

**UKUROVA NİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ**

YÜKSEK LİSANS TEZİ

Engin ARICAN

**NİTEL YANIT DEĞİŞKENE SAHİP REGRESYON MODELLERİNDE
TAHMİN YÖNTEMLERİ**

İSTATİSTİK ANABİLİM DALI

ADANA, 2010

ÇUKUROVA ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

NİTEL YANIT DEĞİŞKENE SAHİP REGRESYON MODELLERİNDE
TAHMİN YÖNTEMLERİ

Engin ARICAN

YÜKSEK LİSANS TEZİ

İSTATİSTİK ANABİLİM DALI

Bu tez 15 / 01 / 2010 Tarihinde Aşağıdaki Jüri Üyeleri Tarafından
Oybirliği/Oyçokluğu İle Kabul Edilmiştir.

İmza
Yrd. Doç. Dr. Mahmude Revan ÖZKALE
DANIŞMAN

İmza
Prof. Dr. Sadullah SAKALLIOĞLU
ÜYE

İmza
Yrd. Doç. Dr. Ebru ÖZGÜR GÜLER
ÜYE

Bu tez Enstitümüz İstatistik Anabilim Dalında hazırlanmıştır.

Kod No:

Prof. Dr. İlhami YEĞİNGİL
Enstitü Müdürü

Not: Bu tezde kullanılan özgün ve başka kaynaktan yapılan bildirişlerin, çizelge, şekil ve fotoğrafların
kaynak gösterilmeden kullanımı, 5846 sayılı Fikir ve Sanat Eserleri Kanunundaki hükümlere
tabidir.

ÖZ
YÜKSEK LİSANS TEZİ

**NİTEL YANIT DEĞİŞKENE SAHİP REGRESYON
MODELLERİNDE TAHMİN YÖNTEMLERİ**

Engin ARICAN

ÇUKUROVA ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ
İSTATİSTİK ANABİLİM DALI

Danışman : Yrd. Doç. Dr. Mahmude Revan ÖZKALE

Yıl : 2010 Sayfa : 105

Jüri : Yrd. Doç. Dr. Mahmude Revan ÖZKALE
: Prof. Dr. Sadullah SAKALLIOĞLU
: Yrd. Doç. Dr. Ebru ÖZGÜR GÜLER

Bu çalışmada, genelleştirilmiş lineer modellere ait parametre tahmini, istatistiksel sonuç çıkarımı ve genelleştirilmiş lineer modellerde çoklu iç ilişkinin tespiti, etkileri ve etkilerini en aza indirmek için kullanılan yanlı tahmin edicilere yer verilmiştir. Ayrıca, yanıt değişkeninin bir sayım verisi olduğu durumlar için Poisson regresyon ve Poisson regresyonda çoklu iç ilişki bulunması durumunda ridge ve temel bileşenler tahmin ediciler anlatılmıştır. Bunların yanı sıra, yanıt değişkeninin Bernoulli dağılımından olduğu durumlar için ise lojistik regresyon anlatılmıştır. Lojistik regresyonda çoklu iç ilişkinin bulunması durumunda ise ridge lojistik tahmin edici, temel bileşenler lojistik tahmin edici ve Stein tahmin edicilerin kullanımı anlatılmış ve hata kareler ortalama değerleri verilmiştir.

Anahtar Kelimeler: Genelleştirilmiş Lineer Modeller, Poisson Regresyon, Lojistik Regresyon, Ridge Tahmin Edici, Temel Bileşenler Tahmin Edici

ABSTRACT
MSc THESIS

**PARAMETER ESTIMATION IN REGRESSION MODELS WITH
QUALITATIVE RESPONSE VARIABLE**

Engin ARICAN

DEPARTMENT OF STATISTICS
INSTITUTE OF NATURAL AND APPLIED SCIENCES
UNIVERSITY OF ÇUKUROVA

Supervisor : Asst. Prof. Dr. Mahmude Revan ÖZKALE

Year : 2010 Pages : 105

Jury : Asst. Prof. Dr. Mahmude Revan ÖZKALE
: Prof. Dr. Sadullah SAKALLIOĞLU
: Asst. Prof. Dr. Ebru ÖZGÜR GÜLER

In this study, parameter estimation and statistical inferences for generalized linear models and detection, effects of multicollinearity in generalized linear models and biased estimators used for minimizing these effects are mentioned. Furthermore, Poisson regression, when response variable is count data, and ridge and principal components estimators when multicollinearity exists in Poisson regression, are explained. Besides, logistic regression, when response variable is from Bernoulli distribution, is described. The use of ridge logistic estimator, principal components logistic estimator and Stein logistic estimator are explained in the circumstances of existing multicollinearity in logistic regression. And the mean square error values are given.

