.

Ortalaması:

$$E(X) = \int_{-\infty}^{\infty} xf(x)dx = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} xe^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx$$

Bu integral $z = \frac{x-\mu}{\sigma}$ değişkeni tanımlanıp değişken değiştirme yaklaşımı ile alınabilir.

$z = \frac{x-\mu}{\sigma} \Rightarrow x = \mu + z\sigma$ ve $dx = \sigma dz$

bulunur. İkinci integralin değeri sıfırdır. Çünkü,

$$E(X) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} (\mu + z\sigma)e^{-\frac{1}{2}z^2} \sigma dz$$

$$E(X) = \frac{1}{\sigma\sqrt{2\pi}} = \int_{-\infty}^{\infty} \mu e^{-\frac{1}{2}z^2} dz + \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z\sigma e^{-\frac{1}{2}z^2} dz$$

$\frac{z^2}{2} = u$ ise $du = zdz$

olur.

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z\sigma e^{-u} \frac{1}{z} du = 0$$

dır.

Aynı şekilde $\frac{z^2}{2} = u$ olmak üzere birinci integrale değişken değiştirme yaklaşımı uygulanırsa,

$$E(X) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} \mu e^{-\frac{1}{2}z^2} dz = \mu$$

şeklinde bulunur.

Varyansı:

$$V(X) = E[(X-\mu)] = \int_{-\infty}^{\infty} (X-\mu)^2 \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx$$

$z = x - \mu$ olarak tanımlanırsa dz=dx olur.

$$V(X) = \frac{1}{\sigma\sqrt{2\pi}} z^2 e^{-\frac{z^2}{2\sigma^2}} dz$$

Bu integrale $u = z$ ve $dv = ze^{-\frac{z^2}{2\sigma^2}}dz$ ile kısmi integral yaklaşımı uygulanırsa,

$du = dz$ ve $v = -\sigma^2 e^{-\frac{z^2}{2\sigma^2}}$

olur.

$$V(X) = \left[\frac{-z}{\sigma\sqrt{2\pi}}\sigma^2 e^{-\frac{z^2}{2\sigma^2}}\right] + \int_{-\infty}^{\infty}\frac{\sigma^2}{\sigma\sqrt{2\pi}} e^{-\frac{z^2}{2\pi^2}}dz$$

$$\int_{-\infty}^{\infty}\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{z^2}{2\pi^2}}dz = 1$$

olduğundan dolayı,

$$V(X) = 0 + \sigma^2\int_{-\infty}^{\infty}\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{z^2}{2\pi^2}}dz = \sigma^2$$

şeklinde elde edilir (Aytaç, 1999:275).

Binom, Poisson ve Normal dağılıma ilişkin yukarıda elde edilen sonuçlar Tablo-1'de özetlenmiştir.

Tablo-1: Binom, Poisson ve Normal Dağılımın Bazı Özellikleri

Dağılımın Adı	***Parametreleri***	***Olasılık Fonksiyonu***	***Ortalaması***	***Varyansı***
Binom	n, p	$p(x) = \binom{n}{x} p^x (1-p)^{n-x}$	np	npq
Poisson	λ	$P(x) = \frac{e^{-\lambda}\lambda^x}{x!}$	λ	λ
Normal	μ, σ	$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$	μ	σ^2

İKİNCİ BÖLÜM

GENELLEŞTİRİLMİŞ DOĞRUSAL MODELLER VE ÜSTEL AİLE KAVRAMI

Basit doğrusal regresyon modellerinden bahsederken, y bağımlı değişkeninin ve hataların normal dağılım gösterdiği varsayımı temel olarak kabul edilir. Fakat her zaman bağımlı değişken ve hata terimi normal dağılım göstermeyebilir.

Eğer anakütle regresyon modeli içindeki hatalar için olasılık dağılım fonksiyonu bir normal dağılım göstermiyorsa genelleştirilmiş doğrusal model kullanılır.

GLM temel olarak parametrik ve parametrik olmayan modellerin bir birleşimidir. Bir GLM aşağıda ifade edilen temel yapıya sahiptir

$$\eta_i = g(\mu_i) = \mathbf{x}_i^T \boldsymbol{\beta}$$

Burada $\mu_i \equiv E(Y_i)$; g bir link fonksiyonu; $\mathbf{x}_i^T$, $\mathbf{X}$ model matrisinin i. satırı; $\boldsymbol{\beta}$, bilinmeyen parametre vektörü ve η_i doğrusal kestirici (lineer predictor) olarak tanımlanır.

GLM iki varsayımı dikkate alır. Bunlardan ilki Y_i'ler bağımsızdır, ikincisi ise Y_i üstel aileden gelen bir dağılıma sahiptir. Yani Y_i normal, binom, poisson, üstel, gamma, ters Gauss tipi dağılım gibi üstel ailesinin üyesi olan bir dağılıma sahip olmalıdır (Montgomery-Peck-Vining).

2.1. Üstel Aile Kavramı

Üstel aile üyeleri olan dağılımlar aşağıdaki genel forma sahiptir:

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

a(.), b(.), c(.) keyfi fonksiyonlardır; ϕ, keyfi ölçek parametresi ve θ, doğal konum (natural location) parametresi olarak tanımlanır.

- Normal Dağılımın Üstel Ailenin Bir Üyesi Olduğunun Gösterilmesi:

$$f_\mu(y) = \frac{1}{\sigma\sqrt{2\pi}} exp\left[-\frac{(y-\mu)^2}{2\sigma^2}\right] = \left[\frac{-y^2+2y\mu-\mu^2}{2\sigma^2} - ln\,(\sigma\sqrt{2\pi})\right]$$

$$f_\mu(y) = exp\left[\frac{y\mu-\mu^2/2}{\sigma^2} - \frac{y^2}{2\sigma^2} - ln\,(\sigma\sqrt{2\pi})\right]$$

$\theta = \mu,$ $\quad b(\theta) = \theta^2/2 \equiv \mu^2/2,$ $\quad a(\phi) = \phi = \sigma^2$ ve

$$c(y,\phi) = -\frac{y^2}{(2\phi)} - ln\left(\sqrt{\phi 2\pi}\right) \equiv \frac{y^2}{(2\sigma^2)} - ln\,(\sigma\sqrt{2\pi})$$

olarak gösterilebildiği için normal dağılım üstel ailenin bir üyesidir denir.

- Binom Dağılımının Üstel Ailenin Bir Üyesi Olduğunun Gösterilmesi:

$$f_\theta(y) = \binom{n}{y} P^y (1-p)^{n-y}$$

$$f_\theta(y) = exp\left\{ln\left[\binom{n}{y} + lnP^y + ln(1-p)^{n-y}\right]\right\}$$

$$f_\theta(y) = exp\left\{ln\binom{n}{y} + ylnP + (n-y)ln\,(1-P)\right\}$$

$$f_\theta(y) = exp\left\{ln\binom{n}{y} + ylnP + n\,ln(1-P) - yln(1-P)\right\}$$

$$f_\theta(y) = exp\left\{y\big(lnP - ln(1-P)\big) + n\ln(1-P) + ln\binom{n}{y}\right\}$$

$$f_\theta(y) = exp\left\{yln\left(\frac{P}{1-P}\right) + n\ln(1-P) + ln\binom{n}{y}\right\}$$

$\theta = ln\left(\frac{P}{1-P}\right)$, $b(\theta) = -n\ln(1-P) = nln\left(\frac{1}{1-P}\right)$, $a(\phi) = \phi = \phi\omega = 1$,

$c(y,\phi) = ln\binom{n}{y}$

olarak gösterilebildiği için normal dağılım üstel ailenin bir üyesidir denir.

- Poisson Dağılımının Üstel Ailenin Bir Üyesi Olduğunun Gösterilmesi:

$$f_\mu(y) = \frac{\mu^y\, exp\,(-\mu)}{y\,!}$$

$$f_\mu(y) = exp\{yln\mu - \mu - ln\,(y!)\}$$

$\theta = ln\mu$, $b(\theta) = e^\theta$, $a(\phi) = \phi = \phi\omega = 1$, $c(y,\phi) = -ln\,(y!)$

olarak gösterilebildiği için normal dağılım üstel ailenin bir üyesidir denir (Montgomery-Vining-Myers, 2002:157:160).

2.2. Üstel Aile Dağılımlarının Özellikleri

Olasılık yoğunluk fonksiyonunun tanımından;

$$\int f_\theta(y)dy = 1$$

olduğu bilinmektedir. Bu eşitliğin her iki tarafının da θ'ya göre türevi alınırsa,

$$\frac{d}{d\theta}\int f_\theta(y)dy = 0$$

elde edilir. Türev alma işlemi de integral altında yapılırsa,

$$\int \frac{df_\theta(y)}{d\theta}dy = 0$$

eşitliği elde edilir. Benzer şekilde,

$$\frac{d^2}{d\theta^2}\int f_\theta(y)dy = 0$$

eşitliği elde edilir. Türev alma işlemi de integral altında yapılırsa da,

$$\int \frac{d^2 f_\theta(y)}{d\theta^2} dy = 0$$

Bu sonuçlar kullanılarak herhangi bir üstel aile rassal değişkeninin beklenen değeri ve varyansı elde ediliebilir.

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

eşitliğinden,

$$\frac{d}{d\theta} f_\theta(y) = \left\{\left(\frac{y - b'(\theta)}{a(\phi)}\right) f_\theta(y)\right\}$$

eşitliği,

$$\int \frac{d f_\theta(y)}{d\theta} dy = 0$$

eşitliğinden de,

$$\int \frac{d}{d\theta} f_\theta(y) = \int \left\{\left(\frac{y - b'(\theta)}{a(\phi)}\right)\right\} f_\theta(y) dy = 0$$

$$\int \left\{\frac{1}{a(\phi)} y f_\theta(y) - \frac{1}{a(\phi)} b'(\theta) f_\theta(y)\right\} dy = 0$$

$$\frac{1}{a(\phi)} \int y f_\theta(y) dy - \frac{1}{a(\phi)} b'(\theta) \int f_\theta(y)\, dy = 0$$

eşitliği elde edilir.

$$\int y f_\theta(y)\, dy = E(Y) \text{ ve } \int f_\theta(y)\, dy = 1$$

olduğundan,

$$\frac{1}{a(\phi)} E(Y) - \frac{1}{a(\phi)} b'(\theta) = 0$$

$$E(Y) = b'(\theta)$$

sonucu elde edilir.

Üstel aile rassal değişkeninin varyansını bulmak için de,

$$\int \frac{d}{d\theta} f_\theta(y) dy = 0$$

$$\frac{d}{d\theta} f_\theta(y) = \left\{\left(\frac{y - b'(\theta)}{a(\phi)}\right) f_\theta(y)\right\}$$

$$\frac{d^2}{d\theta^2} f_\theta(y) = \frac{-b''(\theta)}{a(\phi)} f_\theta(y) + \left\{\left(\frac{y - b'(\theta)}{a(\phi)}\right)^2\right\} f_\theta(y)$$

$$\int \frac{d^2}{d\theta^2} f_\theta(y) = \int \left\{\frac{-b''(\theta)}{a(\phi)} f_\theta(y) + \left(\frac{y - b'(\theta)}{a(\phi)}\right)^2 f_\theta(y)\right\} dy = 0$$

$$\int \frac{d^2}{d\theta^2} f_\theta(y) = \frac{-b''(\theta)}{a(\phi)} \int f_\theta(y)\, dy + \frac{1}{a(\phi)^2} \int (y - b'(\theta))^2 \int f_\theta(y)\, dy = 0$$

$$\int f_\theta(y)\, dy = 1 \text{ ve } var(Y) = \int (y - b'(\theta))^2 \int f_\theta(y)\, dy$$

olduğundan,

$$\frac{-b''(\theta)}{a(\phi)} + \frac{1}{a(\phi)^2} var(Y) = 0$$

$$var(Y) = b''(\theta)a(\phi)$$

sonucu elde edilir.

Sonuç olarak $E(Y) = b'(\theta) = \frac{db(\theta)}{d\theta}$ ve $var(Y) = b''(\theta)a(\phi) = \frac{d^2b(\theta)}{d\theta^2}a(\phi) = \frac{d\mu}{d\theta}a(\phi)$ olarakta ifade edilebilir.

$$var(\mu) = \frac{var(y)}{a(\phi)} = \frac{d\mu}{d\theta}$$

olsun.

Burada var(μ) ifadesi varyansın ortalama üzerindeki bağımlılığını gösterir. Bu özellik üstel ailenin üyesi tüm dağılımların karakteristik özelliğidir (Dobson, 2002:47).

- Normal Dağılımda $E(Y) = b'(\theta)$ ve $var(Y) = b''(\theta)a(\phi)$ Olduğunun Gösterilmesi

$$\theta = \mu$$

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

$$a(\phi) = \sigma^2$$

$$E(Y) = b'(\theta) = \frac{db(\theta)}{d\theta} = \mu$$

$$var(Y) = b''(\theta)a(\phi) = \frac{d^2b(\theta)}{d\theta^2}a(\phi) = \sigma^2$$

- Binom Dağılımda $E(Y) = b'(\theta)$ ve $var(Y) = b''(\theta)a(\phi)$ Olduğunun Gösterilmesi

$$\theta = ln\left(\frac{p}{1-p}\right)$$

$$e^\theta = \frac{p}{1-p}$$

$$P = e^\theta - pe^\theta$$

$$e^\theta = p + pe^\theta$$

$$e^\theta = p(1 + e^\theta)$$

$$p = \frac{e^\theta}{1+e^\theta}$$

$$b(\theta) = nln\left(\frac{1}{1-p}\right)$$

$$b(\theta) = nln\left(\frac{1}{1-\frac{e^\theta}{1+e^\theta}}\right)$$

$$b(\theta) = nln(e^\theta + 1)$$

$a(\phi) = \phi = \phi\omega = 1$

$E(Y) = b'(\theta) = n\frac{e^{\theta}}{1+e^{\theta}} = np$

$var(Y) = b''(\theta)a(\phi) = n\frac{e^{\theta}}{1+e^{\theta}} = \frac{ne^{\theta}(1+e^{\theta})-e^{\theta}e^{\theta}}{(1+e^{\theta})^2}$

$var(Y) = b''(\theta)a(\phi) = n\frac{e^{\theta}}{(1+e^{\theta})^2}$

$var(Y) = b''(\theta)a(\phi) = n\frac{e^{\theta}}{1+e^{\theta}}\frac{1}{1+e^{\theta}} = np(1-p)$

$var(Y) = b''(\theta)a(\phi) = n\frac{e^{\theta}}{1+e^{\theta}}\frac{1}{1+e^{\theta}} = npq$

- Poisson Dağılımda $E(Y) = b'(\theta)$ ve $var(Y) = b''(\theta)a(\phi)$ Olduğunun Gösterilmesi

$\theta = ln\mu$

$\mu = e^{\theta}$

$b(\theta) = \mu$

$a(\phi) = \phi = \phi\omega = 1$

$c(y, \phi) = -ln\ (y!)$

$E(Y) = \frac{db(\theta)}{d\theta} = \frac{db(\theta)}{d\mu}\frac{d\mu}{d\theta}$

$\frac{d\mu}{d\theta} = exp(\theta) = \mu$

$E(Y) = 1\mu = \mu$

$var(Y) = \frac{d\mu}{d\theta} = \frac{dE(y)}{d\theta} = \mu$

(Dobson, 2002,47).

Üstel aile üyesi olan Binom, Poisson ve Normal dağılıma ilişkin yukarıda elde edilen sonuçlar Tablo-2’de özetlenmiştir.

Tablo-2: Bazı Üstel Aile Dağılımlarının Karakteristikleri

	Normal	***Poisson***	***Binom***
Notasyon	$N(\mu,\sigma^2)$	$P(\mu)$	$ln\mu$
$\boldsymbol{f_y}$	$\frac{1}{\sigma\sqrt{2\pi}}exp\left[-\frac{(y-\mu)^2}{2\sigma^2}\right]$	$\frac{\mu^y exp\,(-\mu)}{y!}$	$\binom{n}{y}P^y(1-p)^{n-y}$
Aralık	$-\infty < y < \infty$	$y=0,1,2,\ldots$	$y=0,1,2,\ldots,n$
$\boldsymbol{\Theta}$	μ	$log\,(\mu)$	$ln\left(\frac{P}{1-P}\right)$
$\boldsymbol{b(\theta)}$	$\mu^2/2$	$exp(\theta)$	$\phi(=1)$
$\boldsymbol{\phi}$	σ^2	*1*	*1*
$\boldsymbol{a(\phi)}$	$\phi(=\sigma^2)$	$\phi(=1)$	$\phi(=1)$
$\boldsymbol{c(y,\phi)}$	$-\frac{y^2}{(2\phi)}-log(\sqrt{\phi 2\pi})$	$-log(y!)$	$log\binom{n}{y}$

2.3. Link Fonksiyonları ve Doğrusal Kestiriciler

Daha önceden de bahsedildiği gibi GLM Lojistik ve Poisson gibi doğrusal olmayan modeller ile doğrusal modelleri birleştiren bir yaklaşımdır. Bu birleşmeyi de bağımlı değişkenin beklenen değerinin uygun bir (link) fonksiyonu için doğrusal bir model geliştirerek sağlar. Bunun uygulanabilmesi için y_i bağımlı değişkeni mutlaka üstel ailenin bir üyesi olmalıdır.

Doğrusal kestirici (linear predictor) şu şekilde tanımlanır:

$$\eta_i = g[E(y_i)] = g(\mu_i) = \mathbf{x}_i'\boldsymbol{\beta} = \beta_o + \sum_1^k \beta_i x_i$$

Beklenen yanıt,

$$E(y_i) = g^{-1}(\eta_i) = g^{-1}(\mathbf{x}_i'\boldsymbol{\beta})$$

Burada ki g fonksiyonuna link fonksiyonu denir.

Aslında çoklu lineer regresyon modelinde $g(\mu_i)=\mu_i$ özel bir durumunda link fonksiyonu olarak identity link fonksiyonu kullanılır ve model aşağıdaki şekildedir:

$$\mu_i = \eta_i = \mathbf{x}_i'\boldsymbol{\beta} \qquad i = 1,2,\ldots,n$$

Link fonksiyonlarının çeşitli uygun seçimleri mevcuttur fakat $\eta_i = \theta$ olarak seçilirse μ_i kanonik link olur.

Normal, Poisson ve Binom dağılımlarının kanonik linkleri Tablo-3'de gösterilmiştir.

Tablo-3: Normal, Poisson ve Binom Dağılımları Kanonik Linkleri

Dağılım	Kanonik Link
Normal	$\mu_i = \eta_i$ (identity link)
Binom	$\eta_i = \ln\left(\frac{\pi_i}{1-\pi_i}\right)$ (lojistik link)
Poisson	$\eta_i = \ln(\lambda)$ (log link)

GLM ile kullanılan diğer link fonksiyonları ise şu şekildedir:

- Probit Link:

$$\eta_i = \Phi^{-1}[E(y_i)]$$

Burada Φ kümülatif standart normal dağılım fonksiyonudur.

- Tamamlayıcı (complementary) Log-Log Link:

$$\eta_i = ln\{ln[1-\mu_i]\}$$

- Güç Ailesi Linki:

$$\eta_i = \begin{cases} \mu_i{}^{\lambda}, & \lambda \neq 0 \\ ln[\mu_i], & \lambda = 0 \end{cases}$$

Bütün bunlardan anlaşılacağı üzere bağımlı değişken üzerindeki transformasyon ne kadar önemli ise link fonksiyonunun seçimi de o kadar önemlidir. Çünkü link fonksiyonunun yanlış seçimleri de GLM ile ilgili ciddi problemlere neden olabilir (Montgomery-Peck-Vining).

2.4. GLM'de Parametre Tahmini

GLM için parametre tahmini yaparken ençok-olabilirlik (maximum-likelihood) methodu kullanılır. Bu ençok-olabilirlik (maximum-likelihood) methodunun asıl uygulamasını ise IRLS (iteratively reweighted least squares) adı verilen bir algoritma gerçekleştirir.

IRLS (iteratively reweighted least squares) algoritmasının ürettiği regresyon katsayılarının son değeri $\hat{\beta}$ ise ve link fonksiyonun seçimi doğru ise asimptotik olarak aşağıdaki iafede söylenebilir:

$$E(\hat{\beta}) = \beta \ ve \ var(\hat{\beta}) = a(\phi)(X'V^{-1}X)^{-1}$$

Buradaki V matrisi doğrusal kestiricideki tahminlenen parametrelerin ($a(\phi)$ dışındaki) varyanslarından elde edilen diagonal bir matristir.

Her Y_i için log-olabilirlik fonksiyonunu

$$l_i = \left[\frac{y_i\theta_i - b_i(\theta_i)}{a(\phi)}\right] + c(y_i, \phi)$$

şeklinde ifade edersek;

$$E(Y_i) = \mu_i = b'(\theta_i), var(Y_i) = a(\phi)b''(\theta_i), g(\mu_i) = \mathbf{x}_i'\boldsymbol{\beta} = \eta_i$$

olur.