KeyWords: Generalized Linear Models, Poisson Regression, Logistic Regression, Ridge Estimator, Principal Components Estimator

TEŞEKKÜR

Endüstri mühendisliği altyapım üzerine “istatistik” vizyonu açan; tez çalışmam boyunca bilgi-birikim ve bilim insanı karakterinden faydalandığım danışman Hocam Yrd. Doç. Dr. Mahmude Revan ÖZKALE’ye teşekkürlerimi sunarım.

Ayrıca göstermiş olduğu ilgi ve hassasiyetten dolayı TÜİK Adana Bölge Müdürü Durmuş Ali ŞAHİN’e teşekkürü bir borç bilirim.

Beni her zaman destekleyen anneme, babama, kardeşime, sevgili eşim Eytül’e ve dünyaya henüz 6 ay önce gelmiş olan oğlum Ahmet Talha’ya teşekkür ederim.

İÇİNDEKİLER

SAYFA

ÖZ.....	I
ABSTRACT.....	II
TEŞEKKÜR.....	III
İÇİNDEKİLER.....	IV
TABLolar DİZİNİ.....	VII
ŞEKİLLER DİZİNİ.....	VIII
SİMGELER VE KISALTMALAR.....	IX
1. GİRİŞ.....	1
2. GENELLEŞTİRİLMİŞ LİNEER MODELLER.....	3
2.1. Genelleştirilmiş Lineer Modellerde Parametre Tahmini.....	4
2.2. Çoklu İç İlişki Durumu.....	10
2.2.1. Çoklu İç İlişkinin Sebepleri.....	10
2.2.2. Çoklu İç İlişkinin Tespit Edilmesi.....	11
2.2.3. Çoklu İç İlişkinin Sonuçları.....	12
2.2.4. Çoklu İç İlişkinin Giderilmesi.....	13
2.3. Genelleştirilmiş Lineer Modellerde Bazı Yanlı Tahmin Ediciler.....	13
2.3.1. Ridge GLM Tahmin Edici.....	13
2.3.1.1. Ridge GLM Tahmin Edicide k Parametresinin Seçimi.....	16
2.3.2. Temel Bileşenler GLM Tahmin Edici.....	16
2.3.2.1. Temel Bileşenlerin Atılması.....	21
2.4. Genelleştirilmiş Lineer Modellerde İstatistiksel Sonuç Çıkarımları.....	22
2.4.1. Sapma (Deviance) ve Uyumun İyiliği.....	22
2.4.2. Modeldeki Katsayılar için Önemlilik Testi.....	23
2.4.3. Model Karşılaştırma.....	24
2.4.4. Rezidüler.....	25
2.4.4.1. Sapma (Deviance) Rezidü.....	25
2.4.4.2. Pearson Rezidü.....	26
3. POISSON REGRESYON	27
3.1. Poisson Regresyon Modeli.....	27

3.2. Poisson Regresyon Modelinde Maksimum Likelihood Parametre Tahmini...	29
3.3. Poisson Regresyonda Bazı Yanlı Tahmin Ediciler.....	32
3.3.1. Ridge Poisson Tahmin Edici.....	32
3.3.1.1. Ridge Poisson Tahmin Edicide k Parametresinin Seçimi.....	32
3.3.2. Temel Bileşenler Poisson Tahmin Edici	33
3.4. Poisson Regresyonda İstatistiksel Çıkarımlar.....	34
3.4.1. Sapma (Deviance) ve Uyumun İyiliği.....	34
3.4.2. Modeldeki Katsayılar için Önemlilik Testi.....	35
3.4.3. Model Karşılaştırma.....	35
3.4.4. Poisson Regresyon İçin Rezidüleri.....	35
3.4.4.1. Sapma (Deviance) Rezidü.....	35
3.4.4.2. Pearson Rezidü.....	36
4. LOJİSTİK REGRESYON	37
4.1. Lojistik Regresyon Modeli.....	41
4.2. Lojistik Regresyon Modelinde Maksimum Likelihood Parametre Tahmini...	46
4.2.1. Gruplandırılmamış Veri İçin Lojistik Regresyon Modeli Parametre Tahmini.....	46
4.2.2. Gruplandırılmış Veri İçin Lojistik Regresyon Modeli Parametre Tahmini.....	49
4.3. Lojistik Regresyonda Bazı Yanlı Tahmin Ediciler.....	52
4.3.1. Ridge Lojistik Tahmin Edici.....	53
4.3.1.1. Ridge Lojistik Tahmin Edicide k Parametresinin Seçimi.....	61
4.3.2. Temel Bileşenler Lojistik Tahmin Edici.....	65
4.3.3. Stein Lojistik Tahmin Edici.....	68
4.4. Lojistik Regresyonda İstatistiksel Çıkarımlar.....	69
4.4.1. Sapma (Deviance).....	69
4.4.2. Uyumun İyiliği.....	69
4.4.2.1. Pearson Ki-Kare İstatistiği ve Sapma (Deviance).....	70
4.4.2.2. Hosmer-Lemeshow Testi.....	72
4.4.3. Modeldeki Katsayılar için Önemlilik Testi.....	73
4.4.4. Model Karşılaştırma.....	73