GLM için log-olabilirlik fonksiyonu;

$$L = logl(y,\beta) = \sum_{i=1}^{n}\left\{\left[\frac{y_i\theta_i - b_i(\theta_i)}{a(\phi)} + c(y_i, \phi)\right]\right\}$$

β_j parametresinin ençok-olabilirlik tahminini elde etmek için bahsedilen zincir türev alma kuralı uygulanırsa;

$$\frac{\partial l}{\partial \beta_j} = U_j = \sum_{i=1}^{n}\left[\frac{\partial l_i}{\partial \beta_j}\right] = \sum_{i=1}^{n}\left[\frac{\partial l_i}{\partial \theta_i}\frac{\partial \theta_i}{\partial \mu_i}\frac{\partial \mu_i}{\partial \beta_j}\right]$$

$$\frac{\partial l_i}{\partial \theta_i} = \left[\frac{y_i - b_i{}'(\theta_i)}{a(\phi)}\right] = \frac{[y_i - \mu_i]}{a(\phi)}$$

$$\frac{\partial \theta_i}{\partial \mu_i} = \frac{1}{\frac{\partial \mu_i}{\partial \theta_i}} = \frac{1}{b''(\theta_i)}$$

$$\frac{\partial \mu_i}{\partial \beta_j} = \frac{\partial \mu_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \beta_j} = \frac{\partial \mu_i}{\partial \eta_i}\mathbf{x}_i$$

Bu durumda denklem aşağıdaki şekli alır:

$$\frac{\partial l}{\partial \beta_j} = U_j = \sum_{i=1}^{n}\left[\frac{(y_i - \mu_i)}{a(\phi)}\frac{1}{b''(\theta_i)}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)\mathbf{x}_i\right]$$

Dikkate alınan linkin $\eta_i = \theta$ (kanonik link) olması nedeniyle $\frac{\partial \mu_i}{\partial \eta_i} = \frac{\partial \mu_i}{\partial \theta_i} = b''(\theta_i)$ olur ve denklem aşağıdaki şekli alır:

$$U_j = \sum_{i=1}^{n}\left[\frac{(y_i - \mu_i)}{a(\phi)}\frac{1}{b''(\theta_i)}b''(\theta_i)\boldsymbol{x}_i\right]$$

$$U_j = \frac{1}{a(\phi)}\sum_{i=1}^{n}(y_i - \mu_i)\,\boldsymbol{x}_i = 0$$

$a(\phi)$'nin de bir sabit olması nedeniyle denklem ;

$$\sum_{i=1}^{n}(y_i - \mu_i)\,\boldsymbol{x}_i = 0$$

şeklini alır. Matris notasyonu ile de aşağıdaki şekilde gösterilir.

$$X'(y - \mu) = 0$$

Bu denklemler sistemine skor denklemler denir.

β için bu denklem sistemi IRLS (iteratively reweighted least squares) adı verilen algoritma kullanılarak çözülür.

Dönüşüm bağımlı değişkendeki istenilen sabit varyans ve normallik özelliklerini sağlamakta başarısız ise GLM veri dönüştürmeye karşı en etkili alternatiftir. Yani GLM dönüşüm yapıldıktan sonra problem sabit varyanslı hale geldiğinde dönüşümleri kullanan standart analizlerden daha iyi sonuçlar verir (Dobson, 2002).

2.5. GLM'de Parametrelerin Önem Testi

Lojistik regresyon için tanımlanan çıkarsamalar GLM için de geçerlidir. Yani model sapması modelin uyumunu test etmek amacıyla ve tam model ile daraltılmış model arasındaki sapma (deviance) farkı modeldeki parametrelerin altkümeleri üzerindeki hipotezlerin test edilmesinde kullanılabilir.

- Olabilirlik Oran İstatistiği: Olabilirlik oran istatistiği modelin uyum iyiliğini ve modelleri karşılaştırmanın bir yolu olarak kullanılabilir. Burada tam veya doymuş model olarak adlandırılan model, tahmin edilebilecek maksimum parametreyi içeren daha genel bir modeldir. İlgilenilen model ise daraltılmış (unsaturated) model olarak adlandırılır. Olabilirlik oran istatistiği daraltılmış model ile bu doymuş modeli karşılaştırmaktadır. Doymuş model daraltılmış model ile aynı dağılıma ve link fonksiyonuna sahiptir.

Doymuş model için olabilirlik fonksiyonunu $\mathrm{L}(\mathrm{b}_{\max}, \mathrm{y})$ ve daraltılmış model için de olabilirlik fonksiyonunu $\mathrm{L}(\mathrm{b}, \mathrm{y})$ şeklinde ifade edersek olabilirlik oranı aşağıdaki şekli alır:

$$\gamma = \frac{L(b_{max}; y)}{L(b, y)}$$

Uygulamada olabilirlik oranının logaritması kullanılır:

$$log(\gamma) = l(b_{max}; y) - l(b, y)$$

$2log(\gamma)$ bir χ^2 dağılımına sahip olduğundan $log(\gamma)$ yerine $2log(\gamma)$ daha yaygın olarak kullanılmaktadır. Bu ifadeye ise sapma (deviance) adı verilemektedir.

Sapma (deviance) kavramı 1972 yılında Nelder ve Wedderburn tarafından bulunmuştur. Sapma normal hata doğrusal regresyon modelindeki artık kareler toplamının karşılığıdır (Dobson, 2002).

- Sapma (deviance): Bir modelin sapması (deviance), uyumu yapılan bir modelin log olabilirliği ile doymuş (saturated) bir modelin log-olabilirliğini karşılaştırır.

Log-olabilirlik oran istatistiğinde belirtildiği gibi aşağıdaki şekilde ifade edilir. (Dobson, 2002)

$$D = 2[l(b_{max}; y) - l(b, y)]$$

$$D = 2lnL(doymuş\ model) - 2lnL(\hat{\beta})$$

Örneklem genişliği n büyükse, p modelde yer alan parametre sayısını göstermek üzere model sapması $(n-p)$ serbestlik derecesi ile yaklaşık bir ki kare dağılımına sahiptir.

Model sapmasının büyük çıkması uyumu yapılan modelin doğru model olmadığını gösterir. Model sapmasının küçük çıkması ise uyumu yapılan modelin doymuş model kadar iyi olduğunu gösterir.

2.6. GLM'de Artıkların Analizi

GLM'de artıkların analizi, seçilen link fonksiyonun uygunluğu, varsayımların doğruluğu ve modelin yeterli olup olmadığı hakkında bilgi verici olduğundan oldukça önemlidir.

GLM'den elde edilen ham (raw) artıklar, bağımlı değişken için gerçek değerler ile kestirilmiş değerler arasındaki farka eşittir.

$$e_i = y_i - \hat{y}_i$$

Artıklar uç değerleri belirlemede, zayıf uyum gösteren gözlemleri kestirebilmekte, etkin gözlemleri tesbit etmede ve etkin gözlemleri seçebilmede kullanılabilirler.

Genelde GLM'de artık analizinin sapma (deviance) artıkları kullanılarak yapılması önerilir. i'inci sapma artığı aşağıdaki şekilde ifade edilir:

$$r_{D_i} = \sqrt{d_i} sign(y_i - \hat{y}_i)$$

Burada d_i, i'inci gözlemin sapmaya katkısını ifade eder.

Lojistik regresyon için şunu iafede edebiliriz:

$$d_i = y_i ln\left(\frac{y_i}{n_i\hat{\pi}_i}\right) + (n_i - y_i)\left[\frac{1-\left(\frac{y_i}{n_i}\right)}{1-\hat{\pi}_i}\right] \qquad i = 1, \dots, n$$

$$\hat{\pi}_i = \frac{1}{1+e^{-x_i'\beta}}$$

$\hat{\pi}_i \equiv \frac{y_i}{n_i}$ alındığında sapma (deviance) artıkları sıfıra yaklaşacağından daha düşük çıkacaktır.

Poisson regresyon için de şunu ifade edebiliriz:

$$d_i = y_i ln\left(\frac{y_i}{e^{x_i'\beta}}\right) - \left(y_i - e^{x_i'\beta}\right) \qquad i = 1, \dots, n$$

Yine bağımlı değişkenin gözlenen değeri y_i ile tahminlenen değeri $\hat{y}_i = e^{x_i'\beta}$ birbirine yaklaştıkça sapma (deviance) artıkları sıfıra yaklaşacağın için daha küçük çıkacaktır.

Sapma artıkları normal hata doğrusal regresyon modelindeki sıradan artıkların karşılığıdır. Bu yüzden normal hata olasılık ölçeği (normal probability scale) üzerindeki sapma artıklarının uyumu yapılan değerlere karşı çizimi uygun sonuçlar verecektir (Montgomery-Peck-Vining).

ÜÇÜNCÜ BÖLÜM

LOJİSTİK REGRESYON ANALİZİ

Değişkenlerin birbirleriyle olan ilişkilerinin incelenmesinde en çok kullanılan istatistiki tekniklerden birisi regresyon analizidir.

Doğrusal regresyon analizinde ilişkisi incelenen bağımlı ve bağımsız değişkenler seçilirken; X bağımsız değişkenlerinin ve hata teriminin normal dağılım göstermesi; Y bağımlı değişkeninin ise sürekli olması gerektiği varsayımları dikkate alınır ve model bu varsayımlar doğrultusunda kurulur. Bu varsayımlar doğrultusunda kurulan modele de parametrik teknikler uygulanır. Fakat bu durum her zaman böyle olmayabilir. Yani bağımlı değişken her zaman sürekli olmayabilir, bağımsız değişkenler ve hata terimi de her zaman normal dağılım göstermeyebilir. Bu gibi durumlarda da kurulan modele parametrik olmayan teknikler uygulanır.

Açıklamalardan da anlaşıldığı gibi istatistiksel bir analiz yaparken öncelikli olarak ilişkisi incelenen bağımlı ve bağımsız değişkenlerin yapısına dikkat edilir. Çünkü hangi istatistiksel tekniğin uygulanması gerektiğine karar vermek için verilerin yapısının iyi bilinmesi gerekir (Işığıçok ve Murat, 2007).

Tablo-4 : İstatistiksel Tekniğin Seçimi

		BAĞIMLI DEĞİŞKEN (Y)	
		NİTEL	**NİCEL**
BAĞIMSIZ DEĞİŞKEN (X)	**NİTEL**	Oran testleri Ki-kare testi	t testi,z testi,F testi ANOVA,DOE Basit regresyon
	NİCEL	Diskriminant analizi Lojistik regresyon	Korelasyon Çoklu regresyon

Lojistik regresyon analizi bağımlı değişkenin kategorik, bağımsız değişkenlerin ise kategorik veya sürekli olduğu durumlarda bağımlı değişken ile bağımsız değişken arasındaki ilişkiyi açıklamaya yarayan bir yöntemdir. Bağımlı değişkenin kategorik olduğu durumlarda

Lojistik regresyon analizi dışında Diskriminant analizi (Ünsal, 2005'de yaptığı çalışmasında gruplar arası farklılıkları belirlemede diskriminant analizini kullanmıştır), Probit analizi (Gür, 1995'de yaptığı çalışmasında istatistik kuramının normal dağılıma dönüştürme yöntemlerinden probit analizini uygulamıştır) ve Logaritmik doğrusal regresyon da verileri analiz etmede kullanılan yöntemlerdendir (Oğuzlar, 2005).

Logaritmik doğrusal regresyon tüm bağımsız değişkenlerin kategorik olması gerektiğini varsayarken, Diskriminant analizi tüm bağımsız değişkenlerin sürekli olmasını, kütlelerin ortak varyans kovaryans matrisine sahip olmasını ve çoklu normal dağılım göstermesi gerektiğini varsayar. Diskriminant analizi lojistik regresyon analizine göre daha çok varsayımı gerektirir. Lojistik regresyon analizi kategorik ve sayısal bağımsız değişkenlerin varlığı durumunda daha az varsayım gerektirdiğinden Diskriminant analizi ve Çapraz tablo uygulamalarına alternatif olarak kullanılmaktadır. Zaten bağımlı değişkeninin bir olasılık ifade etmesinden ve binomial dağılım göstermesinden ötürü bu durumlarda Diskriminant analizi kullanmak sakıncalar doğuracaktır. Bununla beraber Diskriminant analizinin varsayımlarının sağlandığı durumlarda Lojistik regresyon analizi de uygulanabilir (Akgül ve Çevik, 2003'den aktaran Oğuzlar, 2005).

Öte yandan Lojistik regresyon analizi bağımlı değişkenin ölçüldüğü ölçek türüne ve bağımlı değişkenin gösterdiği kategorik yapıya göre üçe ayrılmaktadır. Bağımlı değişkenin iki şıklı olması durumunda İkili (binary) Lojistik Regresyon analizi, bağımlı değişkeninin sınıflayıcı (nominal) ölçme düzeyine sahip olduğu ve en az üç şıklı olduğu durumlarda Sınıflayıcı Lojistik regresyon analizi ve bağımlı değişkeninin sıralayıcı (ordinal) ölçme düzeyine sahip olduğu ve yine en az üç şıklı olduğu durumlarda ise Sıralayıcı Lojistik Regresyon analizi kullanılır (Işığıçok ve Murat, 2007). Bu tez çalışmasında bağımlı değişkeninin sadece iki mümkün değer (0 ve 1) alabildiği İkili (binary) Lojistik Regresyon analizi kullanılacaktır. Bağımlı değişken bir birey sigara içiyor olabilir veya içmiyor olabilir ya da bir genç üniversite mezunu olabilir veya olmayabilir şeklinde seçilerek oluşturulabilir. Bu durumda sonuç ya bir başarı ya da bir başarısızlık şeklindedir. Genellikle üzerinde durulan olayın gerçekleşme olasılığı 1, gerçekleşmeme olasılığı da 0 şeklinde kodlanır.

Temelde Lojistik regresyon analizinde de amaç doğrusal regresyon analizinde olduğu gibi bağımlı değişken ile bağımsız değişkenler arasındaki neden sonuç ilişkisini en iyi şekilde açıklayan en sade modeli elde etmektir. Ayrıca regresyon modelinin oluşturulması ve katsayıların yorumlanması açısından da doğrusal regresyon analiziyle benzerlik göstermektedir. Ancak bu iki yöntem arasında üç önemli fark vardır:

- Doğrusal regresyon analizinde bağımlı değişken sürekli bir yapıya sahip iken, Lojistik regresyon analizinde bağımlı değişken kategorik bir yapıya sahiptir.
- Doğrusal regresyon analizinde bağımlı değişkenin alacağı değer tahmin edilirken, Lojistik regresyon analizinde bağımlı değişkeninin alabileceği değerlerden birinin gerçekleşme olasılığı tahmin edilir.
- Doğrusal regresyon analizinde bağımsız değişkenlerin çoklu normal dağılım göstermesi önkoşulu aranırken, Lojistik regresyon analizinde bağımsız değişkenlerin dağılımına ilişkin herhangi bir önkoşul yoktur (Aktaş, 2009).

Lojistik regresyon analizinin diğer bir amacı ise gözlemleri sınıflandırmadır. Bu sınıflandırmayı yapmak için üç analizden faydalanılır:

- Kümeleme analizi
- Diskriminant analizi
- Lojistik regresyon analizi

Kümeleme analizinde verilerin yapısındaki grup sayısı bilinmemektedir ve gözlemler benzerlik ya da uzaklık ölçütlerine göre sınıflandırılmaktadır. Burada amaç, yalnızca gözlemlerin oluşturmuş olduğu kümenin yapısını belirlemektir. Diskriminant ve Lojistik regresyon analizinde ise verilerin yapısındaki grup sayısı bilinmektedir ve bu veriler kullanılarak elde edilen ayrımsama modeli sayesinde veri kümesine yeni alınan gözlemlerin gruplara atanması yapılmaktadır (Başarır, 1990:1'den aktaran Aktaş, 2009).

Lojistik Regresyon analizi son yıllarda biyoloji, tıp, ekonomi, tarım, veterinerlik ve taşıma alanlarında yaygın olarak kullanılmaktadır. Lojistik regresyon analizi 1960'ların sonunda 1970'lerin başında bir alternatif olarak önerildi ve rutin olarak 1980'lerin başında istatistiksel paket programları içinde yer aldı (Peng-Lee-İngersoll, 2002). O günden beri de Lojistik Regresyon analizinin kullanımı sosyal bilimlerde ve eğitimsel araştırmalarda artmıştır.

Lojistik regresyon modelleriyle ilgili Türkiye'de de birçok çalışma yapılmıştır. Bu çalışmalardan bazıları şunlardır:

Aktaş (2009), öğrencilerin sigara içme alışkanlıklarını etkileyen faktörleri belirlemeyi hedefleyen bir çalışma yapmıştır ve bu amaçla da sigara içmeyi etkileyen faktörleri belirlemek için lojistik regresyon analizi ve diskriminant analizini kullanmıştır. SPSS paket programını kullanarak ileri doğru değişken seçme tekniğiyle lojistik regresyon analizini uygulamış ve sonuçta sigara kullanma durumunu etkileyen faktörleri yaş, barınma şekli, babanın sigara içme durumu, alkol kullanma durumu, arkadaş çevresinin sigara kullanma durumu, sigaranın sıkıntı stres ve yanlızlığı giderdiğini ve statü kazandırdığını düşünmesi olarak belirlemiştir. Lojistik regresyon modeliyle elde edilen sonuçların, gözlemlerin varolan gruplardan birine atanması için uygulanan değişken seçme tekniğiyle diskriminant analizi uygulanmış. Sonuç olarak lojistik regresyon modeli için doğru sınıflama yüzdesi oldukça yüksek olduğundan ve ki-kare testine göre de model anlamlı olduğundan lojistik regresyon için belirlenen model en uygun ayrımsama modeli olarak kabul edilmiş. Lojistik regresyon analizi sonuçlarına göre öğrencilerin sigara içmesini etkileyen en önemli faktörün ‘sigaranın statü kazandırdığına inanılması’ olarak belirlenmiş. Bunu sırasıyla ‘sigaranın sıkıntı, stres ve yalnızlığı giderdiğinin düşünülmesi’, ‘arkadaş çevresinin sigara kullanması’, ‘kendisinin alkol kullanması’, ‘babanın sigara içmesi’ ve ‘barınma şeklinin’ izlediği görülmüş. Yaş değişkenin odds oranının değeri ise yaş ilerledikçe sigara içmede azalma olduğunu göstermiştir.

Oğuzlar (2005), Bursa Emniyet Müdürlüğü ile gerçekleştirilen ortak proje kapsamında, ahlak, yankesicilik ve narkotik bürolarına ilişkin verilerden yararlanarak suçluların profilini belirlenmeyi hedefleyen bir çalışma yapmış ve bu amaçla da lojistik regresyon analizini kullanmıştır. Çalışmada analizi yapılacak olan büro 1 diğerleri 0 şeklinde kodlanmış. Bağımlı değişkeni etkileyen bağımsız değişkenler ‘suçun işlendiği olay saati’, ‘suçu işleyen bireyin cinsiyeti’, ‘suçu işleyen bireyin yaşı’, ‘suçu işleyen bireyin doğum yeri’, ‘suçu işleyen bireyin öğenim durumu’, ‘suçu işleyen bireyin mesleği’ olarak belirlenmiş. Ahlak bürosuna ilişkin SPSS’de geriye doğru eleme tekniğiyle belirlenen modelde tüm bağımsız değişkenlerin önemli bulunduğu görülmüş ve tümü modele dahil edilmiş. Ayrıca modele uygulanan Hosmer-Lemeshow uyumun iyiliği ölçüsünün anlamlı çıktığı görülmüş. Yankesicilik bürosuna ilişkin SPSS’de geriye doğru eleme tekniğiyle belirlenmiş modelde yine tüm değişkenlerin önemli bulunduğu ve modele dahil edildiği görülmüş. Yine burada da Hosmer-Lemeshow modelin uyumun iyiliği ölçüsü anlamlı bulunmuş. Narkotik bürosuna ilişkin SPSS’de geriye doğru eleme tekniğiyle belirlenmiş modelde cinsiyet değişkeninin modelden dışlandığını geriye kalan değişkenlerle oluşan modelin uyum iyiliği ölçüsü olan Hosmer-Lemeshow test sonucunun anlamlı çıktığını görülmüş. Lojistik regresyon analizi sonuçlarına göre, ahlak bürosuna ilişkin odds oranları değerlendirildiğinde ahlak suçu işleyenlerin en belirgin özelliği olarak öğrenim durumu öne çıkmış. Özellikle okur yazar ve ilkokul mezunu olanların bu suçu işlemeye daha meyilli olduğu görülmüş. Öne çıkan diğer bir grupta emekli olanlardır. Yankesicilik bürosuna ilişkin odds oranları değerlendirildiğinde öğrenim durumunun yine ön plana çıktığı görülmüş. Özellikle okur yazar ve ilkokul mezunu olanların bu suçu işlemeye daha meyilli olduğu görülmüş. Meslek değişkenlerinin kategorilerine bakıldığında ise ev hanımı olanların bu suçu işlemeye meyilli olduğu görülmüş.

Bir diğer göze çarpan farklılık ise bölgeler içinde sadece Doğu Anadolu Bölgesi doğumlu olanların bahis oranı 1' den büyük çıkmış. Yani Doğu Anadolu Bölgesi doğumlu olanların bu suçu işlemeye daha meyilli oldukları görülmüş. Narkotik bürosuna ilişkin odds oranları değerlendirildiğinde, yine burada öğrenim durumunun öne çıktığı ve bu suçu işlemede okur yazar olmayanların ve ilkokul mezunu olanların oranının daha yüksek olduğu görülmüş. Doğum yeri değişkeninin kategorilerine bakıldığında da Akdeniz Bölgesine ait odds oranının en yüksek değeri aldığı görülmüş. Yani Akdeniz Bölgesi doğumlu olanların bu suçu işlemeye daha meyilli oldukları söylenebilir.

Murat ve Işığıçok (2007), Mart 2007 döneminde yapılan cumhurbaşkanlığı ve genel seçimler öncesi Bursa halkının siyaset ve ekonomi hakkındaki görüş ve düşüncelerini ortaya koymak amacıyla bu çalışmayı gerçekleştirmişlerdir ve bu amaçla da lojistik regresyon analizini kullanmışlardır. Araştırmanın asıl konusunu oluşturan hükümete güvenme durumu (1=güveniyorum 0= güvenmiyorum) ve mevcut ekonomik durumun tatmin ediciliği (1=evet 0=hayır) şeklinde kodlanmış. Lojistik regresyon modelini oluşturan yanıt değişkeni hükümete güvenme durumu ve mevcut ekonomik durumun tatmin ediciliği şeklinde ayrı ayrı modellenmiş ve sonuçta iki ayrı lojistik regresyon modeli elde edilmiş. Kurulan iki ayrı lojistik regresyon modelinin ilki olan mevcut olan hükümete güven durumunda; yaş, mevcut hükümetin ekonomik durum performansı, siyasi durum performansı ve dış ilişkiler performansının önemli faktörler olduğu görülmüş, ikinci model olan mevcut ekonomik durumun tatmin ediciliğinde ise; gelir düzeyi, mevcut hükümetin ekonomik durum performansı, siyasi durum performansının önemli faktörler olduğu görülmüş.

Girginer ve Cankuş (2008), bu çalışmada iki ayrı üniversiteye sahip olan Eskişehir'deki üniversite öğrencilerinin tramvay yolcu memnuniyetini ölçmeyi hedeflemişlerdir ve bu amaçla da lojistik regresyon analizini kullanmışlardır. Çalışmada bağımlı değişken (memnun olanlar = 1 memnun olmayanlar=0) iki şıklı olarak kodlanmış. Lojistik regresyon analizi sonucunda öğrencilerin bir çok açıdan Estram'dan şikayetçi olmalarına rağmen ulaşım kolaylığı sağladığından ve taşıt trafiğinde önceliği bulunduğundan Estram'ı kullandıkları ortaya çıkmıştır. Ünsal ve Güler (2005), çalışmalarında Türk Bankacılık Sektöründeki bankaları sınıflandırmak ve bankaların mali durumlarını öngörmek amacıyla 1997-2003 yılları arasında Türkiye'de faaliyet gösteren ticari bankların mali durumları lojistik regresyon analizi ve diskriminant analizi tekniklerini kullanarak incelemiş ve ilgili veri için hangi tekniğin daha iyi sonuçlar verdiğini belirlemişlerdir. Sonuç olarak diskriminant analizinde de lojistik regresyon analizinde de öngörü aşamasında değişken seçiminde hangi yöntemin kullanılacağına dair kesin bir şey söylenemese de yıl için sınıflandırmalarda tüm değişkenlerle elde edilen fonksiyonların daha iyi sınıflandırma yaptığı görülmüş. Genel olarak değerlendirme yapıldığında ise lojistik regresyon analizinin bankaları sınıflandırmada ve mali durumlarını öngörmede daha başarılı olduğu belirlenmiş.

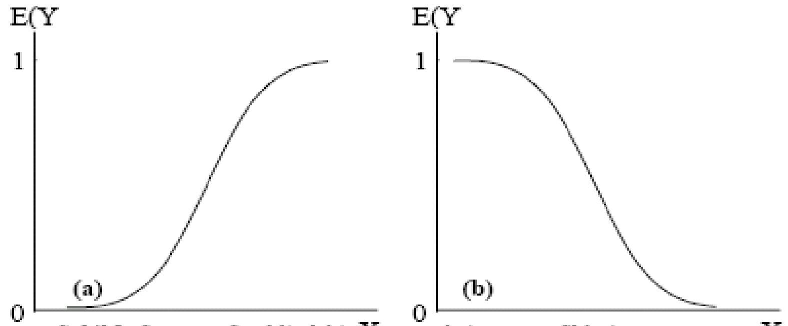
Bircan (2004), bu çalışmasında çocuklarda doğum ağırlığını etkileyen önemli risk faktörlerini belirlemeyi hedeflemiş ve amaçla da lojistik regresyon analizini kullanmıştır. Bağımlı değişken doğum ağırlığı 2,5 kg eşit ve küçük olan bebekler = 1, doğum ağırlığı 2,5 kg üzerinde olan bebekler = 0 şeklinde kategorik olarak kodlanmıştır. Bağımlı değişkeniyle anlamlı ilişki olan cinsiyet, boy, gebelik haftası, gebelik öncesi anne kilosu ve beslenme şekli alınarak uygun model oluşturulmuş.

Ayan ve Kocacık (2009), bu çalışmalarında aile içinde çocuğa uygulanan şiddetin ailenin sosyal, ekonomik, psikolojik, kültürel özelliklerine ne derecede bağlı olduğunu belirlemek amacıyla lojistik regresyon analizi yöntemine başvurmuşlardır. SPSS paket programı kullanılarak ileri doğru değişken seçme tekniğiyle lojistik regresyon analizi uygulanmış ve sonuçta öğrencinin annesi tarafından şiddete maruz kalmasına istatistiksel olarak anlamlı düzeyde etkisi olan değişkenler sırasıyla annenin eğitim durumu, anne baba arasında şiddetin var olma durumu ve annenin çocuğa davranış biçimi olarak belirlenmiş. Öğrencinin babası tarafından şiddete maruz kalmasına istatistiksel olarak anlamlı düzeyde etkisi olan değişkenler ise sırasıyla öğrencinin cinsiyeti, anne baba arasında şiddetin var olma durumu ve babanın çocuğa davranış biçimi olarak belirlenmiş.

Herken ve arkadaşları (2000) bu çalışmalarında orta ve yüksek öğrenim gençlerinin alkol kullanım sıklığı ile alkol kullanımının sosyo-demografik özellikler ve sosyal öğrenme arasındaki ilişkiyi ortaya çıkarmak istemiş ve bu amaçla da lojistik regresyon analizi yöntemine başvurmuşlardır. SPSS paket programı kullanılarak ileri doğru değişken seçme tekniğiyle lojistik regresyon analizi tekniği uygulanmış ve bu analiz sonucunda gençlerin alkol kullanımına istatistiksel olarak anlamlı düzeyde etkisi olan değişkenler babanın, örnek alınan öğrencinin, abisinin ya da ablasının alkol kullanması, gencin yaşı ve alkolü kötü ve zararlı olarak algılama olarak belirlenmiş.

3.1. Lojistik Regresyon Modelleri

Hem teorik hem de deneysel incelemeler sonucunda görülmüştür ki bağımlı değişken binary iken yanıt fonksiyonunun şekli S veya ters S şeklindedir (Bkz: Şekil 1) (Bircan Hüdaverdi, 2004). Fonksiyonun şeklinin S veya ters S şeklinde olması β_1'in işaretine bağlıdır.



Şekil 1: S ve ters S şeklindeki cevap fonksiyonu grafikleri

Şekil 1'de görülen yanıt fonksiyonları lojistik yanıt fonksiyonları olarak bilinir. Bu yanıt fonksiyonları 0 ve 1 değerinde X ve Y eksenlerine asimptottur ve bitiş noktaları dışında doğrusaldır. S ya da ters S şeklinde nitelendirilen bu iki paralel çizgisini iki nedenden dolayı lineer eşitlikle tarif etmek zordur. Birincisi uç noktalar bir lineer trendi takip edemez, ikincisi ise hatalar ne normal dağılır ne de tüm veri aralığınca sabittir. Lojistik regresyon, bağımlı değişkene logit transformasyonu uygulayarak bu sorunu çözer.

Lojistik regresyon ortaya çıkacak riski 0 ile 1 arasında herhangi bir değer olarak tahmin eder. Başka bir deyişle risk 0'ın altında ve 1'in üzerinde olamaz ancak bu aralık içerisinde yer alır. Lojistik regresyonun bu değişim aralığında olması tercih edilmesindeki nedenlerden biridir. Fakat bu durum her model için her zaman doğru olmamaktadır (Hosmer Lemeshow, 1980, 1043-1069'dan aktaran Bircan, 2004).

Lojistik regresyonun temelini oluşturan ana matematiksel kavram bir odds oranının doğal logaritması olan logittir. Yani lojistik model, X'den Y'nin logitini tahmin eder. Bu modelin en basit örneği 2×2 lik kontenjans tablosundan türetilmiştir (Peng-Lee-İngersoll, 2002).

k bağımsız değişken ve N gözlem olduğunda doğrusal regresyon modelinin genel formu i. gözlem için,

$$y_i = \beta_0 + \beta_1 x_{1i} + \cdots\cdots\cdots + \beta_k x_{ki} + \varepsilon_i$$

şeklindedir.

Örneklem büyüklüğü n olduğunda ise doğrusal regresyon modeli,

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \cdots\cdots\cdots + \hat{\beta}_k x_{ki} + e_i$$

şeklinde ifade edilir (Aktaş Cengiz, 2009).

Bağımlı değişkenin alabileceği değerlerin 0 ve 1 arasında olmasını sağlayan eğrisel ilişkiyi veren model aşağıdaki şekilde ifade edilir:

$$y_i = \mathbf{x}_i' \beta + \varepsilon_i$$

Burada $\mathbf{x}_i' = [1, x_{i1} x_{i2}, \ldots, x_{ik}]$, $\beta' = [\beta_0, \beta_1, \ldots, \beta_k]$'dır ve bağımlı değişken y_i ya 0 ya da 1 değerini almaktadır. y_i bağımlı değişkenin aşağıda verilen olasılık dağılımı ile bir bernoulli rassal değişkeni olduğunu varsayacağız (Montgomery-Peck-Vining, 1992:443).

Y_i	Olasılık
1	$P(y_i = 1) = \pi_j$
0	$P(y_i = 0) = 1 - \pi_i$

$E(\varepsilon_i) = 0$ olduğu için bağımlı değişkeninin beklenen değeri,

$E(y_i) = 1(\pi_i) + 0(1 - \pi_i)$

$E(y_i) = \pi_i$

Bu şunu ifade eder:

$E(y_i) = E[\mathbf{x}_i'\beta + \varepsilon_i]$

$E(y_i) = \mathbf{x}_i'\beta$

$E(y_i) = \pi_i$

Yani yanıt fonksiyonu $E(y_i) = \mathbf{x}_i'\beta$ ile verilen beklenen yanıt, sadece yanıt değişkeni 1 değerini alan bir olasılıktır. Özetle lojistik regresyonda sonuçlar π_i'ye eşit olduğu için bir olasılıktır.

$y_i = \mathbf{x}_i'\beta + \varepsilon_i$ denklemi bazı problem içerir:

- Yanıt binary ise ε_i hata terimi sadece iki değer alabilir.

$\varepsilon_i = 1 - \mathbf{x}_i'\beta \qquad y_i = 1 \ ise$

$\varepsilon_i = -\mathbf{x}_i'\beta \qquad y_i = 1 \ ise$

Bu nedenle bu modeldeki hatanın normal dağılım göstermesi mümkün değildir.

- Hata varyansı sabit değildir. Çünkü;

$\sigma_{y_i}^2 = E\{y_i - E(y_i)\}^2$

$\sigma_{y_i}^2 = (1 - \pi_i)^2 \pi_i + (0 - \pi_i)^2 (1 - \pi_i)$

$\sigma_{y_i}^2 = \pi_i (1 - \pi_i)$

$E(y_i) = \mathbf{x}_i'\beta = \pi_i$ olduğu için son ifade,

$\sigma_{y_i}^2 = E(y_i)(1 - E(y_i))$

şeklinde ifade edilir. Yani gözlemlerin varyansı (bu hataların varyansı ile aynıdır çünkü $\varepsilon i = y_i - \pi_i$ 'dir ve π_i bir sabittir) ortalamanın bir fonksiyonudur.

- Yanıt fonksiyonu üzerinde bir kısıtlama mevcuttur.

$0 \leq E(y_i) = \pi_i \leq 1$

Çünkü π_i bir olasılık olduğu için 0 ve 1 aralığında yer alır (Montgomery-Peck-Vining, 1992:443).

Bu bilgiler doğrultusunda β_1'in işaretine göre S veya ters S (bakınız şekil 1) şeklinde olan eğrileri sağlayan lojistik fonksiyon aşağıdaki formülle ifade edilir:

$$E(y_i) = \pi_i = \frac{exp\,(\mathbf{x}_i'\beta)}{1+exp\,(\mathbf{x}_i'\beta)} \quad \text{veya} \quad E(y_i) = \pi_i = \frac{exp(\beta_0+\beta_1 x_{1i}+\cdots+\beta_k x_{ki})}{1+exp(\beta_0+\beta_1 x_{1i}+\cdots+\beta_k x_{ki})}$$

Eşdeğer olarakta;

$$E(y_i) = \pi_i = \frac{1}{1+exp\,(-\mathbf{x}_i'\beta)}$$

şeklinde tanımlanır.

Lojistik fonksiyon kolaylıkla doğrusallaştırılabilir. Lojistik regresyon analizinde ilgilenilen durumun (P) olma olasılığının diğer durumun $(1-P)$ olma olasılığına oranın logaritması açıklayıcı değişkenlere doğrusal olarak bağlanmaktadır.

η doğrusal kestirici olsun. Burada η,

$\eta = ln\frac{\pi_i}{1-\pi_i}$ transformasyonu ile tanımlanır. İspatı ise aşağıdaki şekilde gösterilir:

$$E(y_i) = \pi_i = \frac{1}{1+exp\,(-\mathbf{x}_i'\beta)}$$

$$\eta = \mathbf{x}_i'\beta$$

$$\pi_i = \frac{1}{1+e^{-\eta}}$$

$$1 + e^{-\eta} = \frac{1}{\pi_i}$$

$$e^{-\eta} = \frac{1-\pi_i}{\pi_i}$$

$$-\eta = ln\frac{1-\pi_i}{\pi_i}$$

$$\eta = -ln\frac{1-\pi_i}{\pi_i}$$

$$\eta = ln\frac{\pi_i}{1-\pi_i} = ln\left(\frac{P(Y=1|X)}{P(Y=0|X)}\right) = \beta_0 + \beta_1 X = X'\beta$$

Bu trasformasyonlar genellikle π_i olasılığının logit[1] transformasyonu olarak adlandırılır ve $\frac{\pi_i}{1-\pi_i}$ oranına da *odds* oranı denir (Montgomery-Peck-Vining, 1992:443). İki

[1] Ln odds dönüşümü ise lojit olarak adlandırılır.

şıklı bağımlı değişkenin iki kategorisinin görülme olasılıklarının birbirine oranlanmasına bahis ya da odds adı verilir. Örneğin ilgilendiğimiz türden bir olayın olma olasılığı (P) ise, diğer olayın olma olasılığı $(1-P)$ olacaktır. Odds oranı ise iki odds'un oranlanması ile bulunur. Bağımlı değişken 0 ve 1 değerleri verilerek kodlanırsa, π_i bağımlı değişkenin 1 değerini alma olasılığını, $1-\pi_i$ de bağımlı değişkenin 0 değerini alma olasılığını göstermektedir.

Lojistik fonksiyona benzer olan başka fonksiyonlarda vardır. Bunlar da π transforme edilerek elde edilir. Bunlardan biri de probit transformasyonudur. Bu bir probit regresyon modelini oluşturur. Probit regresyon modeli logit regresyon modelinden daha az esnektir ve bu yüzden de lojistik regresyona göre kullanım alanı daha dardır.

3.2. Lojistik Regresyon Modellerinde Parametre Tahmini

Bağımlı değişkeni kategorik olan lojistik regresyon modellerinde parametreleri tahmin etmekte kullanılan çeşitli yöntemler mevcuttur. Diskriminant analizi, Ağırlıklı En Küçük Kareler ve En Çok (maksimum likelihood) Olabilirlik yöntemi bunlardan bazılarıdır. Bu yöntemlerin arasında en bilineni ve en sık kullanılanı ise En Çok Olabilirlik (maksimum likelihood) yöntemidir.

En çok olabilirlik (maksimum likelihood) tekniğinin kuralı, p tane açıklayıcı değişkene ilişkin β kestirimini Ybağımlı değişkenin gözlenme olasılığını mümkün olduğunca büyük kılacak şekilde bulmaktır. Yani En Çok Olabilirlik (maksimum likelihood) yöntemi, gözlemlenmiş veri kümesinden elde edilmenin olasılığını en büyük yapacak bilinmeyen parametrelerin değerlerini verir (Aktaş, 2009).

Lojistik regresyon modelleri genel olarak,

$$y_i = E(y_i) + \epsilon_i$$

şeklinde ifade edilir.

Burada $\mathrm{y_i}$ gözlemleri,

$E(y_i) = \pi_i = \frac{exp\ (\mathbf{x}_i'\beta)}{1+exp\ (\mathbf{x}_i'\beta)}$ beklenen değeri ile bağımsız bernoulli rassal değişkenleridir.

$\mathbf{x}_i'\beta$ doğrusal kestiricisindeki parametreleri tahmin etmek için çoğunlukla en çok olabilirlik (maximum likelihood) yöntemi tercih edilir ve bu amaçlada ilk olarak olabilirlik foksiyonu olarak isimlendirilen bir fonksiyon kurulur.

Her bir y_i gözlemi bir bernoulli tesadüfü değişkenidir. Bu nedenle her bir örnek gözlemin olasılık dağılımı da,

$f_i(y_i) = \pi_i^{y_i}(1-\pi_i)^{1-y_i}$ *i=1,2,.....,n*

Her bir gözlem değeri 0 ya da 1 değerini alır. $\mathrm{y_i}$ gözlemleri bağımsızdır ve en çok olabilirlik (maximum likelihood) fonksiyonu aşağıdaki gibidir:

$$L(\mathrm{y_1, y_2, \ldots, y_n}; \beta) = \prod_{i=1}^{n} f_i(y_i)$$

$$L(\mathrm{y_1, y_2, \ldots, y_n}; \beta) = \prod_{i=1}^{n} \pi_i^{y_i}(1-\pi_i)^{1-y_i}$$

şeklindedir.

En çok olabilirlik (maximum likelihood) yönteminde, olabilirlik fonksiyonu yerine olabilirlik fonksiyonunun logaritmasını en büyük yapmak daha kolay olacağından bu işlemi yapmak daha uygundur. Bu durumda olabilirlik fonksiyonunun logaritması:

$$lnL(\mathrm{y_1, y_2, \ldots, y_n}; \beta) = ln\prod_{i=1}^{n} f_i(y_i)$$

$$lnL(\mathrm{y_1, y_2, \ldots, y_n}; \beta) = ln\prod_{i=1}^{n} \pi_i^{y_i}(1-\pi_i)^{1-y_i}$$

$$lnL(\mathrm{y_1, y_2, \ldots, y_n}; \beta) = \sum_{i=1}^{n} ln\,\pi_i^{y_i}(1-\pi_i)^{1-y_i}$$

$$\mathrm{ln}L(\mathrm{y_1, y_2, \ldots, y_n}; \beta) = \sum_{i=1}^{n}[ln\,\pi_i^{y_i} + ln(1-\pi_i)^{1-y_i}]$$

$$\mathrm{ln}L(\mathrm{y_1, y_2, \ldots, y_n}; \beta) = \sum_{i=1}^{n}[y_i ln\,\pi_i + (1-y_i)ln(1-\pi_i)]$$

$$lnL(\mathrm{y_1, y_2, \ldots, y_n}; \beta) = \sum_{i=1}^{n}[y_i ln\pi_i - y_i\,ln(1-\pi_i) + ln(1-\pi_i)]$$

$$lnL(\mathrm{y_1, y_2, \ldots, y_n}; \beta) = \sum_{i=1}^{n} y_i(\,ln\,\pi_i - ln(1-\pi_i)\,) + \sum_{i=1}^{n} ln(1-\pi_i)$$

$$lnL(\mathrm{y_1, y_2, \ldots, y_n}; \beta) = \sum_{i=1}^{n}[y_i\,ln\frac{\pi_i}{1-\pi_i}] + \sum_{i=1}^{n} ln(1-\pi_i)$$

$(1-\pi_i) = [1 + exp(\mathbf{x}_\mathbf{i}'\beta)]^{-1}$ ve $\eta_i = ln\frac{\pi_i}{1-\pi_i} = \mathbf{x}_\mathbf{i}'\beta$

olduğu için log-olabilirliği aşağıdaki gibi yazabiliriz;

$$lnL(\mathbf{y}; \boldsymbol{\beta}) = \sum_{i=1}^{n} y_i\,\mathbf{x}_\mathbf{i}'\beta - \sum_{i=1}^{n} ln[1 + exp(\mathbf{x}_\mathbf{i}'\beta)]$$

$L(\beta)$‘yı maksimum yapan β değerlerini bulmak için, $L(\beta)$’nın β_0 ve β_1’e göre türevleri alınarak sıfıra eşitlenir. Elde edilen eşitlikler:

$$\sum_{i=1}^{n}[y_i - \pi_i] = 0$$

$$\sum_{i=1}^{n} x_i[y_i - \pi_i] = 0$$

şeklindedir. Bu eşitlikler olabilirlik denklemleri olarak adlandırılır. Böylelikle β için bu denklem sistemleri çözülerek parametrelerin En çok olabilirlik (maximum likelihood) tahminleri bulunabilir.

Matris notasyonu ile ifadesi,

$$\mathrm{X}'(\mathrm{y} - \mu) = 0$$

şeklindedir. Sonuç olarak bu aslında her bir model parametresi için $p = k + 1$ denklemden oluşan bir denklem sistemidir ve bu denklem sistemi skor denklemleri adını almaktadır (Montgomery-Vining-Myers, 2002:105).

Doğrusal regresyon analizinde β'ya göre türevinden elde edilen eşitlikler bilinmeyen parametreleri içeren doğrusal ifadeler oldukları için kolayca çözümlenebilirler. Fakat lojistik regresyon için elde edilen eşitlikler β_0 ve β_1'de doğrusal değildir. Bu yüzden eşitliklerin çözümlenebilmesi için özel yöntemlere ihtiyaç vardır. En çok olabilirlik (maximum likelihood) tahmincilerini gerçek olarak bulmak için İteratif Olarak Yeniden Ağırlıklandırılan En Küçük Kareler yöntemini kullanabiliriz. İterasyonlar arasında fark olmaması durumunda yakınsama sağlanır ve iterasyon işlemine yakınsama sağlanıncaya kadar devam edilir (Hosmer-Lemeshow, 1989'dan aktaran Bircan, 2004).

$\hat{\beta}$'yı bu yöntemlerin ürettiği model parametrelerinin son tahmini olarak kabul edersek ve eğer model parametreleri doğru ise asimptotik olarak aşağıdaki eşitlikler gösterilebilir:

$$E(\hat{\beta}) = \beta$$

ve

$$var(\hat{\beta}) = (X' \mathrm{V}^{-1} X)^{-1}$$

Doğrusal kestiricinin tahminlenen değeri $\hat{\eta}_i = \mathbf{x}_i'\hat{\beta}$'dır ve uyumu yapılan lojistik regresyon modeli aşağıdaki gibi yazılır:

$$\hat{y}_i = \hat{\pi}_i = \frac{exp\ (\hat{\eta})}{1 + exp\ (\hat{\eta})}$$

$$\hat{y}_i = \hat{\pi}_i = \frac{exp\ (\mathbf{x_i}'\hat{\beta})}{1 + exp\ (\mathbf{x_i}'\hat{\beta})} \quad \text{ya da} \quad \hat{y}_i = \hat{\pi}_i = \frac{1}{1 + exp\ (-\mathbf{x_i}'\hat{\beta})}$$

β katsayısının değeri x ve y'nin logiti arasındaki ilişkinin yönünü belirler. β, 0'dan büyük olduğu zaman daha büyük (ya da daha küçük) x değerleri y'nin daha büyük (ya da daha küçük) logitleri ile birleşir. Aksine β, 0' dan daha küçük olursa daha büyük (ya da daha küçük) x değerleri y'nin daha küçük (ya da büyük) logitleri ile birleşir (Peng-Lee-İngersoll, 2002).