4.4.5. Lojistik Regresyonda Tanılama (Diagnostics).....	73
5. ÖRNEK UYGULAMA.....	76
6. SONUÇ VE ÖNERİLER.....	93
KAYNAKLAR.....	95
ÖZGEÇMİŞ.....	101
EKLER.....	102

TABLolar DİZİNİ**SAYFA**

Tablo 4.1. Bernoulli olasılık dağılımı.....	41
Tablo 4.2. Gruplandırılmış veri yapısı.....	50
Tablo 5.1. YAKIT değişkeni için kukla değişken seviyeleri.....	77
Tablo 5.2. EGİTİM değişkeni için kukla değişken seviyeleri.....	78
Tablo 5.3. Sadece sabit terimi içeren başlangıç modeli.....	80
Tablo 5.4. Sadece sabit terimi içeren başlangıç modeli katsayı değeri.....	80
Tablo 5.5. Açıklayıcı değişkenler ve Wald istatistikleri-1.....	82
Tablo 5.6. Açıklayıcı değişkenler ve Wald istatistikleri-2.....	84
Tablo 5.7. Senaryolar.....	85
Tablo 5.8. Hosmer-Lemeshow testi.....	85
Tablo 5.9. İyi uyum sağlamayan gözlemlere ait değerler.....	92
Tablo 5.10. Gözlemler atılmadan önce ve sonra uyumun iyiliği ölçütleri.....	92

ŞEKİLLER DİZİNİ

SAYFA

Şekil 4.1. Lineer regresyonda, uydurulmuş yanıt değişken ve x değişkeni.....	43
Şekil 4.2. Lojistik regresyon yanıt fonksiyonu (S-Shape).....	44
Şekil 5.1. ΔD_j ye karşın $\hat{p}_j$ grafiği.....	86
Şekil 5.2. ΔX_j^2 ye karşın $\hat{p}_j$ grafiği.....	87
Şekil 5.3. $\Delta \hat{b}_j$ ye karşın $\hat{p}_j$ grafiği.....	88
Şekil 5.4. $\Delta \hat{b}_j$ ye karşın h_j grafiği.....	89
Şekil 5.5. ΔX_j^2 ye karşın h_j grafiği.....	90
Şekil 5.6. ΔD_j ye karşın h_j grafiği.....	91

SİMGELER VE KISALTMALAR

COICOP	: Amaca Yönelik Kişisel Tüketim Sınıflaması (Classification of Individual Consumption by Purpose)
EKK	: En Küçük Kareler
GLM	: Genelleştirilmiş Lineer Modeller
HBA	: Hanehalkı Bütçe Anketi
IRLS	: Yeniden Ağırlıklandırılmış İteratif En Küçük Kareler Yöntemi (Iteratively Reweighted Least Squares)
MLE	: Maksimum Likelihood Tahmin Edici (Maximum Likelihood Estimator)
MSE	: Hata Kareler Ortalaması (Mean Square Error)
TÜİK	: Türkiye İstatistik Kurumu
VIF	: Varyans Şişirme Faktörü (Variance Inflation Factor)

1. GİRİŞ

Açıklayıcı değişken veya değişkenler ile yanıt değişken arasındaki ilişkiyi inceleyebilmek amacıyla regresyon analizinden yararlanılır. Regresyon modeli oluşturulurken yanıt değişkenin hata teriminin sabit varyanslı ve normal dağılıma sahip olduğu varsayılır.