3.3. Lojistik Regresyon Modelindeki Parametrelerin Önem Testi

Lojistik regresyon analizinde, katsayıların tahmininden sonra oluşturulan modelin içerdiği değişkenlerin anlamlılığı test edilmektedir. Bu değerlendirme genelde modelde bulunan bağımsız değişkenlerin bağımlı değişken ile önemli bir şekilde ilişki içinde olup olmadığının testi şeklindedir. Lojistik regresyon analizinde değişkenlerin anlamlı olup olmadıklarını sınayan ve yaygın olarak kullanılan üç test mevcuttur. Bunlar Olabilirlik Oran Testi (Likelihood Ratio Test), Sapma (Deviance), Wald Testi (Wald Test) ve Skor Testi (Score Test)'dir.

- Model Sapması: Bir modelin sapması (deviance), uyumu yapılan bir modelin log olabilirliği ile doymuş (saturated) bir modelin log-olabilirliğini karşılaştırır.[2]

Lojistik regresyon modeli için bunun anlamı, π_i olasılıkları tamamiyle sınırlandırılmadığı için π_i'yi y_i'ye (y_i= 0 ya da 1) eşit almak olabilirliği maksimum yapacaktır. Bu durum doymuş model için log-olabilirlik fonksiyonunun bir maksimum değerinde sonuçlanır. Bu nedenle log-olabilirlik fonksiyonunun maksimum değeri sıfırdır. (Montgomery-Peck-Vining, 1992) Maksimum olabilirlik tahmincileri $\hat{\beta}$'lar uyumu yapılan modelin log-olabilirlik fonksiyonunda kullanıldığında, fonksiyon maksimum değerine ulaşır. Bu maksimum-olabilirlik fonksiyonu aşağıdaki gibidir:

$$lnL(\hat{\beta}) = \sum_{i=1}^{n} y_i \mathbf{x}_i'\hat{\beta} - \sum_{i=1}^{n} ln[1 + exp(\mathbf{x}_i'\beta)]$$

Uyumu yapılan model daha az parametre içereceğinden, bu model için log-olabilirlik fonksiyonunun değeri asla doymuş model için log-olabilirlik fonksiyonunun değerini geçemez.

Model sapması aşağıdaki gibi tanımlanır:

$$D(\beta) = -2ln\left[\frac{\mathcal{L}(\beta)}{\mathcal{L}(\hat{\beta})}\right]$$

ya da

$$D(\beta) = -2ln\left[\frac{uyumu\ yapılan\ modelin\ olabilirliği}{doymuş\ modelin\ olabilirliği}\right]$$

$(-2\ ln)$ katının alınması matematiksel olduğu kadar aynı zamanda dağılımı bilinen bir değer elde etmek içindir. Eğer lojistik regresyon modeli doğru regresyon fonksiyonu ise ve örneklem genişliği n büyükse, p modelde yer alan parametre sayısını göstermek üzere model sapması $(n-p)$ serbestlik derecesi ile yaklaşık bir χ^2 dağılımına sahiptir.

$$D(\beta) \sim {\chi^2}_{n-p}$$

Test istatistiği ve karar kriteri şu şekildedir:

$$D(\beta) \le \chi^2_{\alpha,n-p} \Rightarrow uyumu\ yapılan\ model\ yeterli$$

[2] n tane gözlem değeri için n tane parametre tahmini yapılıyorsa buna doymuş model denir.

$D(\beta) > \chi^2_{\alpha,n-p} \;\; \Rightarrow uyumu\ yapılan\ model\ yeterli\ değil$

Model sapmasının büyük çıkması uyumu yapılan modelin doğru model olmadığını gösterir. Model sapmasının küçük çıkması ise uyumu yapılan modelin doymuş model kadar iyi olduğunu gösterir (Montgomery-Vining-Myers, 2002:113).

- Olabilirlik Oran Testi: Lojistik regresyon analizinde (ve genel olarak GLM için) hipotez testi olabilirlik oran testine dayanır. Bu bir büyük örneklem prosedürü olduğu için test prosedürleri asimptotik teoriye dayanır.

Doğrusal regresyon modelinde hipotezleri test etmek için regresyon (ya da hata) kareler toplamlarındaki farkı kullandığımız gibi model parametrelerinin anlamlı olup olmadıklarının belirlenmesinde, bağımsız değişkeni içeren modelin sapması bağımsız değişkeni içermeyen modelin sapması ile karşılaştırılmaktadır. (Montgomery-Peck-Vining, 1992) Yani modelde yer alan bağımsız bir değişkenin önemine karar vermek için, denklemde bağımsız değişkenlerin yer aldığı durumdaki sapma değeri ile bağımsız değişkenlerin yer almadığı durumdaki sapma değeri karşılaştırılır.

D değerindeki bu değişim G istatistiği olarak da adlandırılmaktadır ve aşağıdaki gibi tanımlanır:

$$G = D(değişkeni\ içermeyen\ model\ için) - D(değişkeni\ içeren\ model\ için)$$

ya da

$$G = -2ln\left[\frac{değişkensiz\ modelin\ olabilirliği}{değişkenli\ modelin\ olabilirliği}\right]$$

Modelin aşağıdaki gibi olduğunu varsayalım.

$\eta = x\beta$

$\eta = x_1\beta_1 + x_2\beta_2$

Burada tam model p sayıda parametreye sahiptir. β_1 bu parametrelerin $p-r$ adedini, β_2 ise r adedini içerir ve x_1, x_2 matrislerinin sütunları bu parametrelerle ilişkili değişkenleri içerir.

Aşağıdaki hipotezleri test etmek istediğimizi varsayalım:

$H_0 : \beta_2 = 0$

$H_1 : \beta_2 \neq 0$

Bu durumda daraltılmış (reduced) model,

$\eta = X_1\beta_1$ olur.

Bu ifade $n-(p-r)-(n-p) = r$ serbestlik derecesine sahiptir. Eğer sıfır hipotezi gerçekse ve eğer n büyükse sapmadaki fark r serbestlik derecesi ile bir χ^2 dağılımına sahiptir.

Test istatistiği ve karar kriteri şu şekildedir:

$G \geq \chi^2_{\propto,r} \Rightarrow$ sıfır hipotezi reddedilir.

$G < \chi^2_{\propto,r} \Rightarrow$ sıfır hipotezi reddedilemez.

Daraltılmış model daha az parametre içereceğinden daraltılmış model için sapma, her zaman tam model için sapmadan daha büyük olacaktır. Buna karşın eğer daraltılmış model için sapma, tam model için sapmadan çok büyük değilse daraltılmış model neredeyse tam model kadar iyidir ve bu nedenle β_2'deki parametreler muhtemelen sıfıra eşittir yorumu yapılır. Yani bu durumda sıfır hipotezini reddedemeyiz (Montgomery-Peck-Vining, 1992).

G istatistiği tüm katsayıların testinde kullanılabileceği gibi eğim parametresi $H_0: \beta_1 = 0$ hipotezinin testinde 1 serbestlik dereceli bir χ^2 dağılımına sahiptir (Oğuzlar Ayşe, 2005).

Olabilirlik oran testi için test istatistiği log-olabilirlik oranının -2 katına eşittir.

$$-2\ln\left[\frac{\mathcal{L}(\text{daraltımış model})}{\mathcal{L}(\text{full model})}\right] \sim \chi^2$$

Fakat bu sapmadaki farkla tam olarak aynıdır. (Myers-Montgomery-Vining, 2002, 112)

- Wald Testi: Bireysel model katsayıları üzerine bir diğer yaklaşımda en çok olabilirlik tahmincileri teorisine dayanmaktadır. Büyük örneklem için bir maksimum olabilirlik tahmincisinin dağılımı, küçük sapmayla ya da hiç sapmasız yaklaşık olarak normaldir. Maksimum olabilirlik tahmincilerinin bir kümesinin varyans ve kovaryansları log-olabilirlik fonksiyonunun model parametrelerine göre ikinci mertebeden türevleri alınarak bulunabilir. Wald testide bu temele dayanmaktadır. Wald testi, β_j'in en çok olabilirlik tahmini ile bu tahminin standart hatasını karşılaştırır (Montgomery-Peck-Vining, 1992).

G, log-olabilirlik fonksiyonunun ikinci mertebeden kısmi türevlerinin $p \times p$ boyutlu matrisi olsun. Yani,

$$G_{ij} = \frac{\partial^2 \mathcal{L}(\beta)}{\partial\beta_i\, \partial\beta_j} \qquad \text{i,j=0,1,.....,k olsun.}$$

G matrisi Hessian Matrisi olarak adlandırılır. Eğer Hessian Matrisinin elemanları $\beta = \hat{\beta}$ maksimum olabilirlik tahmincileriyle değerlendirilirse, regresyon katsayılarının büyük örneklem tahmin kovaryans matrisi aşağıdaki şekilde gösterilir:

$$\text{var}(\hat{\beta}) = \hat{\Sigma} = -G(\hat{\beta})^{-1}$$

Yani Hessian Matrisinin tersinin negatifi varyans-kovaryans matrisine eşittir. Bu nedenle,

$H_0: \beta_j = 0$

$H_1: \beta_j \neq 0$

ile ifade edilen sıfır hipotezi için test istatistiği,

$Z_j = \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)}$ ‘dir.

$\hat{\beta}_j$ ‘in standart hatası, varyans kovaryans matrisindeki köşegen elemanlarının kareköklerinin alınmasıyla elde edilmektedir. Her değişken için listedeki standart hatalar kullanılarak Z testi yapılır. Bu istatistik için referans dağılım, standart normal dağılımdır.

Normal rassal bir değişkenin karesinin alınması 1 serbestlik dereceli bir χ^2 rassal değişkenine eşit olacağından, Wald istatistiği şu şekilde de gösterilebilir (Myers-Montgomery-Vining, 110):

$$Z_j^{\,2} = \left(\frac{\hat{\beta}_j}{SE(\hat{\beta}_j)}\right)^2 \sim \chi^2$$

- Skor Testi: Parametrelerin anlamlılığının testinde kullanılan bir başka testte skor testidir. Bu test en çok olabilirlik denklemlerinin türevlerinin koşullu dağılımlarına dayanmaktadır ve şu şekilde gösterilir:

$$ST = \frac{\sum_{i=1}^{n} x_i(y_i - \bar{y})}{\sqrt{\bar{y}(1-\bar{y})\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

Bu istatistik eğim parametresini gösteren $\beta_1 = 0$ hipotezi için standart normal dağılım göstermektedir (Oğuzlar, 2005).

3.4. Lojistik Regresyon Modelindeki Parametrelerin Yorumlanması

Lojistik regresyon modelinde, tahmin edilen regresyon katsayılarının yorumlanması doğrusal regresyon modelindeki katsayıların yorumlanması kadar kolay değildir. Lojistik regresyon modelinde x değişkenindeki bir birimlik artışı ölçmek zordur. $\hat{\beta}_1$ katsayısı yorumlanırken x’de ki 1 birimlik artış için $\frac{\pi_i}{1-\pi_i}$ odds oranı ile $exp\ (\hat{\beta}_1)$ çarpılarak elde edilen sonuca bakılarak değerlendirme yapılır. Aynı şekilde kestirici değişkende d birimlik bir değişim söz konusu ise odds oranındaki tahminlenen artış, $exp\ (d\hat{\beta}_1)$’dır.

Her bir parametrenin $exp\ (\hat{\beta}_k)$ değeri odds oranıdır. Böylece $exp\ (\hat{\beta}_k)$, y değişkeninin x_k değişkeninin etkisi ile kaç kat daha fazla gözlenme olasılığına sahip olduğunu belirtir.

$exp\ (\hat{\beta}_k)$ değeri eğer 1'den büyükse katsayının anlamlı olması koşulu ile ilgili değişkenin önemli bir etken olduğuna karar verilir ve modele dahil edilir. $exp\ (\hat{\beta}_k)$ değeri 0'a yakın ise de bu değişkenin y üzerinde önemli bir etkisi olduğu fakat bu etkinin negatif yönlü olduğuna karar verilir.

İlk olarak modelimizin bir bağımsız değişkene sahip olduğunu düşünürsek, bu durumda x'in özel bir değerinde $(\mathrm{x}_i \text{ gibi})$ uyumu yapılan modelin değeri;

$$\hat{\eta}_{(x_i)}=\hat{\beta}_0+\hat{\beta}_1 x_i$$

olur.

x_i+1 ' de uyumu yapılan modelin 1 dönem sonraki değeri ise, uyumu yapılan modelin değeri;

$$\hat{\eta}_{(x_i+1)}=\hat{\beta}_0+\hat{\beta}_1(x_i+1)$$

şeklinde olur.

Bu iki tahminlenen modelin birbirine oranı,

$$\hat{\eta}_{(x_i+1)}-\hat{\eta}_{(x_i)}=\hat{\beta}_1$$

olur.

Bu istatistik bize x=1 olan başarı olasılığının x=0 olan başarısızlık olasılığına göre bağımlı değişkenin kaç kere daha fazla 1 olarak görüldüğü sonucunu verir. Başka bir ifadeyle,

$$\hat{\eta}_{(x_{i+1})}-\hat{\eta}_{(x_i)}=ln\left(odds_{x_{i+1}}\right)-ln\left(odds_{x_i}\right)$$

$$\hat{\eta}_{(x_{i+1})}-\hat{\eta}_{(x_i)}=ln\left(\frac{odds_{x_{i+1}}}{odds_{x_i}}\right)$$

$$\hat{\eta}_{(x_{i+1})}-\hat{\eta}_{(x_i)}=\hat{\beta}_1$$

Eğer antilog alınırsa aşağıdaki odds oranı elde edilir: (Montgomery-Peck-Vining)

$$\hat{O}_r=\frac{odds_{x_{i+1}}}{odds_{x_i}}=e^{\hat{\beta}_1}$$

Bu istatistik, x=1 olan bireylerin x=0 olan bireylere göre bağımlı değişkende kaç kere daha fazla 1 olarak görüldüğünün sonucunu verir. Odds oranı 1 olduğunda logit değeri 0'a eşittir. Odss oranı 1'den büyük değerler aldığında logit değeri yavaşça artarken, 1'den küçük olduğunda logit değeri hızla azalır.

Çoklu lojistik regresyon modelindeki regresyon katsayılarının yorumu, sadece bir bağımsız değişken içeren doğrusal model için yapılan yorumla aynıdır. Yani diğer tüm değişkenlerin sabit olduğu varsayıldığında, $exp\ (\hat{\beta}_j)$ çokluğu x_j değişkeni için odds oranıdır. (Montgomery-Peck-Vining, 1992)

3.5. Modelin Uyum İyiliği

Uyum iyiliği (Goodness of fit), bağımlı değişken için kurulan modelin ne kadar etkin olduğunun göstergesidir.

Doğrusal regresyon analizinde modelin önemliliği test etmede varyans analizinden yararlandığımız gibi lojistik regresyon analizinde de modelin uyum iyiliğini test etmede Hosmer-Lemeshow testinden yararlanabiliriz. Ayrıca Pearson ve Sapma (deviance) istatistikleri de modelin uyum iyiliğinin testinde kullanılan yöntemlerdendir (Montgomery-Vining-Myers, 2002:113).

Hosmer-lemeshow testinde, teorik frekansların 5'den büyük olması sağlanmaya çalışılmakta ve böylece serbestlik derecesinin düşmesi sağlanmaktır. Serbestlik derecesi de χ^2 dağılımına uygunluk sağlamakta, böylece de güvenilir bir uyum iyiliği ölçütü olmaktadır. Testte tahmin edilen olasılık değerleri gruplandırılmaktadır. Örneğin $g = 10$ grup için sınır değerler belirlendikten sonra, olasılığı 0,1'den küçük olanlar 1. ve olasılığı 0,9'dan büyük olanlar 10. grup şeklinde atama yapılmaktadır. Testin güvenirliği için gözlenen ve beklenen frekanslar tablosunda 5'ten küçük değer olmaması gerekmektedir, ayrıca grupların sayısı 6'dan az olmamalıdır (Arabacı, 2002).

Hosmer-Lemeshow uyum iyiliği istatistiği $\hat{C}$, gözlenen ve teorik frekanslardan oluşan $2 \times g$ tablosu için χ^2 değeri olarak bulunur ve aşağıdaki şekilde hesaplanır:

$$\hat{C} = \sum_{k=1}^{g} \frac{(o_k - n_k \bar{\pi}_k)^2}{n_k \bar{\pi}_k (1 - \bar{\pi}_k)}$$

Formülde yer alan o_k, gözlenen frekanslardır ve $o_k = \sum_{j=1}^{n_k} y_j$ olarak bulunur.

$\bar{\pi}_k$; tahmin edilen olasılık değerlerinin ortalamasıdır ve $\bar{\pi}_k = \sum_{j=1}^{n_k} \frac{m_j \hat{\pi}_j}{n_k}$ şeklinde hesaplanır.

Hosmer-lemeshow $\hat{C}$ istatistiği, $g - 2$ serbestlik dereceli χ^2 dağılımı göstermektedir.

3.6. Lojistik Regresyonda Aşırı Yayılım

Lojistik regresyon modelini seçmek bazı durumlarda hatalı olabilir. Bu hatalı durumlara neden olan başlıca sorunlar şunlar olabilir:

- Binomial varsayım hatalı olabilir.
- Seçilen logit model hatalı olabilir.
- Kullanılan doğrusal kestirici yapısı hatalı olabilir.

- Verilerin içerisinde aşırı uç değerler (outliners) olabilir.

Bazen de seçilen dağılım ve model yeterince uygun ve veri seti aşırı uç değerlerden yoksun olmasına rağmen hala ortalama sapma göze çarpan bir problem olabilir. Yani gözlenen varyans beklenen varyanstan büyük olabilir. Bu duruma aşırı yayılım (overdispersion) ya da ekstra-binomial varyasyon denir (Myers-Montgomery-Vining,2002:126).

Lojistik regresyonda aşırı yayılıma (overdispersion) başarı olasılıkları arasındaki varyasyon ve ikili yanıt değişkenleri arasındaki korelasyon neden olmaktadır.

Aşırı yayılımın söz konusu olmadığı durumda lojistik regresyonda Y_i bağımlı değişkeninin beklenen değerini,

$$E(Y_i) = n_i \pi_i$$

şeklinde ve varyansını da,

$$var(Y_i) = n_i \pi_i (1 - \pi_i)$$

şeklinde ifade edebiliriz.

Aşırı yayılımı analitik örneklerle açıklamanın basit bir yolu da binomial parametre π_i'ya çeşitli değerler vermektir. Gerçekte deneysel şartların sabit olduğu verilen bir grupta, π_i, μ ortalama ve $\emptyset > 1$ varyansla temsil edilir.

Eğer Y rassal bir binomial değişken ise beklenen değer değişiklik göstermez ve aşağıdaki şekilde gösterilir

$$E(Y_i) = E[E(Y|\pi_i)] = nE(\pi_i) = \pi_i$$

Fakat varyans aşağıdaki şekilde değişikliği uğrar:

$$var(Y_i) = var[E(Y|\pi_i)] + E[var(Y|\pi_i)]$$

$$var[E(Y|\pi_i)] = var[n\pi_i] = n^2\phi$$

$$E[var(Y|\pi_i)] = nE[\pi_i(1 - \pi_i)]$$

$$nE[\pi_i(1 - \pi_i)] = n[E(\pi_i) - E(\pi_i{}^2)] = n[\mu - (\phi + \mu^2)]$$

$$var(Y_i) = n^2\phi + n\mu - n\phi - n\mu^2$$

$$var(Y_i) = n\mu(1 - \mu) + n\phi(n - 1)$$

$$var(Y_i) > n\mu(1 - \mu)$$

Eşitliklerden de görüldüğü gibi ancak $\phi = 0$ olduğu durumlarda beklenen varyans gözlenen eşit olur. Bu durumda da aşırı yayılım (overdispersion) söz konusu değildir yorumu yapılır. $\phi < 1$ olduğu durumda gözlenen varyans beklenen varyansa eşit olmayıp, daha küçük olacaktır. Bu durumunda veri kümesinde eksik yayılım (underdispersion) vardır yorumu yapılır. $\phi > 1$ olduğu durumda ise gözlenen varyans beklenen varyansa eşit olmayıp, daha büyük olacaktır. Bu durumunda da veri kümesinde de aşırı yayılım (overdispersion) vardır yorumu yapılır. (Myers-Montgomery-Vining, 2002:127)

Lojistik regresyonda yaygın olarak kullanılan iki uyum istatistiği; sapma ve pearson ki-kare uyum istatistikleridir. Bu uyum istatistikleri aşırı yayılımı (overdispersion) belirlemede de kullanılmaktadır. Bu iki uyum istatistiğinin değerinin 1'e eşit ya da çok yakın çıkması regresyonda aşırı yayılımın (overdispersion) olmadığını gösterir.

- Model Sapması: Bir modelde aşırı yayılımın (overdispersion) olup olmadığını anlamak için model sapmasının (deviance) değeri serbestlik derecesine bölünür. Çıkan sonuç 1'i çok fazla aşıyorsa aşırı yayılımın (overdispersion) olduğu söylenebilir.

Lojistik regresyonda modelin sapması, standart lineer regresyon modellerindeki ortalama kare hataya paraleldir.

- Pearson Ki Kare İstatistiği: Serbestlik derecesine bölünmüş bir χ^2 istatistiğidir.

$$\frac{1}{n-p}\sum_{i=1}^{m}\left[\frac{(y_i-\hat{y}_i)^2}{n_i\hat{\pi}_i(1-\hat{\pi}_i)}\right]=\frac{\chi^2}{n-p}$$

Kalıntı karelerinin binomial varyans tarafından bölünmesi binomial varyans standardizesi içindir. Çıkan sonuç 1'i çok fazla aşıyorsa aşırı yayılımın (overdispersion) olduğu söylenebilir.

Uygun lojistik regresyonun aşırı yayılıma (overdispersion) sahip sonuçlarının etkisi bildiğimiz standart lineer regresyondan oldukça şüpheli görünebilir. Standart lineer regresyonda eğer kusurlu model artıklarının değişkenliği tarafından tamamen şişirilmişse, ortalama hata kareyi yüksek verir ve sonrasında da regresyon katsayılarının standart hataları düşük tahmin edilir.

Tahmincilerin varyans kovaryans matrisi $(X'X)^{-1}s^2$ tarafından tahminlenir. s^2 ortalama hata karedir ve şişirilmiştir.

Lojistik regreyonda aşırı yayılım (overdispersion) olduğu taktirde ölçek parametresi $\sigma^2 > 1$, varyans kovaryans matrisine aynı şekilde girer.

$$var(b) = (X'V^{-1}X)^{-1}\sigma^2$$

Böylelikle standart hatalar $\sigma^2 > 1$ olmadığı sürece düşük tahminlidirler.

Uygun model ile aşırı yayılım durumunda β'ların maksimum olabilirlik tahmincileri asimptotik olarak yansız kalacaktır (Myers-Montgomery-Vining, 2002, 128).

Eğer uyum istatistikleri sonucunda aşırı yayılım (overdispersion) durumu söz konusu ise bilinen lojistik regresyonu kullanmak doğru olmayacaktır. Bunun yerine aşırı yayılımı (overdispersion) açıklayan ve bunu modele dahil ederek bu sorunu ortadan kaldıran modelleri tercih etmek doğru olacaktır. Bu yöntemlerden biri de Williams metodudur.

DÖRDÜNCÜ BÖLÜM

POİSSON REGRESYON ANALİZİ

Bir olayın sabit bir zaman aralığında yapılan denemeler sonucunda meydana gelme sayısına sayma verileri denir. Sayma verilerine doğrusal regresyonu uygulamak doğru sonuçları vermeyecektir. Bu yüzden sayma verilerini analiz etmede en çok kullanılan ve en basit yöntem Poisson regresyon analizidir.

Poisson regresyon analizi, açıklayıcı değişkenler ile sayıma dayalı olarak elde edilen bağımlı değişkenler arasındaki ilişkiyi açıklar. Bağımlı değişkenin 0,1,2,3... gibi kesikli değerler aldığı fakat kategorik olmadığı durumlarda Poisson regresyon analizi sayma verilerini analiz etmede kullanılır. Poisson regresyon analizi bağımlı değişkenin Poisson dağılım göstermesi esasına dayanır. Poisson dağılım bir ayrık olasılık dağılımı olup belli bir sabit zaman birim aralığında meydana gelme sayısının olasılığını ifade eder.

Poisson regresyon modellerinde bilinen ilk gelişmeler aktüeryal bilimler, biyoistatistik ve demografide gözlenmiştir. Son yıllarda da bu modeller iktisat, politik bilimler ve sosyolojide sıkça kullanılmaya başlanmıştır. Türkiye'de Poisson regresyon analiziyle ilgili olarak yapılmış çok fazla sayıda çalışma bulunmamaktadır.

Şahin (2002), çalışmasında 1964-1998 yılları arasında Türkiye'de yapılan grev sayısını etkileyen faktörleri ortaya koymayı hedeflemiş ve bu amaçla da Poisson regresyon analizi yöntemini kullanmıştır. Grevin belirleyicileri olarak da işsizlik oranı, sendikalaşma oranı, çalışan başına milli gelirin değişim oranı ve kukla değişken kullanılmış. Analiz sonucunda, milli gelirin değişim oranının negatif yönde fakat istatistiki olarak anlamlı olmayan bir şekilde grev sayısını etkilediği görülmüş. Yine sendikalaşma oranının grev sayısı üzerinde etkili olmadığı sonucuna varılmış. Buna karşın işsizlik oranı ve kukla değişkenin grev sayısını etkilediği görülmüş. Sendikalaşma oranı grev sayısını etkilemediğinden ve mevcut anayasa ortalama olarak daha fazla greve neden olduğundan bu analiz sonucunda önerilebilecek politikalar, işçi sendikalaşmalarına engel konulmaması ve anayasal düzenlemelerin greve gitme konusundaki kısıtlamalarının tekrar gözden geçirilmesi şeklinde değerlendirilmiştir.

4.1. Poisson Regresyon Modelleri

Y rassal değişkeni belirli bir zaman içerisinde bir olayın gerçekleşme sayısını göstersin. Bu rassal değişken Poisson dağılımına sahipse $\mathbf{x}_i$'ye bağlı y_i için regresyon modeli;

$$P(y_i) = \frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!} \qquad y_i = 0,1,2,\ldots.$$

olarak gösterilir.

Ortalama parametresi ise;

$$E(y_j) = \mu_i = exp\ (x'_i\beta)$$

şeklinde gösterilir.

Burada $x'_i = [1, x_{i1}, x_{i2}, \ldots., x_{ik}]$'dır. Bu fonksiyona aynı zamanda log-doğrusal modelde denir. Çünkü koşullu ortalamanın logaritması parametreleri doğrusal olarak verir.

$$lnE(y_j|x_i) = \mu_i = x'_i\beta$$

y_i bağımlı değişkeni,

$E(y_i) = \mu_i$ ve $var(y_i) = \mu_i$

ile bağımsız Poisson dağılmaktadır (Montgomery-Myers-Vining, 2002:131).

Bundan da anlaşıldığı üzere Poisson regresyonda bağımlı değişkeninin koşullu ortalaması ile koşullu varyansı birbirine eşittir. Ancak bazen uygulamada koşullu varyansın koşullu ortalamayı aştığı durumlar ortaya çıkabilir. Bu gibi durumlarda da negatif binom regresyon modelleri kullanılmaktadır. Bununla birlikte y'nin dağılımı yanlış belirlenmiş olsa bile Quasi maksimum olabilirlik tahmin edicileri elde edilen tahmin edicilerin istatistiki yorumda kullanılmasına olanak sağlar (Şahin, 2002). Çünkü varsayılan dağılımın yanlış olması halinde bile Quasi maksimum olabilirlik tahmin edicileri tutarlı ve asimptotik olarak normal dağılıma sahiptir (Mittelhamer ve diğerleri, 2000:245). Bu anlamda Selim ve Üçdoğruk (2003) yaptıkları çalışmada, yayılım durumu ele alındığında genellikle çocuk sayısı verileri için eksik yayılım (underdispersion) durumu ile karşılaşıldığından hanehalkı çocuk sayısı belirleyicilerini modellemek için Standart Poisson Regresyon Modeli (PRM) yerine tutarlı tahminciler veren Poisson Quasi Maksimum Olabilirlik (PQML) tahmini kullanmışlardır. Bununla birlikte dağılımın doğru olması halinde Maksimum Olabilirlik yerine Quasi Maksimum Olabilirlik kullanmak asimptotik etkinlikten vazgeçmektir.

4.2. Poisson Regresyon Modellerinde Parametre Tahmini

Poisson regresyon modellerinde paramatre tahmini için çeşitli yöntemler kullanılabilmektedir. En çok olabilirlik yöntemi, doğrusal ve karesel varyans fonksiyonları ile negatif binom, yapay en çok olabilirlik bu yöntemlerden bazılarıdır. Bunlardan poisson regresyon için en sık kullanılan yöntem en çok olabilirlik yöntemidir. Poisson regresyonda en çok olabilirlik kestiriminin uygulanabilmesi için y bağımlı değişkeni poisson dağılım göstermeli, koşullu varyans ve koşullu ortalamanın birbirine eşit olması gerekir. Eğer koşullu varyans ve ortalama eşit değilse en çok olabilirlik tahmin edicisinden daha etkin tahmin ediciler kullanılabilir. Ancak en çok olabilirlik yöntemi kullanılacaksa da tahmincilerin istatistiksel olarak geçerli olabilmesi için en azından koşullu ortalama doğru tanımlanmalıdır (Deniz, 2005).

Poisson regresyon için en çok olabilirlik fonksiyonu;

$$lnL(y_1, y_2, \dots, y_n; \beta) = ln \prod_{i=1}^{n} \left[\frac{e^{-\mu} \mu_i{}^{y_i}}{y_i!}\right]$$

$$lnL(y_1, y_2, \dots, y_n; \beta) = \sum_{i=1}^{n} \{lne^{-\mu_i} + ln\mu_i^{y_i} - lny_i!\}$$

$$lnL(y_1, y_2, \dots, y_n; \beta) = \sum_{i=1}^{n} \{-\mu_i + y_i ln\mu_i - lny_i!\}$$

$$lnL(y_1, y_2, \dots, y_n; \beta) = \{-e^{\mathbf{x}'_i\boldsymbol{\beta}} + y_i lne^{\mathbf{x}'_i\boldsymbol{\beta}} - lny_i!\}$$

$$lnL(\beta) = \sum_{i=1}^{n} \{-e^{\mathbf{x}'_i\boldsymbol{\beta}} + y_i \mathbf{x}'_i\boldsymbol{\beta} - lny_i!\}$$

$$\frac{\delta lnL(\beta)}{\delta\beta} = \sum_{i=1}^{n} \{-\mathbf{x}_i e^{\mathbf{x}'_i\boldsymbol{\beta}} + y_i \mathbf{x}_i\}$$

$$\frac{\delta lnL(\beta)}{\delta\beta} = \sum_{i=1}^{n} \{y_i \mathbf{x}_i - e^{\mathbf{x}'_i\boldsymbol{\beta}} \mathbf{x}_i\}$$

$$\frac{\delta lnL(\beta)}{\delta\beta} = \sum_{i=1}^{n} \{y_i \mathbf{x}_i - \mu_i \mathbf{x}_i\} = 0$$

$$\frac{\delta lnL(\beta)}{\delta\beta} = \sum_{i=1}^{n} (y_i - \mu_i) \mathbf{x}_i = 0$$

$$\sum_{i=1}^{n} (y_i - \mu_i) \mathbf{x}_i = 0$$

Bu denklem skor denklemidir ve matris notasyonu ile de;

$$\mathbf{X}'(\mathbf{y} - \boldsymbol{\mu}) = \mathbf{0}$$

şeklinde gösterilir.

Doğrusal regresyon analizinde β'ya göre türevinden elde edilen eşitlikler bilinmeyen parametreleri içeren doğrusal ifadeler olduklarından kolayca çözümlenebilirler. Fakat Poisson regresyon için elde edilen eşitlikler β'ya göre doğrusal değildir. Bu yüzden eşitliklerin çözümlenebilmesi için özel yöntemlere ihtiyaç vardır. En Çok Olabilirlik tahmincilerini gerçek olarak bulmak için kullanılan standart yöntem Fisher iterasyon yöntemidir. İteresyonlar arasında fark olmaması durumunda yakınsama sağlanır ve iterasyon işlemine yakınsama sağlanıncaya kadar devam edilir.

4.3. Poisson Regresyon Modelindeki Parametrelerin Önem Testi

Poisson regresyon modelinde katsayıların tahmin edilmesinin ardından sıra modele dahil edilen bu değişkenlerin anlamlılıklarının sınanmasına gelir. Poisson regresyon analizinde değişkenlerin anlamlı olup olmadıklarını sınayan ve yaygın olarak kullanılan üç test mevcuttur. Bunlar Olabilirlik Oran Testi (likelihood ratio test), Wald Testi (wald test) ve Skor Testi (score test)'dir.

Olabilirlik oran istatistiği ve Wald çıkarsaması için kullanılan hipotez testleri lojistik regresyon modellerinde açıklananlar ile aynıdır. Her iki model için de varyans ortalamanın bir fonksiyonudur ve skor denklemi β'nın tahminini verir (Myers-Montgomery-Vining, 2002:134).

- Model Sapması:

Doymuş (saturated) modelde log-olabilirlik;

$$lnL(\beta) = -\sum_{i=1}^{n} y_i + \sum_{i=1}^{n} y_i ln y_i - \sum_{i=1}^{n} y_i!$$

Poisson regresyon modeli için;

$$lnL(\beta) = -\sum_{i=1}^{n} \hat{\mu}_i + \sum_{i=1}^{n} y_i ln\hat{\mu}_i - \sum_{i=1}^{n} y_i!$$

$$D(\beta) = -2ln\,[\mathcal{L}(\beta)/\mathcal{L}(\mu)]$$

$$D(\beta) = -2\{ln(\mathcal{L}(\beta)) - ln(\mathcal{L}(\mu))\}$$

$$D(\beta) = -2\{[-\sum_{i=1}^{n} y_i + \sum_{i=1}^{n} y_i ln y_i - \sum_{i=1}^{n} y_i!] - [-\sum_{i=1}^{n} \hat{\mu}_i + \sum_{i=1}^{n} y_i ln\hat{\mu}_i - \sum_{i=1}^{n} y_i!]\}$$

$$D(\beta) = -2[-\sum_{i=1}^{n}(y_i - \hat{\mu}_i) + \sum_{i=1}^{n} y_i(ln y_i{}_{-} ln\mu_i)]$$

$$D(\beta) = -2[-\sum_{i=1}^{n}(y_i - \hat{\mu}_i) + \sum_{i=1}^{n} y_i ln\left(\frac{y_i}{\hat{\mu}_i}\right)]$$

$$\sum_{i=1}^{n}(y_i - \hat{\mu}_i) = 0$$

olduğundan,

$$D(\beta) = 2[\sum_{i=1}^{n} y_i ln\left(\frac{y_i}{\hat{\mu}_i}\right)]$$

$$D(\beta) \leq \chi^2{}_{\alpha,n-p} \Rightarrow \text{ uyumu yapılan model yeterli}$$

$D(\beta) > \chi^2{}_{\alpha,n-p} \Rightarrow$ *uyumu yapılam model yeterli değil* (Myers-Montgomery-Vining, 2002:135).

4.4. Poisson Regresyon Modelindeki Katsayıların Yorumlanması

Poisson regresyon üstel bir yapıya sahip olduğu için üstel koşullu beklenen değer;

$$E[y|x] = exp(x'\beta)$$

şeklinde olduğu için bu ifadedeki x göre türev alınıp β değeri yanlız bırakıldığında,

$$\frac{\delta E[y|x]}{\delta x_j} = \beta_j\, exp(x'\beta)$$

sonucu elde edilir.

Bu sonuca göre de ‘j’inci bağımsız değişkende meydana gelen bir birimlik artış (ya da azalış), koşullu ortalamayı. $\widehat{\beta}_j exp(x'\widehat{\beta})$ kadar artırmaktadır (ya da azalmaktadır) şeklinde yorum yapılır.

4.5. Poisson Regresyonda Modelin Uyum İyiliği

Genelleştirilmiş doğrusal modeller için uyum iyiliğinin ölçülmesinde en sık kullanılan iki yöntem Pearson ve Sapma istatistikleridir.

- Sapma İstatistiği:

Genel bir kural aynen lojistik regresyonda olduğu gibi model sapması D’yi serbestlik derecesi$(n-p)$’ye bölmektir. Eğer $\frac{D(\beta)}{(n-p)}$ 1’e yakınsa model yeterli olarak değerlendirilir. Bu oran 1’den büyük ise modelin yeterli olmadığı söylenebilir (Montgomery-Vining-Myers, 2002:113).

- Pearson χ^2 İstatistiği:

$$\chi^2 = \sum_{i=1}^{n} \left[\frac{(y_i-\widehat{\mu_i})^2}{\widehat{\mu_i}}\right]$$

Pearson istatistiği asimptotik olarak ${\chi^2}_{n-p}$ dağılımına sahiptir.

Uyum iyiliği için karar kriteri şu şekildedir:

$H_0: Veriler\ poisson\ regresyona\ uygunluk\ göstermektedir$

$H_1: Veriler\ poisson\ regresyona\ uygunluk\ göstermemektedir$

Hesaplanan değer $(n-p)$ serbestlik dereceli χ^2 değeri ile karşılaştırılır. Hesaplanan değer ${\chi^2}_{n-p}$ değerini aşıyorsa H_0 hipotezi reddedilir ve veriler poisson modele uygunluk göstermemektedir denir (Myers-Montgomery-Vining, 2002:134).

4.6. Poisson Regresyonda Aşırı Yayılım

Poisson dağılımın en önemli özelliği ortalaması ile varyansının eşit olmasıdır. Bazı durumlarda ortalaması varyanstan daha küçük çıkabilir. Bu duruma aşırı yayılım (overdispersion) denir.

Poisson regresyonda da aşırı yayılımı (overdispersion) belirlemede lojistik regresyonda olduğu gibi sapma ve pearson ki-kare uyum istatistikleri kullanılabilir. Bu iki uyum istatistiğinin yayılım parametre değerleri, hesap değerlerinin kendi serbestlik derecelerine bölünmesiyle elde edilmektedir. Her iki uyum istatistiği için elde edilen yayılım parametre değerinin 1’den büyük çıkması regresyonda aşırı yayılımın (overdispersion) varlığına işaret eder.

Aşırı yayılımın (overdispersion) söz konusu olduğu durumlarda bilinen poisson regresyonun kullanılması sakıncalı sonuçlar ortaya çıkarır. Bunun yerine bu durumu açıklayan ve yayılım parametresini modele dahil eden analizler kullanılmalıdır. Bu analizlere örnek olarak karışık poisson regresyon ve karışımlı poisson regresyon analizleri verilebilir.

Karışımlı poisson regresyon analizinde, Poisson dağılımda olduğu gibi varyans ile ortalama birbirine eşit olmaz. Bu sorunu önlemenin en etkili yolu, Poisson dağılımına sahip ancak birden fazla populasyonu temsil eden heterojen veri setini alt populasyonlara ayırarak, her alt populasyona poisson regresyon yöntemini uygulamaktır.

Veri setinin heterojen olmadığı ve aşırı yayılımın söz konusu olmadığı durumda ortalama ve varyans arasındaki ilişki;

$$E(Y_i) = Var(Y_i)$$

$$Var(Y_i | K = k) = 0$$

şeklindedir.

Karışımlı poisson regresyonun ortalama ve varyansı ise,

$$E(y_i) = E(E(y_i/K = k) = \sum_{k=1}^{K} \pi_k \lambda_k$$

$$Var(Y_i) = E\{Var(Y_i/K = k)\} + VarE\{(Y_i/K = k)\}$$

$$Var(Y_i) = \sum_{k=1}^{K} \pi_k \lambda_k + \{\sum_{k=1}^{K} \pi_k {\lambda_k}^2 - (\sum_{k=1}^{K} \pi_k \lambda_k)^2\}$$

şeklindedir (Yeşilova ve arkadaşları, 2004).

BEŞİNCİ BÖLÜM

UYGULAMA

Tütün, yapraklarında tanen, zamk, nişasta, reçine ve nikotin gibi kimyasal maddeler bulunduran bir bitkidir.

Tütünün ana yurdu Amerika'dır. Kuzey Amerika'da, Meksika'da ve Haiti'de yerliler tapınaklarda tütünü yakarak dumanını çekerler ve tedavi amaçlı kullanırlardı.

Tütünün Avrupa'ya girişi ise Christophe Colomb'un Amerika'yı keşfi ile başlar. Colomb'un, yerlilerin tütün içtikleri saz borunun adı olan "Tobacco" adını tütün bitkisine vermesiyle "Nicotina Tobaccum" adı doğmuştur. Fakat tütünün Avrupa'ya yayılması 1559'da Toledu'lu Francisco Hernandez Boncolo'nun tütün tohumlarını İspanya'ya getirmesiyle olur ve bir yıl sonra da İspanya ve Portekiz'de tütün ekimi tamamen gelişir. 1560 yılında da Fransa büyük elçisi Jean NİCOT Fransız sarayını şifa verici bitki olarak tütünle tanıştırdı[3]. Fransa'dan sonra tütün bütün Avrupa'ya ve Asya'ya yayıldı.

Tütün ülkemize 1601 yılında Osmanlı İmparatorluğu döneminde İngiliz gemiciler tarafından getirilmiş ve bazı göğüs hastalıklarına iyi geldiği söylentisiyle kullanımı yaygınlaştırılmıştır. İlk olarak İstanbul'a getirilen tütün süratle tüm Türkiye'ye yayılmıştır.

1856 Kırım Savaşı esnasında gazete kağıdına sarılarak içilen tütünler Türk, İngiliz, Fransız ve yerli ordulara mensup askerler arasında büyük rağbet görmüş ve dönüşlerinde bu alışkanlıklara devam etmeleri sigara sanayisinin temelini oluşturmuştur (Uzunca, 2002:24).

Tütün genellikle sigara şeklinde tüketilmektedir. Bir paket sigarada bulunan nikotin bir kişiye damardan verildiğinde ölüme sebebiyet verecek dozdadır. Bu da sigaranın bireyin ve toplumun sağlığına son derece zararlı olduğunu göstermektedir. Bilim adamları sigaranın ihtiva ettiği nikotinin ve sigara dumanının bünyede kanserden, sinir sistemlerinde bozukluğa kadar bir dizi zarar ve hastalığa yol açtığından söz etmektedir.

Sigaradaki nikotin temelde bağımlılık yapan bir maddedir ve bu da sigaranın kısa sürede alışkanlığa dönüşmesindeki en önemli etkenlerden biridir. Sigara bağımlılığının çok küçük bir oranını fiziksel yön oluşturmakla birlikte asıl büyük orana sahip olan kısmını alışkanlık oluşturmaktadır. Sigaranın çok kısa sürede alışkanlık haline gelmesi, dünyanın her yerinde çok kolaylıkla temin edilebilir olması, zararlı etkilerinin kısa sürede ortaya çıkmaması gibi nedenlerde günden güne artan sigara tüketiminde etkilidir. Ayrıca özellikle sigara bağımlılarının sigaranın sıkıntılarını azalttığı ve zevk verdiği yönündeki düşünceleri de sigaranın zararlı etkilerini önemsemediklerini açıkça ortaya koymaktadır.