Lineer regresyonda, hata terimlerinin sabit varyanslı olmama durumunda ise lineer dönüşümler yapılır. Yanıt değişken dönüşümleri, normallik varsayımının sağlanamadığı ve sabit varyanslı olmama durumlarında etkili bir yoldur. Ağırlıklandırılmış en küçük kareler ise sabit varyanslı olmama durumu için en küçük kareler yöntemine alternatif bir yol olmuştur. Eğer yanıt değişken üstel bir dağılıma (normal dağılım, binom dağılım, Bernoulli dağılımı, gama dağılımı, Poisson dağılımı gibi) sahipse, ağırlıklandırılmış EKK yöntemine alternatif olabilecek diğer bir yaklaşım ise Genelleştirilmiş Lineer Modeller (GLM)'dir (Montgomery ve ark., 2001). Yine bu bağlamda, Myers (1990) ağırlıklandırılmış en küçük kareler yönteminin de kabul edilemez olmadığını ancak genelleştirilmiş lineer modellerin daha kullanışlı ve popüler olduğunu belirtmiştir. Ayrıca, Agresti (2002) de genelleştirilmiş lineer modellerin kullanımı, ağırlıklandırılmış en küçük karelerin, Bayes çıkarımının, minimum ki-kare tahmin edicilerin, minimum ayrıştırma bilgisinin, Kernel düzleştirmesinin ve cezalı likelihood yönteminin kullanımlarına alternatif olarak gösterilmiştir.

GLM hem lineer hem de lineer olmayan regresyon modelleri için kullanılabilir. GLM'de yanıt değişken Poisson dağılımına ait bir sayım verisi olduğu durumlarda Poisson regresyon, Bernoulli dağılımına sahip olduğu durumlarda ise lojistik regresyon kullanılır.

Regresyon varsayımlarına uygun olarak, açıklayıcı değişkenler arasında çoklu iç ilişkinin olmadığı durumlarda GLM'de tahmin yöntemi olarak, maksimum likelihood (en çok olabilirlik) metodu kullanılmaktadır. Açıklayıcı değişkenler arasında iç ilişkinin bulunduğu durumlarda maksimum likelihood tahmin edici kararsız parametre tahminleri üretecektir ve kurulan modelden yapılacak çıkarımlarda hataya sebep olacaktır. Bu durum, araştırmacıları oluşan kararsız

tahminleri daha kararlı hale getirecek yeni tahmin metotları geliştirmeye sürüklemiştir.

Bu çalışmada da, genelleştirilmiş lineer modeller için maksimum likelihood tahmin edici (MLE) ve çoklu iç ilişki olması durumunda kullanılan ridge tahmin edici ve temel bileşenler tahmin edici verilecektir. Daha sonra, Poisson ve lojistik regresyon için özel formları yazılacaktır. Ayrıca, lojistik regresyon için Stein tahmin edici de incelenecektir.

Bu tez çalışmasında ilk olarak, genelleştirilmiş lineer modeller anlatılacak ve daha sonra çoklu iç ilişkinin sebepleri, tespiti, sonuçları ve giderilmesi açıklanacaktır. Bölüm 3’de ise Poisson regresyon ve çoklu iç ilişkinin bulunduğu durumlarda kullanılan ridge Poisson tahmin edici ve temel bileşenler Poisson tahmin edici verilecektir. Bir sonraki bölümde ise lojistik regresyonun kullanım alanlarından bahsedildikten sonra teorik açıdan lojistik regresyonun gelişimi özetlenecektir. Lojistik regresyon modeli verildikten sonra, lojistik regresyonda çoklu iç ilişki durumu ele alınacak, nasıl tespit edileceği ve doğurduğu sonuçlar tartışılacaktır. Daha sonra, çoklu iç ilişkiden kaynaklanan etkileri en aza indirmek için geliştirilmiş yanlı tahmin edicilerden ridge lojistik tahmin edici, temel bileşenler lojistik tahmin edici ve Stein lojistik tahmin edici anlatılacaktır. Aynı zamanda, lojistik regresyonda istatistiksel sonuç çıkarımları ele alınacaktır. Bölüm 5’de ise lojistik regresyon tahmin edicileri ile ilgili örnek bir uygulama sunulacaktır.

2. GENELLEŞTİRİLMİŞ LİNEER MODELLER

Genelleştirilmiş lineer modeller, genellikle yanıt değişkenin nitel olduğu durumlarda kullanılan bir yöntemdir.

Genelleştirilmiş lineer modellerde kilit varsayım, yanıt değişkenin üstel aileden bir dağılıma sahip olmasıdır. Bernoulli, binom, normal, gama, Poisson dağılımları ise üstel aileye ait bazı dağılımlardır. $E(e_i) = 0$ olmak üzere, model

$$y_i = x_i' b + e_i \quad (2.1)$$

olarak yazılabilir. Burada $x_i' = [1, x_{i1}, x_{i2}, \dots, x_{ip}]$, $b' = [b_0, b_1, b_2, \dots, b_p]$ ve y_i ise üstel dağılım ailesinden rasgele değişken olsun.