[3] Bilim adamları tarafından Jean NICOT sebebiyle tütünün içindeki zehirli maddeye NİKOTİN adı verilmiştir.

Ergenlik döneminde sigaraya başlayan çocuklar, hayatları boyunca sigara bağımlısı olma riski taşırlar. Gençlerin 1/3'ü sigarayı denemektir ve bunların yarısı sigara bağımlısı olma riski taşımaktadırlar. Ne yazık ki, hayatındaki ilk iki sigarasını tamamen bitiren gençlerin %85'i sigara bağımlısı olmaktadır. 1999 yılında yapılan bir araştırmada, gelişmekte olan ülkelerde yaşayan 13-15 yaşındaki gençlerin %10-33'ünün tütün içtiği saptanmıştır.

Günümüzde, her yıl 4 milyon kişi tütün içmeyle ilgili hastalıklar nedeniyle ölmektedir. Bu alışkanlığın yaygınlığı değişmezse, 2020 yılında bu sayının 8 milyon 400 bin kişiye ulaşacağı tahmin edilmektedir. Dünyada yaşı 15'in üstünde olan 1 milyar 200 milyon kişi tütün içmektedir. Dünya sağlık örgütünün verilerine göre dünya genelinde, erkeklerin %47'si, kadınların ise %12'si tütün içmektedir. Türkiye'de ise 1997 yılına ait verilere göre yetişkin erkeklerin %51'i, kadınların ise %49'u günlük düzenli içicidir. Kısacası bir kere başlayınca bir daha zor bırakılan bu alışkanlığa ülkemizde her yıl 100.000 insanımızı erken yaşta kurban vermekteyiz (Bilgel, 2002:63).

Bütün bunlara bağlı olarak sigaraya başlamayı önleme, bırakmayı destekleme ve sigara dumanını kontrol altına almak amacıyla ülkemizde, kamu hizmet binalarıyla koridorlar dahil olmak üzere her türlü eğitim, sağlık, spor ve eğlence yerlerinin kapalı alanlarında sigara içilmesini yasaklayan Tütün Ürünlerinin Zararlarının Önlenmesi ve Kontrolü Hakkında Kanun 19 Mayıs'ta uygulanmaya başlandı. Yasakla birlikte taksiler dahil olmak üzere tüm toplu taşıma araçlarında, okul öncesi eğitim kurumları ile dershanelerde, özel eğitim ve öğretim kurumları dahil olmak üzere ilk ve orta öğrenim kurumlarının kültür ve sosyal hizmet binalarının kapalı ve açık alanlarında sigara içilmesi yasaklandı. Yasa kahvehanelerde, lokanta ve kafeteryalarda ise 19 Temmuz 2009′da yürürlüğe girdi.

Bu tez çalışmasının amacı İnönü Üniversitesi İktisadi ve İdari Bilimler Fakültesi öğrencilerinin sigara içme alışkanlığını etkileyen faktörleri ortaya koymakla birlikte, gençlerin son zamanlarda ülkemizde uygulanmaya başlanan sigara yasağına olan bakış açılarını ortaya çıkarmak ve bu yasağın sigara kullanımına olan etkisini değerlendirmektir. Bağımlı değişken iki şıklı kategorik bir değişken olduğundan, bu tür verilerin analizinde uygulanan lojistik regresyon analizi yardımıyla sigara içmede etkili olan en önemli faktörler belirlenmeye çalışılmıştır.

Sigara içmede etkili olduğu düşünülen bağımsız değişkenler aşağıdaki gibi belirlenmiştir:

bölümü

sınıfı

boy

kilo

cinsiyet

Beğitim= Babanın eğitim düzeyi

Aeğitim= Annenin eğitim düzeyi

Mdurum= Anne ve babanın medeni/sosyal durumu

Ksayısı= Kardeş sayısı

Tharçlık= Aylık ortalama toplam harçlık

Bşekli= Barınma şekli

Skullanma= Sigara kullanma durumu

Sbaşyaşı= Sigaraya başlama yaşı

Siçmesüresi= Sigara içme süresi

Smiktarı= Günde ortalama içilen sigara miktarı

Smasrafı= Ortalama aylık sigara masrafı

Sbırakma= Sigarayı bırakmayı deneme

Bsigara= Babanın sigara içme durumu

Asigara= Annenin sigara içme durumu

Ksigara= Kardeşin sigara içme durumu

Arkçevsigara= Arkadaş çevresinin sigara içme durumu

Alkkullanma= Alkol kullanma durumu

Alkkulnedeni= Alkol kullanma nedeni

Alkkulmamanedeni= Alkol kullanmama nedeni

Arkçevalkol= Arkadaş çevresinin alkol kullanma durumu

Sıkıntıstres= Sigaranın sıkıntı, stres ve yalnızlığı giderdiğinin düşünülmesi

Statü= Sigaranın insanların statüleri üzerindeki etkisi

Szammı= Sigaraya yapılan zamların kullanılan sigara markası üzerindeki etkisi

Syasağı= Kapalı ortamlar için getirilen sigara yasağının sigarayı bırakmak üzerine etkisi

Syasağınıdes= Kapalı ortamlar için getirilen sigara yasağını destekleme durumu

Bağımlı değişken y ise,

0= Sigara kullanmıyor

1= Sigara kullanıyor

şeklinde kodlanmıştır.

Yukarıdaki değişkenler ile ilgili veriler İnönü Üniversitesi İktisadi ve İdari Bilimler Fakültesinde okuyan 2281 öğrenciden 964 'üne anket uygulanarak elde edilmiştir.

Örneklem genişliğini doğru belirlemek araştırmanın kısa sürede tamamlanması ve doğru sonuçlar elde etme açısından oldukça önemlidir. Bu anket çalışması sırasında basit rassal örnekleme yapılmıştır. Elde edilen geçerli 964 anketin yeterli bir örneklem genişliği olduğunu Çıngı, 1994:60 referans alınarak, $\propto = 0{,}01$ ve $P = 0{,}50$ için 0,05 hoşgörü miktarına göre yeterli bulunmuştur.

Sorulan sorulara verilen cevaplar ile ilgili elde edilen sonuçlar şu şekildedir:

Tablo-5: Öğrencinin Cinsiyeti

Cinsiyet	Frekans	Yüzde
☐ Kız	461	47,8
☐ Erkek	503	52,2

Tablo-6: Öğrencinin Boyu

Boy	Ortalama
☐ Kız	1,64
☐ Erkek	1,76

Tablo-7: Öğrencinin Kilosu

Kilo	Ortalama
☐ Kız	55
☐ Erkek	71

Demografik bilgilere göre ankete katılan öğrencilerin % 52,2'si erkek, % 47,8'i kız'dır. Erkeklerin boy ortalamaları 1,76 ve kilo ortalamaları 71, kızların boy ortalamaları 1,64 ve kilo ortalamaları ise 55 olarak belirlenmiştir.

Ankete katılan öğrencilerden %30,1'i birinci, %31,1'i ikinci, %20,9'u üçüncü ve % 17,9'u da dördüncü sınıfta okumaktadır. Yine bu öğrencilerin %11,7'si ekonometri, %29,4'ü iktisat, %27,8'i kamu yönetimi, %28,9'u işletme, %2,2'si uluslararası ilişkiler bölümü öğrencileridir.

Tablo-8: Babanın Eğitim Düzeyi

Babanın eğitim düzeyi	Frekans	Yüzde
☐ Lise	300	31,1
☐ İlkokul	294	30,5
☐ Ortaokul	169	17,5
☐ Üniversite	163	16,9
☐ Yok	38	3,9

Ankete katılan öğrencilerin %31,1'inin babası lise mezunu iken bunu %30,5 ile ilkokul mezunu olanlar izlemektedir. Bu iki sonucun birbirine olan yakınlığı dikkat çekicidir. Babası üniversite mezunu olanların oranının %16,9 ve babası ortaokul mezunu olanların oranının ise %17,5 olduğu görülmüştür. Yine bu iki sonucun birbirine olan yakınlığı dikkat çekicidir.

Tablo-9: Annenin Eğitim Düzeyi

Annenin eğitim düzeyi	Frekans	Yüzde
☐ İlkokul	425	44,1
☐ Yok	211	21,6
☐ Ortaokul	151	15,7
☐ Lise	136	14,1
☐ Üniversite	41	4,3

Ankete katılan öğrencilerin %44,1'inin annesi ilkokul mezunu iken %4,3'ünün annesi üniversite mezunudur. Öğrencilerin %21,9'unun annesi ise eğitim görmemiştir.

Tablo-10: Anne ve Babanın Medeni/Sosyal Durumu

Anne ve babanın medeni/sosyal durumu	Frekans	Yüzde
☐ Birlikte yaşıyorlar.	880	91,3
☐ Babam öldü.	45	4,7
☐ Boşandılar, ayrı yaşıyorlar.	22	2.3
☐ Annem öldü.	15	1,6
☐ Üvey annem / babam var	2	0,2

Ankete katılan öğrencilerin % 91,3'ünün anne ve babası birlikte yaşıyor. Bu da gençlerin çok büyük bir oranla aile bütünlüğü parçalanmamış ortamlarda yetiştiğini göstermektedir. Öğrencilerin %4,7'sinin ise babası ölmüştür.

Tablo-11: Kardeş Sayısı

Kardeş sayısı	Frekans	Yüzde
☐ **Evet 1-2**	365	37,9
☐ **Evet 2-4**	286	29,7
☐ **Evet 4-6**	161	16,7
☐ **Evet 6'dan fazla**	137	14,2
☐ **Hayır**	15	1,6

Ankete katılan öğrencilerin % 37.9'u 1 ya da 2 kardeşe sahiptir. Kardeş sayısı arttıkça oranların düşmesi dikkat çekicidir.

Tablo-12: Barınma Şekli

Barınma şekli	Frekans	Yüzde
☐ **Ev**	490	50,8
☐ **Devlet Yurdu**	264	24,7
☐ **Evde ailemle birlikte**	119	12,3
☐ **Özel Yurt**	91	9,4

Ankete katılan öğrencilerin % 50,8'i evde kalırken bunu %27,4 ile devlet yurdunda kalanlar izlemektedir. Bu sonuçlara bakılarak ailesiyle birlikte yaşayan öğrenciler hariç büyük çoğunluğunun ev yerine yurtta kalmayı tercih ettiği görülmektedir.

Tablo-13: Sigara İçme Alışkanlığı

Sigara içme alışkanlığı	Frekans	Yüzde
☐ Hiç içmedim.	393	40,8
☐ Düzenli içiyorum.	208	21,6
☐ Bir kez denedim, hiç sevmedim.	180	18,7
☐ Ara sıra içiyorum.	137	14,2
☐ İçiyordum, bıraktı.	45	4,7

Ankete katılan öğrencilerin % 40,8'i hiç sigara kullanmamışken %21,6'sı düzenli olarak sigara içmektedir. Yine bu öğrencilerin %4,7'si ise sigarayı bırakmıştır.

Tablo-14: Sigarayı Deneme ya da Sigaraya Başlama Nedeni

Sigarayı deneme ya da sigaraya başlama nedeni	Frekans	Yüzde
☐ Sigara içmeyenler	618	64,1
☐ Stresten	118	12,2
☐ Keyif verdiği için	107	11,1
☐ Özendim	55	5,7
☐ Merak ettim	43	4,5
☐ Arkadaşlarımın ısrarı ile	24	2,5

Sigara kullanan öğrencilerin %12,2'si sigarayı stresten, %11,1'i ise sigarayı keyif verdiği için kullanıyor. Arkadaş çevresinin etkisiyle sigarayı deneyen ya da sigaraya başlayanların % 2,5 ile en düşük yüzdeye sahip olması dikkat çekicidir.

Tablo-15: Sigaraya Başlama Zamanı

Sigaraya başlama zamanı	Frekans	Yüzde
☐ Sigara içmeyenler	618	64,1
☐ 15-20 Yaş arası	228	23,7
☐ 10-15 Yaş arası	62	6,4
☐ 20-25 Yaş arası	53	5,5
☐ 25 Yaş ve üstü	3	0,3

Sigara kullanan öğrencilerin % 23,7'si sigaraya 15-20 yaş arası başlamış.

Tablo-16: Sigarayı Bırakmayı Deneme

Sigarayı bırakmayı deneme	Frekans	Yüzde
☐ Sigara içmeyenler	618	64,1
☐ Hayır	175	18,2
☐ Evet	171	17,7

Sigara içen öğrencilerin % 18,2'si sigarayı bırakmayı denememiş, %17,7 ise en az bir kere sigarayı bırakmayı denemiştir.

Tablo-17: Aile Üyelerinizin Sigara İçme Alışkanlıkları

Aile üyelerinizin sigara içme alışkanlıkları	Frekans	Yüzde
Anne		
☐ İçmiyor	767	79,6
☐ içiyor	150	15,6
☐ içiyordu, bıraktı	40	4,1
☐ Anne yok	7	0,7
Baba		
☐ İçiyor	418	43,4
☐ içmiyor	353	36,6
☐ içiyordu, bıraktı	160	16,6
☐ Baba yok	33	3,4

Ankete katılan öğrencilerin % 79,6'sının annesi sigara kullanmazken %15,6'sının annesi sigara içiyor. Yine bu öğrencilerin % 43,4'ünün babası sigara içerken %36,6'sının babası sigara kullanmıyor.

Tablo-18: Arkadaş Çevrenizin Sigara İçme Alışkanlığı

Arkadaş çevrenizin sigara içme alışkanlığı	Frekans	Yüzde
☐ Genelde içiyor	539	55,9
☐ Genelde içmiyor	414	42,9
☐ Genelde bıraktı.	11	1,1

Ankete katılan öğrencilerin %55,9'unun arkadaş çevresi sigara içerken %42,9'unun arkadaş çevresi sigara kullanmıyor. Bu iki sonucun birbirine olan yakınlığı yine dikkat çekicidir.

Tablo-19: Alkol Kullanma Durumu

Alkol kullanma durumu	Frekans	Yüzde
☐ Hiç içmedim.	585	60,7
☐ Ara sıra içiyorum.	213	22,1
☐ Bir kez denedim, hiç sevmedim.	117	12,1
☐ Düzenli içiyorum.	33	3,4
☐ İçiyordum, bıraktım	16	1,7

Ankete katılan öğrencilerin %60,7'si hiç alkol kullanmazken, %22,1'i ara sıra da olsa alkol almaktadır. Düzenli olarak alkol kullananların oranı ise % 3,4 ile en düşük sırada yer almaktadır.

Tablo-20: Alkol Kullanma Nedeni

Alkol kullanma nedeni	Frekans	Yüzde
☐ Alkol kullanmayanlar	706	73,2
☐ Keyif verdiği için	161	16,7
☐ Merak ettim.	34	3,5
☐ Stresten	33	3,4
☐ Özendim.	16	1,7
☐ Arkadaşlarımın ısrarı ile.	14	1,5

Alkol kullanan öğrencilerin %16,7'si alkolü keyif verdiği için kullanıyor. Yine arkadaş çevresinin ısrarı ile alkol kullananların oranı %1,5 ile dikkat çekicidir.

Tablo-21: Alkol Kullanmama Nedeni

Alkol kullanmama nedeni	Frekans	Yüzde
☐ Alkol kullananlar	248	25,7
☐ Dini görüşümden dolayı	367	38,1
☐ Sağlığımı olumsuz yönde etkilediğinden dolayı	122	12,7
☐ Sevmediğimden dolayı	227	23,5

Ankete katılan alkol kullanmayan öğrencilerin %38,1'i dini görüşünden ötürü alkol kullanmazken %23,5'i sevmediğinden dolayı alkol kullanmıyor.

Tablo-22: Arkadaş Çevresinin Alkol Kullanma Alışkanlığı

Arkadaş çevresinin alkol kullanma alışkanlığı	Frekans	Yüzde
☐ Genelde içmiyor	709	73,5
☐ Genelde içiyor	241	25,0
☐ Genelde içiyordu, bıraktı	14	1,5

Ankete katılan öğrencilerin % 73,5'inin arkadaş çevresi alkol kullanmıyor.

Tablo-23: Sigaranın Sıkıntı Stres ve Yalnızlığı Giderdiğinin Düşünülmesi

Sigaranın sıkıntı stres ve yalnızlığı giderdiğinin düşünülmesi	Frekans	Yüzde
☐ Hayır	701	72,7
☐ Evet	263	27,3

Ankete katılan öğrencilerin %72,7'si sigaranın sıkıntı, stres ve yalnızlığı giderdiğini düşünmezken %27,3'ü sigaranın sıkıntı, stres ve yalnızlığı giderdiğini düşünmektedir.

Tablo-24: Sigara İçmenin İnsanların Statülerinin Üzerinde Ne Tür Etki Yaptığı

Sigara içmenin insanların statülerinin üzerinde ne tür etki yaptığı	Frekans	Yüzde
☐ Olumsuz etki	535	55,5
☐ Hiçbir etkisi yok	374	38,8
☐ Olumlu etki	55	5,7

Ankete katılan öğrencilerin % 55,5'i sigaranın statü üzerinde olumsuz etkisi olduğunu düşünürken, %38,8'i sigaranın statü üzerinde hiçbir etkisinin olmadığını düşünüyor.

Tablo-25: Yapılan Zamların Kullanılan Sigara Markasında Değişikliğe Neden Olması

Yapılan zamların kullanılan sigara markasında değişikliğe neden olması	Frekans	Yüzde
☐ Sigara içmeyenler	238	64,1
☐ Hayır	618	24,7
☐ Evet	108	11,2

Sigara kullanan öğrencilerin %11,3'si sigara yapılan zamlar nedeniyle kullandığı sigara markasında değişikliğe gidiyor. %25,7'si ise böyle bir değişiklik yapmıyor.

Tablo-26: Kapalı Ortamlar İçin Getirilen Sigara Yasağının Sigarayı Bırakmak Üzerine Etkisi

Kapalı ortamlar için getirilen sigara yasağının sigarayı bırakmak üzerine etkisi	Frekans	Yüzde
☐ Evet	574	59,5
☐ Hayır	390	40,5

Ankete katılan öğrencilerin %59,5'i kapalı ortamlara getirilen sigara içme yasağının sigarayı bırakmak üzerinde olumlu bir etkisi olduğunu düşünüyor. %40,5 ise bu görüşe katılmıyor.

Tablo-27: Kapalı Ortamlar İçin Getirilen Sigara Yasağını Desteklenmesi

Kapalı ortamlar için getirilen sigara yasağını desteklenmesi	Frekans	Yüzde
☐ **Evet**	794	82,4
☐ **Hayır**	170	17,6

Ankete katılan öğrencilerin %84,4'ü bu kapalı ortamlar için getirilen sigara yasağını destekliyor. %17,6'sı ise sigara yasağını desteklemiyor.

Bir ankette güvenilirlik analizi bütün soruların birbirleriyle tutarlılığını, ele alınan oluşumu ölçmede türdeşliğini ortaya koyan önemli bir kavramdır. Güvenilirlik, araştırmaların tutarlılığını ve etkinliğini bazı güvenilirlik testlerinin sonuçlarına göre değerlendirmektir. En çok kullanılan testler, Cronbach Alpha, İkiye Bölme (split), Paralel, Mutlak Kesin Paralel (strict) olarak sayılabilir (Acı ve Sezgin, 2007).

Cronbach Alfa katsayısı istatistik temelleri tutarlı ve tüm soruları dikkate alarak hesaplandığından genel güvenirlik yapısını diğer katsayılara göre en iyi yansıtan katsayıdır.

Alfa katsayısının değerlendirilmesinde uyulan değerlendirme kriteri;

$0.00 \leq \alpha < 0.40$ *ise ölçek güvenilir değil*dir.

$0.40 \leq \alpha < 0.60$ *ise ölçek düşük güvenirlikte*dir.

$0.60 \leq \alpha < 0.80$ *ise ölçek oldukça güvenilir*.

$0.80 \leq \alpha < 0.100$ *ise ölçek yüksek derecede güvenilir*.

şeklindedir (Özdamar, 1997:500).

Uygulanan anketin tüm soruları göz önüne alınarak güvenirlik analizi yapıldığında $\alpha = 0{,}7420$ olarak hesaplanmıştır. Bu sonuçta anketin güvenirliğinin oldukça yüksek olduğunu göstermektedir.

5.1. Lojistik Regresyon Modellerinin R Kullanılarak Kurulması ve Analizi

R, Becker and Chembers tarafından geliştirilen S dilinin bir çeşididir. R programı temelde bir veri işleme ve grafik çizme programıdır ve kod yazımı tabanlıdır. R yazılımı kullanılarak istatistiksel analiz, grafik çizme ve veri işleme işlemleri yapılabilir.

Temel olarak yapılması gereken komutların yazılarak çıktıların görüntülenmesi işlemidir. Komutların girilebilmesi için kullanılan bölge “R Console” olarak adlandırılmıştır.

R yazılımının bazı temel komutları şu şekildedir:

<-; elde edilen veriler bir değişkene atanmak istenirse bu atama işlemi için “<-“ işlem yapısı kullanılır.

help(); R ile çalışırken herhangi bir fonksiyon ile ilgili yardım alınmak istendiğinde kullanılan komuttur.

c(); R yazılımının en büyük özelliklerinden biriside değişkenler ile çalışırken vektör ve matris kullanımına izin vermesidir. Bir vektörü yaratmak içinde kullanılan fonksiyon da c() fonksiyonudur. Bu fonksiyon aynı zamanda birden fazla vektörün tek bir vektör olarak birleştirilmesinde veya karakter değişkeninin sayılarla birleştirilmesinde de kullanılmaktadır.

length(); önceden tanımlanmış olan bir vektörün birim sayısını öğrenmek için kullanılan fonksiyondur.

seq(); Belirli bir düzene sahip vektör yaratılmak istendiğinde kullanılan fonksiyondur.

rep(); İlgilenilen değişkenin her seviyesi için birim sayısı kadar isim girilmesi gerektiğinde kullanılan fonksiyondur.

matrix(); Matris oluşturulması fonksiyonu ile sağlanır.

mantık operatörleri; Doğru(T) ve Yanlış(F) olmak üzere iki mantıksal değer vardır. Mantık operatörleri karşılaştırma yaparken ve vektörler ile matrislerin belirli elemanlarını belirlerken çok kullanılırlar. R yazılımı mantık operatörlerinin yanında mantık fonksiyonları da sunmaktadır. Bu fonksiyonlar yardımıyla ilgilenilen değişkenin bir karakter değişkeni mi yoksa sayısal bir değişken mi olduğu anlaşılabilir.

List nesneleri; istatistiksel analizler için oluşturulan farklı nesnelerin bir araya getirilmesinde List nesnelerinden faydalanılır.

data.frame; R yazılımında veri seti içerisindeki “faktör listeleri ve gözlem birimleri” data.frame olarak bir araya getirilirler.

mean(), sqrt(), sd(), var(), quantile(), median(); Değişkenin ortalamasının, karekökünün, standart hatasının, varyansının, kantillerinin ve medyanının hesaplanmasında kullanılan fonksiyonlardır. Tüm aritmetik işlemler bu şekilde kolaylıkla yapılabilir.

summary(); Bir veri setindeki tüm değişkenlerin özet istatistiklerini veren fonksiyondur. Bu özet istatistikler sırasıyla değişkene ait minimum değer, 1. kantil, medyan, aritmetik ortalama, 3. kantil, maksimum değer ve eksik gözlem sayısı olarak R yazılımı tarafından verilir.

plot(); R yazılımında saçılım grafiği çizmekte kullanılan fonksiyondur.

hist(); R yazılımında histogram grafiği çizmekte kullanılan fonksiyondur.

stem(); R yazılımında dal-yaprak grafiği çizmekte kullanılan fonksiyondur.

barplot(); R yazılımında bar grafiği çizmekte kullanılan fonksiyondur.

boxplot(); R yazılımında kutu grafiği çizmekte kullanılan fonksiyondur.

qqnorm(); Bu grafik veri setinin ilgili dağılıma sahip ana kütleden veya herhangi iki veri setinin aynı dağılıma sahip bir ana kütleden gelip gelmediğini test etmek için kullanılır.

pie();R yazılımında pasta grafiği çizmekte kullanılan fonksiyondur.

sample(); Rassal örneklem seçme işlemi için kullanılan fonksiyondur.

permn(); Permütasyon işlemi için kullanılan fonksiyondur.

choose(); Kombinasyon işlemi için kullanılan fonksiyondur.

combn(); Belirli birim sayısı için oluşturulacak bütün alt örneklemlerin gösterimini sağlamasıdır (Sönmez, 2006).

GLM normal-olmayan yanıt dağılımını doğrusal modellere uyumlu hale getiren aynı zamanda açık ve kesin bir biçimde doğrusallaştıran bir gelişmedir (Mc Cullagh ve Nelder, 1989).

Regresyon analizinde GLM için kullanılan R fonksiyonlarından bazıları şunlardır:

family; Aile nesnesi GLM gibi fonksiyonlar tarafından kullanılan modellerin detaylarının özelleştirilmesine elverişli bir yol sağlar. R'nin sağladığı olanaklar GLM sınıfının gaussian, binomal, poisson, ters gaussian ve gamma yanıt dağılımlarını ayrıca yanıt dağılımının tam olarak özelleştirilmediği quasi-likelihood modelini de kapsar. Son durumda varyans fonksiyonu ortalamanın bir fonksiyonu olarak özelleştirilmelidir, fakat diğer durumlarda bu fonksiyon yanıt dağılımı tarafından uygulanmalıdır. Her bir yanıt dağılımı doğrusal tahminci ile ortalamaya bağlanmak için çeşitli bağlantı fonksiyonlarını içerir. Yanıt dağılımının bileşimi, bir bağlantı fonksiyonu ve bilginin diğer çeşitli diğer parçalarından oluşur. Bu iki bileşene GLM'nin ailesi olarak bilinen model tatbiki yürütmek için ihtiyaç duyulur.

Family fonksiyonu için kullanılan argümanlar:

name: Ailenin ismidir.

link: Tersi ve türemişi bir dizin ile link fonksiyonudur.

variance: varyans ve sapma fonksiyonları ile bir dizindir.

Family name	**Link functions**
`binomial`	`logit, probit, log, cloglog`
`gaussian`	`identity, log, inverse`
`Gamma`	`identity, inverse, log`
`inverse.gaussian`	`1/mu^2, identity, inverse, log`
`poisson`	`identity, log, sqrt`
`quasi`	`logit, probit, cloglog, identity, inverse, log, 1/mu^2, sqrt`

glm.nb: Uygun negatif binomial bir genelleştirilmiş doğrusal modeldir.

glm: Doğrusal tahmincinin sembolik tanımlayıcısı ve hata dağılımının bir tanımlayıcısı olarak verilen genelleştirilmiş doğrusal modeli düzeltmek için kullanılır.

polr: Orantılı odds lojistik regresyonudur.

Deviance: Uygun model nesnesinin sapmasıdır.

AIC: Sapmaya alternatif olarak ve uygun olarak kullanılan kriterdir.

coef: Model fonksiyonlar tarafından elde edilen model katsayılarıdır.

predict: Lineer model tahmin değerleridir.

residuals: Model fonksiyonlar tarafından elde edilen model kalıntılarıdır.

gfit: Modelin uyum iyiliğini gösterir.

leverage.plots: Leverage regresyon grafiğidir.

influence.plot: İnfluence regresyon grafiğidir (Ricci, 2005).

Uygun temel GLM fonksiyonu için temel argümanlar şu şekildedir:

$glm(formula, family, data, weights, control, subset)$

formula: Doğrusal kestiricidir.

control: İteratif süreçtir.

family: Ek bilgi ile aile ismini verir. Örneğin, uygun binomial yanıt değişkeni ile probit link gibi.

The binomial family:

Uygun binomial model yanıt için GLM'de 3 olasılık kullanıyor:

- Eğer yanıt bir vektör ise bunun binary veri olduğu kabul edilir ve bu yüzden vektör 0/1 olmak zorundadır.
- Eğer yanıt iki-sütun matrisi ise birinci sütunun denemeler için başarı sayısı olduğu kabul edilir ve ikinci sütun da başarısızlık sayısına dayanır.
- Eğer yanıt bir faktör ise ilk seviye başarısızlık olarak (0) alınır ve diğer tüm seviyeler başarı(1) olarak alınır (McCullagh & Nelder , 1989).

5.2. Lojistik Regresyon Modeli İçin R'de Bir Uygulama

Model oluşturulurken göz önünde bulundurulan en temel kural, en az sayıda bağımsız değişken yardımıyla bağımlı değişkendeki değişimi en fazla biçimde açıklayacak bir model kurmaktır. Eğer modele gereksiz değişkenler ilave edilirse standart hata tahminleri büyüyecek ve aynı zamanda modeli tahmin etme süreci karmaşıklaşacaktır. Yine modeli büyük ölçüde etkileyecek bir değişkeni model dışında bırakmamız modeli hatalı kurmamıza neden olacaktır. Tüm bu sebeplerden dolayı modeli kurarken değişken seçimine çok dikkat etmeliyiz.

Lojistik regresyon modelinde değişken seçimi yaparken kullanılan yöntemler temelde tek değişkenli ve çok değişkenli analiz olmak üzere 2'ye ayrılır. Çok değişkenli analizde kendi içerisinde adımsal (stepwise) yöntem ve en iyi alt setler yöntemi olmak üzere ikiye ayrılmaktadır. En iyi alt setler yöntemi lojistik regresyonda çok tercih edilmemektedir.

Adımsal yöntem ise yine kendi içerisinde ileriye doğru seçim (forward selection) ve geriye doğru eleme (backward elimination) olmak üzere ikiye ayrılır (Oğuzlar, 2005).

Geriye doğru eleme yönteminde ise işleme tüm bağımsız değişkenlerin olduğu model ile başlanır ve modele katkısının en az olan değişkenler teker teker modelden çıkarılır. Modele katkısı en az olan değişkenler belirlenirken seçilen α yanılma düzeyinde kısmi F istatistiği en küçük olan seçilir ve modelden çıkarılır. (Alpar, 2003:346)

İleriye doğru seçim yönteminde değişken seçme işlemine, modelde sadece sabit terimin bulunduğu bir denklemle başlanır ve değişkenler modele yeni eklenen bir değişkenin modele katkısı olmayana kadar eklenir ve her bir adımda modele katkısı en çok olan değişken seçilir. Modele katkısı en çok olan bağımsız değişkeni belirlerken de y bağımlı değişkeni ile en yüksek korelasyona sahip (negatif ya da pozitif) bağımsız değişken hangisi ise o seçilir. Modele ilk olarak alınan değişken aynı zamanda y bağımlı değişkeni ile en yüksek F istatistiğine sahip değişkendir. Belirlenen α yanılma düzeyinde hesaplanan F istatistiği anlamlı ise bu değişken modele alınır ve ileriye doğru seçim işlemi devam eder. Yapılan test sonucunda bu değişken modele alınmazsa seçim süreci sona erer. İlk değişkenin modele alınmasından sonra modele alınacak diğer değişkenler bağımlı değişkenle en yüksek kısmi F değerine sahip olan (ya da bağımlı değişkenle en yüksek kısmi korelasyon katsayına sahip olan) değişkenlerdir. İkinci değişkene ait kısmi F testi α yanılma düzeyinde anlamlı ise değişken modele alınır ve ileriye doğru seçme yöntemi devam eder. Yine aynı şekilde eğer F testi sonucu anlamlı bulunmaz ise seçim işlemi sona erer (Alpar, 2003:345).

Tezin uygulamasında kurulacak olan model lojistik regresyon modeli olduğundan yazılan komut da lojistik regresyona göre olmalıdır.

Lojistik regresyon için $family = binomial(link = logit)$ alındığında lojistik regresyon modeli oluşturulmuş olur.

Geriye doğru eleme yöntemi kullanılarak tüm değişkenler modele dahil edilmiş ve sırasıyla modele katkısı en az olan değişken modelden çıkarılmak suretiyle en uygun model oluşturulmuştur. Bu amaçla da ilk olarak tüm değişkenlerin oluşturduğu model,

```
model<-glm(Skullanma~cinsiyet+bölümü+sınıfı+boy+kilo+Beğitim+Aeğitim+
Mdurum+Ksayısı+Tharçlık+Bşekli+Skulnedeni+Sbaşyaşı+Siçmesüresi+Smiktarı+Sma
srafı+Sbırakma+Asigara+Bsigara+Ksigara+Arkçevsigara+Alkkullanma+Alkkulneden
```

i+Alkkulmamanedeni+Arkçevalkol+Sıkıntıstres+Statü+Szammı+Syasağı+Syasağınıd es,family='binomial')

şeklinde tanımlanmıştır.

Bu modelden elde edilen p olasılıklarına bakılarak olasılığı en yüksek değişken olan Ksigara modelden çıkarılmış ve model yeniden kurulmuştur. Aynı işlem tekrarlanarak modelden sırasıyla Ksayısı, Statü, Smasrafı, Tharçlık, kilo, Szammı, Sıkıntıstres, Syasağınıdes, Alkkullanma, Bsigara, Arkçevalkol, Skulnedeni, Alkkulmamanedeni, Alkkulnedeni, Bşekli, sınıfı, Smiktarı, Asigara, Aeğitim, Mdurum, Arkçevsigara, bölümü, Beğitim, Sbırakma, cinsiyet, boy değişkenleri çıkarılmıştır.

Sonuçta geriye kalan değişkenler ile en anlamlı model,

```
Model1<-glm(Skullanma ~ Sbaşyaşı + Siçmesüresi + Syasağı, family = "binomial")
```

Tablo-28: Model1 için orijinal R çıktısı:

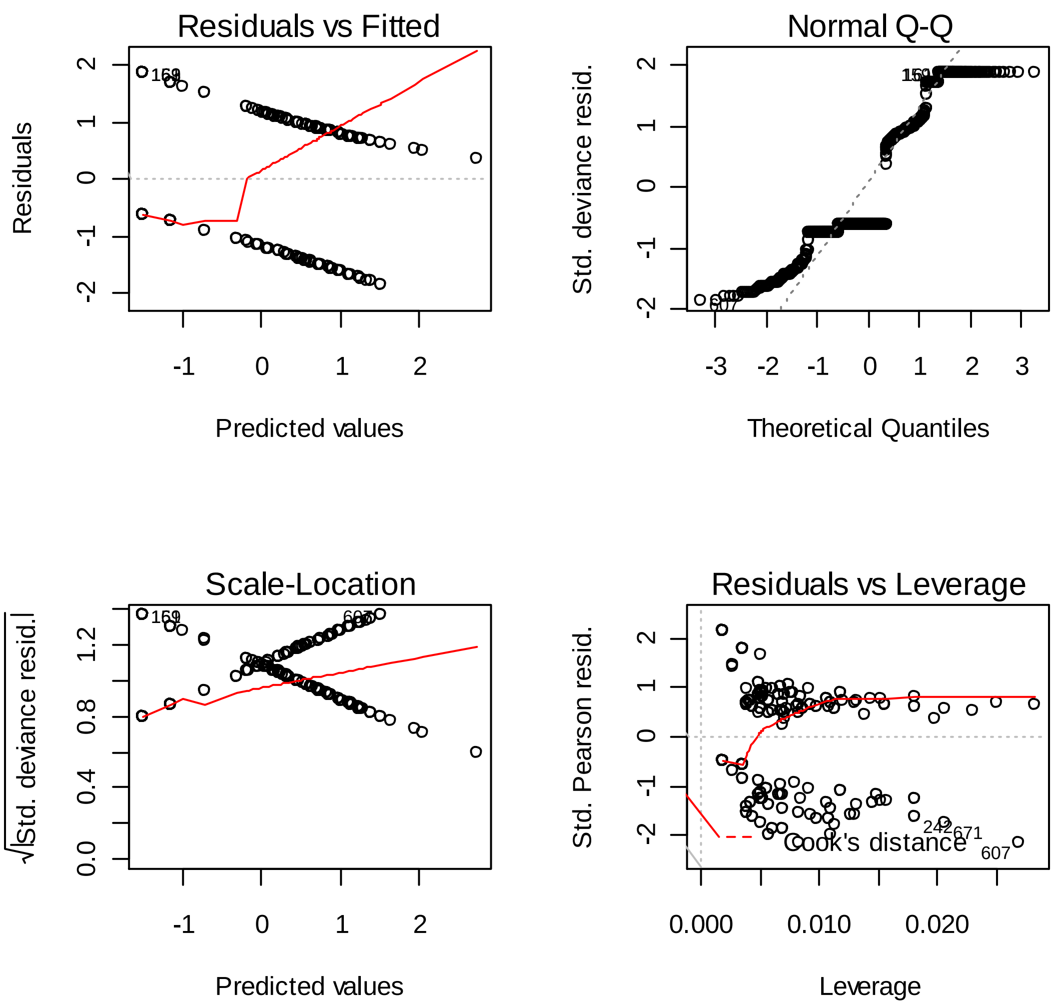
```
Call:

glm(formula = Skullanma ~ Sbaşyaşı + Siçmesüresi + Syasağı, family = "binomial")

Deviance Residuals:

    Min       1Q    Median       3Q      Max

-1.8491  -0.7389  -0.6275   0.8925   1.8558

Coefficients:

            Estimate  Std. Error z value Pr(>|z|)

(Intercept) -1.89148    0.23377  -8.091 5.92e-16 ***

Sbaşyaşı     0.67168    0.09001   7.462 8.51e-14 ***

Siçmesüresi  0.13253    0.03233   4.099 4.15e-05 ***

Syasağı      0.36639    0.15545   2.357   0.0184 *

Null deviance: 1258.6  on 963  degrees of freedom

Residual deviance: 1045.1  on 960  degrees of freedom

AIC: 1053.1
```

sonuçlarına da bakılarak,

```
Skullanma = -1.89148 + 0.67168 Sbaşyaşı + 0.13253 Siçmesüresi + 0.36639 Syasağı
```

şeklinde tanımlanır.

Kurduğumuz modelden de görüldüğü üzere sigara kullanımını etkileyen sigaraya başlama yaşı, sigara içme süresi ve sigara yasağının sigarayı bırakma üzerine etkisi değişkenlerinin hepsi en az yüzde 5 anlamlılık seviyesinde anlamlı bulunmuşlardır. Deviance (sapma), modelin uyum iyiliğinin ölçüsü olarak da kullanılmaktadır. Kurulan modellerde yer alan Residual Deviance değeri kurulan geçerli modelin sapma değerini ifade ederken, Null Deviance değeri modelde yalnızca sabit terimin olduğu değişkenlerin olmadığı sapma değerini ifade eder. Residual Deviance değeri Null Deviance değerinden daha düşük olmalıdır. Çünkü modele anlamlı değişkenler eklendikçe sapma değerinde düşüş olmalıdır. Aynı zamanda yine AIC değeri de iki modelin karşılaştırılmasında ve en iyi modelin belirlenmesinde önemlidir. AIC değeri düşük olan model uygun model olarak seçilir.



(Everitt ve Torsten).

Değişkenlerin tek tek veya birkaçının sigara kullanımını üzerine etkisini görmek için birçok model kurmak mümkündür. Sigara kullanımını büyük ölçüde etkileyeceği düşünülen değişkenleri seçerek aşağıdaki modelleri kurabiliriz:

```
Model2<-glm(Skullanma~cinsiyet,family='binomial')
```

Tablo-29: Model2 için orijinal R çıktısı:

```
Call:
glm(formula = Skullanma ~ cinsiyet, family = "binomial")
Deviance Residuals:
    Min       1Q     Median       3Q      Max
-1.0890    -1.0890  -0.7803    1.2685   1.6356
Coefficients:
            Estimate   Std. Error z value  Pr(>|z|)
(Intercept)   0.6101     0.2083    2.930  0.00339 **
cinsiyet     -0.8216     0.1387   -5.922  3.17e-09 ***
Null deviance: 1258.6  on 963  degrees of freedom
Residual deviance: 1222.4  on 962  degrees of freedom
AIC: 1226.4
```

Bu modelde cinsiyetin sigara kullanımı üzerine etkisinin oldukça anlamlı olduğunu görüyoruz. AIC ve Residual Deviance değerlerine baktığımızda ilk kurduğumuz modele göre bu değerlerin daha yüksek çıktığını görüyoruz. Bu da kurulan ilk modelin bu modele göre daha iyi bir model olduğunu gösteriyor. Zaten geriye doğru eleme yöntemiyle tüm değişkenlerden anlamsız olanları çıkararak elde ettiğimiz model kurabileceğimiz modeller içinde en iyisi olacaktır.

Model3<-glm(formula = Skullanma ~ Smasrafı + Szammı,family = "binomial")

Tablo-30: Model3 için orijinal R çıktısı:

```
Call:

glm(formula = Skullanma ~ Smasrafı + Szammı, family =
"binomial")

Deviance Residuals:

    Min       1Q   Median       3Q      Max

-2.0798  -0.6802  -0.6802   0.9584   1.7762

Coefficients:

             Estimate Std. Error z value Pr(>|z|)

(Intercept) -1.346057   0.097860 -13.755  < 2e-16 ***

Smasrafı     0.007701   0.001945   3.958 7.55e-05 ***

Szammı       0.730760   0.111616   6.547 5.87e-11 ***

Null deviance: 1258.6  on 963  degrees of freedom

Residual deviance: 1070.1  on 961  degrees of freedom

AIC: 1076.1
```

Kurulan bu üçüncü modelde de sigaraya yapılan zamlar ile aylık ortalama sigara masrafının sigarayı kullanmak üzerine etkisinin oldukça anlamlı olduğunu görüyoruz. Model 3'ün AIC ve Redidual Deviance değerlerinin Model 2'ye göre daha düşük çıktığını görmekteyiz. Bu da sigaraya yapılan zamların ve aylık ortalama sigara masrafının sigarayı kullanmak üzerine etkisinin cinsiyetin tek başına yaptığı etkiden daha fazladır.

Model4<-glm(Skullanma~Arkçevsigara,family='binomial')

Tablo-31: Model4 için orijinal R çıktısı:

```
Call:

glm(formula = Skullanma ~ Arkçevsigara, family = 
"binomial")

Deviance Residuals:

    Min       1Q    Median        3Q       Max

-1.0690  -1.0690  -0.7852     1.2899   1.9664

Coefficients:

             Estimate Std. Error  z value   Pr(>|z|)

(Intercept)    0.4978     0.2014   2.471   0.0135 *

Arkçevsigara  -0.7583     0.1365  -5.557   2.75e-08 ***

Null deviance: 1258.6  on 963  degrees of freedom

Residual deviance: 1226.2  on 962  degrees of freedom

AIC: 1230.2
```

Kurulan bu dördüncü modelde ise arkadaş çevresinin sigara kullanma durumunun sigarayı kullanmak üzerine etkisi oldukça anlamlı bulunmuştur.

Model5<-glm(Skullanma~Asigara,family='binomial')

Tablo-32: Model5 için orijinal R çıktısı:

```
Call:
glm(formula = Skullanma ~ Asigara, family = binomial)
Deviance Residuals:
    Min      1Q     Median     3Q      Max
-1.1254  -0.9304  -0.9304   1.4464   1.5566
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.1235     0.2789  -0.443   0.6579
Asigara      -0.2449     0.1456  -1.681   0.0927 .
Null deviance: 1258.6  on 963  degrees of freedom
Residual deviance: 1255.8  on 962  degrees of freedom
AIC: 1259.8
```

Bir önceki model gibi bu modelde de annenin sigara içme alışkanlığının tek başına sigara kullanımı üzerine etkisine bakılmış ve yüzde 5 anlamlılık düzeyinde anlamlı bulunmuştur. Aynı şekilde babanın sigara içme durumunu da bakılmış fakat sonuç anlamsız bulunmuştur.

Model6<-glm(Skullanma~Alkkulnedeni+Arkçevalkol,family='binomial')

Tablo-33: Model6 için orijinal R çıktısı:

```
Call:
glm(formula = Skullanma ~ Alkkulnedeni + Arkçevalkol, family
= binomial)


Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.2836  -0.8584  -0.8584   1.3273   1.6764
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.17878    0.31076  -0.575 0.565102
Alkkulnedeni  0.14800    0.04338   3.412 0.000645 ***
Arkçevalkol  -0.31498    0.16126  -1.953 0.050782 .
Null deviance: 1258.6  on 963  degrees of freedom
Residual deviance: 1232.0  on 961  degrees of freedom
AIC: 1238.0
```

Kurulan bu modelde ise alkol kullanma nedeni ve arkadaş çevresi alkol kullanımının sigara kullanımı üzerinde anlamlı bir şekilde belirleyici olduğunu görmekteyiz. Özellikle öğrencinin alkol kullanma nedeni sigara kullanımını oldukça etkilemektedir.

Model7<-glm(Skullanma~Sbırakma+Syasağınıdes,family='binomial')

Tablo-34: Model7 için orijinal R çıktısı:

```
Call:
glm(formula = Skullanma ~ Syasağınıdes + Sbırakma,
family = "binomial")
Deviance Residuals:
    Min      1Q    Median      3Q      Max
-1.8043  -0.6816  -0.6816     0.8418   1.7740
Coefficients:
             Estimate  Std. Error z value   Pr(>|z|)
(Intercept)   -1.8952     0.2391  -7.926   2.27e-15 ***
Syasağınıdes   0.5540     0.2036   2.721   0.0065 **
Sbırakma       1.0983     0.1039  10.574   < 2e-16 ***
Null deviance: 1258.6  on 963  degrees of freedom
Residual deviance: 1065.6  on 961  degrees of freedom
AIC: 1071.6
```

Bu modelde sigarayı bırakmayı denemek ve sigara yasağını desteklemek değişkenlerinin sigara kullanımı üzerinde ki etkisi oldukça anlamlı bulunmuştur. AIC ve Residual Deviance değerleri Model 1 hariç diğer modellerden daha düşük bulunmuştur. Çıkan bu sonuçta kurulan bu modelin diğer modellere göre sigara kullanımı etkileyen faktörleri daha iyi belirlemiştir.

Aynı modeli bu kez binomial olarak değil de yine geriye doğru eleme yöntemiyle linear model olarak yeniden kuracak olursak;

Model8<-
lm(Skullanma~cinsiyet+boy+Sbaşyaşı+Siçmesüresi+Sbırakma+Syasağı)

Tablo-35: Model8 için orijinal R çıktısı:

```
Call:

lm(formula = Skullanma ~ cinsiyet + boy + Sbaşyaşı +
Siçmesüresi + Sbırakma + Syasağı)

Residuals:

   Min      1Q    Median     3Q      Max

-0.8151 -0.2229 -0.1654  0.3308  0.8974

Coefficients:

             Estimate Std. Error t value Pr(>|t|)

(Intercept)  1.003490   0.442843   2.266 0.023673 *

cinsiyet    -0.072092   0.039889  -1.807 0.071027 .

boy         -0.462374   0.235125  -1.967 0.049529 *

Sbaşyaşı     0.107650   0.026709   4.030 6.01e-05 ***

Siçmesüresi  0.023600   0.006851   3.445 0.000596 ***

Sbırakma     0.075585   0.038201   1.979 0.048143 *

Syasağı      0.066272   0.028752   2.305 0.021382 *

Multiple R-squared: 0.2266,

Adjusted R-squared: 0.2217

F-statistic: 46.72 on 6 and 957 DF,

p-value: < 2.2e-16
```

Kurulan doğrusal modelde de binomial modelde olduğu gibi sigaraya başlama yaşı, sigara içme süresi ve sigara yasağının sigarayı bırakmak üzerine etkisi anlamlı bulunmuştur.

Ayrıca bunlara ek olarak doğrusal modelde cinsiyet, boy ve sigarayı bırakmayı deneme değişkenleri de sigara kullanımı etkileyen faktörler arasında yer almıştır. Cinsiyet değişkeni yüzde on anlamlılık düzeyinde anlamlı bulunurken diğer tüm değişkenler yüzde beş anlamlılık düzeyinde anlamlı bulunmuştur.

SONUÇ VE ÖNERİLER

Tez çalışmasında, bağımlı değişkenin iki şıklı olduğu, bağımsız değişkenlerin ise kategorik ve sürekli olduğu durumlarda alternatif olarak geniş bir uygulama alanı bulan Lojistik Regresyon Analizi ve Poisson Regresyon Analizi detaylı bir şekilde incelenmiştir.

Son zamanlarda kapalı mekanlarda uygulanan sigara içme yasağı ve gençlerin sigara içme oranlarındaki artan hız dikkate alınarak, gençlerin sigara içmelerine neden olan faktörlerin belirlenebilmesi için ve uygulanan sigara yasağına olan bakış açılarını ölçmek amacıyla İnönü Üniversitesi İktisadi ve İdari Bilimler Fakültesi öğrencilerine bir anket uygulaması yapılmıştır.

Öğrencilerin anketlere verdiği cevaplara dayanarak elde edilen sonuçlara göre; öğrencilerin %21,6’sı düzenli olarak sigara içerken %60’ bir şekilde sigara kullanmıştır. Sigaraya başlama yaşının 15-20 yaş aralığında yüksek olması sigaraya başlama yaşının özellikle ergenlik dönemine denk geldiğini ve bu durumun son derece dikkat çekici olduğunu ortaya koymaktadır. Öğrencilerin arkadaş çevresinin yarıya yakın kısmı sigara kullanmaktadır. Sigara içenlerin sigarayı kullanmakta gösterdikleri en yüksek neden olarak stres, ikinci olarak da keyif öne çıkarken alkol kullanma nedeni olarak keyif öne çıkmaktadır. Alkol kullanmayanların kullanmama sebebinde ise dini görüş faktörü belirleyici olmaktadır. Öğrencilerin arkadaş çevrelerinin büyük kısmı alkol kullanmamaktadır. Sigara kullananların yarıya yakın bir kısmı sigarayı bırakmayı denemiştir. Öğrencilerin büyük kısmı sigaranın sıkıntı, stres ve yalnızlığı giderdiğini düşünmezken yine çok büyük bir kısmı sigaranın statü üzerinde olumsuz bir etkisi olduğunu düşünmektedir. Öğrencilerin %60’a yakını kapalı ortamlar için getirilen sigara yasağının sigarayı bırakmak üzerinde olumlu bir etki yarattığını düşünürken yine öğrencilerin %82,4 gibi çok yüksek bir oranı bu yasağı desteklemektedir. Bu da gösteriyor ki sigara kullananların da büyük bir kısmı bu yasağa destek vermektedir. Bu gelişme sigara ile mücadelede aşama kaydedildiğini ve gençlerin bu konuda gittikçe bilinçlendiğini ortaya koymaktadır.

Lojistik regresyon analizi sonuçlarına göre de İnönü Üniversitesi İktisadi ve İdari Bilimler Fakültesi öğrencilerinin sigara içmelerine neden olan en önemli faktörler kurulan birçok model arasından en uygunu seçilerek sigaraya başlama yaşı, sigara içme süresi ve sigara yasağının sigarayı bırakma üzerine etkisi olarak belirlenmiştir. Ayrıca değişkenlerin sigara kullanımı üzerine etkilerine tek tek bakıldığında cinsiyet, sigara masrafı ve sigaraya yapılan zamlar, arkadaş çevresinin ve annenin sigara içme durumu, alkol kullanma nedeni ve arkadaş çevresinin alkol kullanma durumu, sigarayı bırakmayı deneme ve sigara yasağının desteklenmesi faktörlerinin sigara kullanımı üzerinde etkili olduğu görülmüştür.

Modeli binomial olarak değil de doğrusal model olarak kurduğumuzda ise cinsiyet, boy, sigaraya başlama yaşı, sigara içme süresi, sigarayı bırakmayı deneme ve sigara yasağının sigarayı bırakmak üzerine etkisi faktörlerinin sigara kullanımı üzerinde etkili olduğu görülmüştür.

Cinsiyet değişkenin başlı başına sigara kullanımını etkilemesi oldukça dikkat çekicidir. Dünya Sağlık Örgütü, artan sigara tüketimine dikkat çekmek, tütün kullanımının sağlık etkilerini ve tütün kullanımını azaltmaya yönelik etkili politikaların altını çizmek amacıyla her yıl 31 Mayıs gününü "**Dünya Tütünsüz Günü**" olarak kutlamaktadır. DSÖ, bu seneki Dünya Tütünsüz Gününde "**Tütün Endüstrisinin Hedefi: Kadın**" temasıyla tütün ürünlerinin kadınlara yönelik pazarlama taktiklerine ve sigaranın kadınlar ve genç kızlar üzerindeki zararlı etkilerine özellikle dikkat çekmektedir. Kadınlar arasındaki tütün salgınını kontrol altına almak, kapsamlı tütün kontrol stratejilerinin önemli bir parçasıdır. Kadınlar, dünyada sigara içen bir milyar kişinin yaklaşık beşte birini oluşturmaktadır. Bununla birlikte, tütün kullanım yaygınlığı özellikle gelişmekte olan ülkelerde kadınlar arasında her geçen gün artmaktadır. Bu nedenle yeni müşterilere ihtiyaç duyan tütün endüstrisinin ana hedefi, tütün kullanımına bağlı hastalıklardan ölecek olan hali hazırdaki tütün kullanıcılarının yarısının yerine geçecek potansiyele sahip olan kadınlardır. Tütün endüstrisi bugüne kadar kadınların ve genç kızların özgür, bağımsız ve çekici olmaları gerektiğini vurgulamıştır. Günümüzde direkt ya da indirekt reklâmlarda da aynı temayı kullanarak kadınları hedef almaya devam etmektedirler. Sigara, bağımsızlık, zarafet, çekicilik, kilo kontrolü, kültür ve güçle bağdaştırılarak reklâmı yapılmakta ve kadınlar arasında sigara içiciliğini arttırmaya yönelik yeni pazarlama taktikleri geliştirilmektedir.

Elde edilen veriler doğrultusunda sigaraya başlama yaşının 15-20 yaş arasında olması ve modelimizde bulunan sigaraya başlama yaşı ve sigara içme süresinin de anlamlı bulunması gösteriyor ki sigara bağımlılığında genç yaşta sigarayla tanışmış olmak oldukça etkilidir. Sigara tiryakilerinin tamamına yakınının 20 yaşından bu nesneye bağımlı hale geldiği görülmektedir. Bu yaştan sonra sigara bağımlılığının gelişmesi ise çok nadir bir durumdur. Özellikle insan neden sigaraya başlar sorusunun yanıtını almak üzere yapılan anket sonuçları incelendiğinde bu durumun tamamen gençlik dönemine has psikolojik motivasyonlardan ibaret olduğu görülmektedir. Kendini ispatlama, arkadaşları içerinde yer edinme, otoriteye (anne-baba, okul vs) karşı gelme, yasağı delme arzusu, akranlarından görüp merak ve deneme dürtüsü, arkadaşlarının teklifini geri çevirememe, içmediği için alaya alınmaktan, aşağılanmaktan korkma, büyüklerine özenme, hayranlık duyduğu kişilere benzeme, gizli ya da açık sigara reklamlarından etkilenme gibi dile getirilen nedenlerin o dönem için ne denli önemli olduğu kabul edilebilir gerçeklerdir.

Bu gerçekleri en iyi bilen sigara sektörüdür. Onlar için kendilerini besleyecek ölçüde geniş bir tiryaki pazarı oluşturmak için fazla zamanlarının olmadığını ve ne yapıp edilip 10-20 yaş arasında gençlerin sigara müptelası haline getirilmeleri gerektiğini çok iyi bilmektedirler. Bir günde dünyada imal edilen sigara sayısı, dünya nüfusunun iki katıdır. Dünyada bir yılda sigara satın almak için ödenen para 300 milyar dolar tutmaktadır. Bu kadar karlı bir pazar, ancak yeni tiryakiler oluşturmakla korunabilir (Özlü, 2000:77).

Bunun yanı sıra pasif içicilik ile istemleri dışında sigara içmek zorunda bırakılan bir kitle mevcuttur. Bu kitlenin de büyük çoğunluğunu kadınlar ve çocuklar oluşturmaktadır. Anne, baba vb büyüklerin sınır tanımayan sigara içme tutkuları karşısında, çocuklar ve gençler çoğu zaman kendilerini savunma şansına sahip değillerdir. Özellikle gençlerin sıklıkla gittikleri kafe, kahvehane, gazino, bar vb ortamlarda pasif içiciliğe maruz kalmaları ilerki yaşlarda aktif içicilik ve tiryakiliği geliştirmektedir. Çevresel sigara dumanına devamlı ve uzun süreli maruziyetin, aktif içicilikle benzer riskler taşıdığı ise artık çok iyi bilinmektedir (Özlü, 2000:80).

Tüm bu nedenlerden dolayı ülkemizde sigarayla mücadele konusunda son bir buçuk yılda önemli gelişmeler olmuştur. 4207 sayılı kanuna göre sigara reklamları yasaklanmıştır. Tütün üretici firmaları, isim, marka ve amblemler ile reklam ve tanıtım yapamayacak. Kampanya düzenleyemeyecek. Yardımlar dahil hiçbir etkinliğe isimlerini kullanarak destek olamayacak. Firmalar ve ürünlerine ilişkin markalar, tişört gibi kıyafet, takı ve aksesuar olarak taşınamayacak. Sigara, puro gibi ürünler, hediye, eşantiyon ve promosyon olarak dağıtılamayacak. Ayrıca geçen yıl ülkemizde sigaraya başlamayı önleme, sigara bırakmayı destekleme ve sigara dumanını kontrol altına almak amacıyla kamu hizmet binalarıyla koridorlar dahil olmak üzere her türlü eğitim, sağlık, spor ve eğlence yerlerinin kapalı alanlarında sigara içilmesini yasaklayan Tütün Ürünlerinin Zararlarının Önlenmesi ve Kontrolü Hakkında Kanun yürürlüğe girmiştir. Yasakla birlikte taksiler dahil olmak üzere tüm toplu taşıma araçlarında, okul öncesi eğitim kurumları ile dershanelerde, özel eğitim ve öğretim kurumları dahil olmak üzere ilk ve orta öğrenim kurumlarının kültür ve sosyal hizmet binalarının kapalı ve açık alanlarında sigara içilmesi yasaklandı. Yasak kapsamında, yasaklanan yerlerde sigara içenlere 50 YTL, izmarit, paket, ağızlık ve benzeri atıkları çevreye atanlara 20 YTL ceza verilecek. Yasaklara ilişkin yükümlülükleri yerine getirmeyen işletmeler önce yazılı olarak uyarılacak. Verilen sürede yükümlülüklerini yerine getirmeyen işletmelere 500 YTL'den 5 bin YTL'ye kadar para cezası uygulanacak. Yasaya aykırı ilan, reklam, pazarlama yapan ve satan tütün üretici firmalarına 50 bin YTL ile 250 bin YTL arasında ceza verilecek. Dizi, film, klip gibi yayınlarında sigara yasağına uymayan TV'ler 50 bin YTL ile 250 YTL arasında ceza ödeyecek. Hastane, okul gibi sağlık, eğitim ve kültür alanlarında sigara satanlara 1000 YTL ceza uygulanacak. 18 yaşından küçük çocuklara sigara satanlara 6 aydan 1 yıla kadar hapis cezası verilecek. Sigarayı tane ile satanlara 250 YTL; sakız, şeker, oyuncak gibi ürünleri sigarayı anımsatacak şekilde üretenlere 20 bin YTL ile 100 bin YTL arasında ceza uygulanacak. Sigara paketlerinde uyarıcı yazılara yer vermeyen veya yanlış yer verenlere, piyasaya sürdükleri mal kadar ceza verilecek ancak bu ceza 250 bin YTL'den az olmayacak. Bu cezaları vermeyen görevliler hakkında disiplin hükümleri uygulanacak. Sigara içilmesinin yasaklandığı yerlerde uyarılar, salonlarda en az 10 cm'lik puntolarla, toplu taşıma araçlarında ise en az 3 cm'lik puntolarla herkesin göreceği yerlere asılacak. Sigara paketlerinin üzerine en geniş iki yüzünden birinin yüzde 30, diğerinin yüzde 40'ından az olmamak üzere tütün ürünlerinin zararlarına ilişkin uyarılar Türkçe yazılacak.

Özellikle pasif içiciliğe önlem olması bakımından kapalı mekanlar için getirilen sigara içme yasağı bu konuda atılan en önemli adımdır. Elde edilen sonuçlara göre sigara içenlerin de arasında bulunduğu büyük bir oranın bu yasağa destek vermesi, getirilen yasağın sigarayı bırakmakta etkili olduğunu düşünmesi ve sigara kullananların büyük oranının bırakmayı denemesi bu uygulamanın başarıya ulaşmasında ki en büyük etkenlerdir.

Tüm bunlar gösteriyor ki toplum sigarayla mücadele konusunda her geçen daha da bilinçlenmektedir ve sigarayla mücadeleye destek vermektedir. Bu sebeple son günlerde bu yasağın kaldırılmasının gündeme getirilmesi ve yeniden her yerde sigara içilmesinin önünün açılmaya çalışılması son derece yersizdir.

KAYNAKÇA

A. Kitaplar

AKDENİZ, Fikri, (2007), ***Olasılık ve İstatistik,*** Eskişehir: Nobel Kitabevi.

ALPAR, Reha, (2003), ***Uygulamalı Çok Değişkenli İstatistiksel Yöntemlere Giriş***, Ankara: Nobel Yayın Dağıtım.

AYTAÇ, Mustafa, (1999), ***Matematiksel İstatistik,*** Bursa: Ezgi Kitabevi.

ÇINGI, Hülya, (1994), ***Örnekleme Kuramı***, Ankara: H.Ü. Fen Fakültesi Basımevi.

Dobson, A.J. (2002), ***An Introduction to Generalized Linear Models***, Chapman&Hall/CRC, U.S.A.

Everitt, B.S., Hothern, T., ***A Handbook Of Statistical Analyses Using R***.

Faraway, J. Julian, (2006), ***Extending The Linear Model With R***, Chapman and Hall/CRC.

McCullagh, P. and J.A. Nelder, (1989), ***Generalized Linear Models***, Chapman and Hall:London

Montgomery, D.C., Peck, E.A., Vining, G.,G. (2001), ***Introduction to Linear Regression Analysis***, *WILEY*, New York.

Myers, H. Raymond, Douglas C. Montgomery ve Geoffrey G.VINING, (2002), ***Generalized Linear Models***, A Wiley-Interscience Publication, John Wiley ve Sons, Inc.

ÖZLÜ, Tevfik, (2000), ***Sigara ve Bırakmanın Yolları***, İstanbul: Timaş Yayınları.

Ricci, Vito, (2005), ***R Function For Regression Analysis***.

Sönmez, Harun, (2006), ***R Yazılımı İle İstatistiksel Analiz***, Eskişehir: Kaan Kitabevi.

B. Tezler

Arabacı, Özer, (2002), ***"Lojistik Regresyon Analizi ve Bir Uygulama Denemesi"***, (Yayımlanmamış Yüksek Lisans Tezi), Uludağ Üniversitesi Sosyal Bilimler Enstitüsü, Bursa.

C. Makaleler

Acı, E., Sezgin, F., (2007), 4. Uluslararası Sivil Toplum Kuruluşları - Küresel Yoksulluk Kongresine (19-21 EKİM 2007, Çanakkale-Türkiye) Sunulmuş Ve Yayınlanmış Olan ***"İnsani Yoksullukla Mücadelede Bir Sivil Toplum Proje Örneği: İnsanca Yaşam Projesi"*** Adlı Bildirim.

Aktaş Cengiz, (2009), ***"Lojistik Regresyon Analizi: Öğrencilerin Sigara İçme Alışkanlığı Üzerine Bir Uygulama",*** Eskişehir Osmangazi Üniversitesi Sosyal Bilimler Enstitüsü Dergisi, Sayı:26, pp.107-121.

Ayan, Sezer ve Faruk Kocacık, (2009), ***"Çocuk İstismarı: Sivas (Türkiye) Örneği"***, Uluslararası İnsan Bilimleri Dergisi, Cilt:6, Sayı:1.

Bilgel, Nazan, (2002), ***"Türkiye'de Sigara İçme Yaygınlığı"*** Sigara ve Sağlık adlı kitapta yer alan makale, Bursa.

Bircan, Hüdaverdi, (2004), ***"Lojistik Regresyon Analizi: Tıp Verileri Üzerine Bir Uygulama"***, Kocaeli Üniversitesi Sosyal Bilimler Enstitüsü Dergisi, pp.185-208.

Deniz, Özlem, (2005), ***"Poisson Regresyon Analizi"***, İstanbul Ticaret Üniversitesi Fen Bilimleri Dergisi, Yıl:4, Sayı:7, Bahar, pp. 59-72.

Girginer, Nuray ve Bülent Cankuş, (2008), ***"Tramvay Yolcu Memnuniyetinin Lojistik Regresyon Analiziyle Ölçülmesi: Estram Örneği"***, Yönetim ve Ekonomi, 15(1), 181-194.

Işığıçok, E., Murat, D., (2007), ***"2007 Seçim Döneminde Ekonomik Ve Siyasi Duruma İlişkin Beklentiler: Bursa Uygulaması"***, 8. Türkiye Ekonometri ve İstatistik Kongresi, İnönü Üniversitesi, Malatya.

Oğuzlar, Ayşe, (2005), ***"Lojistik Regresyon Analizi Yardımıyla Suçlu Profilinin Belirlenmesi"***, Uludağ Üniversitesi İktisadi ve İdari Bilimler Dergisi, Cilt: 19, Sayı:1.

Peng, Chao-Ying Joanne, Kuk Lida Lee ve Gary M. İngersoll, (2002), ***"An İ ntroduction to Logistik Regression Analysis and Reporting"***, The Journal of Educational Research, Vol: 96, No:1.

Yeşilova, Abdullah ve İsmail Kasap, (2008), ***"Lojistik Regresyonda Meydana Gelen Aşırı Yayılımın İncelenmesi"***, Yüzüncü Yıl Üniversitesi, Ziraat Fakültesi, Tarım Bilimleri Dergisi, pp. 21-25.