Üstel dağılım ailesinin olasılık yoğunluk fonksiyonunun genel formu, $m_i = E(y_i)$ ve $g(m_i) = x_i' b$ olmak üzere ve $h_i = x_i' b$ link fonksiyonu olarak adlandırılmak üzere,

$$f(y_i, h_i, j) = \exp\left(\frac{y_i h_i - b(h_i)}{a(j)} + h(y_i, j)\right) \quad (2.2)$$

şeklinde verilmektedir (Montgomery ve ark, 2001). Bu genel form kullanılarak, üstel aile dağılımlarından olan Poisson ve Bernoulli dağılımları için olasılık yoğunluk fonksiyonlarının özel durumları yazılabilir. Poisson dağılımının olasılık yoğunluk fonksiyonu,

$$\begin{aligned} f(y_i) &= \frac{l_i^{y_i} e^{-l_i}}{y_i!} \\ &= \exp(y_i \ln l_i - l_i - \ln(y_i!)) \end{aligned} \quad (2.3)$$

şeklinde üstel formda yazılabilir. Eşitlik (2.2)'de yer alan olasılık yoğunluk fonksiyonunun genel formunu oluşturan terimler, eşitlik (2.3)'de

$$h_i = \ln l_i \text{ ve } l_i = e^{h_i},$$

$$b(h_i) = l_i,$$

$$a(j) = 1,$$

$$h(y_i, j) = -\ln(y_i!)$$

olur. Bernoulli dağılımı için ise olasılık yoğunluk fonksiyonu,

$$\begin{aligned} f(y_i) &= p_i^{y_i} (1-p_i)^{1-y_i} \\ &= \exp(y_i \ln p_i + (1-y_i) \ln(1-p_i)) \\ &= \exp(y_i \ln p_i - \ln(1-p_i) - y_i \ln(1-p_i)) \\ &= \exp\left(y_i \ln\left(\frac{p_i}{1-p_i}\right) + \ln(1-p_i)\right) \end{aligned} \quad (2.4)$$

şeklinde üstel formda yazılabilir. Eşitlik (2.2)'de yer alan olasılık yoğunluk fonksiyonunun genel formunu oluşturan terimler, eşitlik (2.4)'de

$$h_i = \ln\left(\frac{p_i}{1-p_i}\right) \text{ ve } p_i = \frac{e^{h_i}}{1+e^{h_i}},$$

$$b(h_i) = -\ln(1-p_i),$$

$$a(j) = 1,$$

$$h(y_i, j) = 0$$

olur.

2.1. Genelleştirilmiş Lineer Modellerde Parametre Tahmini

GLM için olasılık yoğunluk fonksiyonunun genel formu eşitlik (2.2) ile verilmişti. Parametre tahmini yapmak için, likelihood fonksiyonu yazılır:

$$L(y_i, b) = \prod_{i=1}^n \exp \left(\frac{y_i h_i - b(h_i)}{a(j)} + h(y_i, j) \right) \quad (2.5)$$

Buradan, log-likelihood fonksiyonu,

$$\mathbf{l}(y_i, b) = \sum_{i=1}^n \left(\frac{y_i h_i - b(h_i)}{a(j)} + h(y_i, j) \right) \quad (2.6)$$

olarak ifade edilir. Log-likelihood fonksiyonun b 'ya göre zincir kuralı kullanılarak kısmi türevi alınıp sıfıra eşitlenirse,

$$\frac{\partial \mathbf{l}}{\partial b} = \frac{\partial \mathbf{l}}{\partial h_i} \frac{\partial h_i}{\partial b} = 0 \quad (2.7)$$

$$\begin{aligned} \frac{\partial \mathbf{l}}{\partial h_i} &= \frac{\partial}{\partial h_i} \left(\sum_{i=1}^n \left(\frac{y_i h_i - b(h_i)}{a(j)} + h(y_i, j) \right) \right) \\ &= \sum_{i=1}^n \left(\left(y_i - \frac{\partial b(h_i)}{\partial h_i} \right) \frac{1}{a(j)} + \frac{\partial h(y_i, j)}{\partial h_i} \right) \\ &= \frac{1}{a(j)} \sum_{i=1}^n \left(y_i - \frac{\partial b(h_i)}{\partial h_i} \right) \end{aligned} \quad (2.8)$$

ve

$$\frac{\partial h_i}{\partial b} = \frac{\partial}{\partial b} (x_i' b) = x_i \quad (2.9)$$

olur. Eşitlik (2.8) ve (2.9) birlikte, (2.7)'da yerine yazılırsa,

$$\frac{\partial \mathbf{l}}{\partial b} = \frac{\partial \mathbf{l}}{\partial h_i} \frac{\partial h_i}{\partial b} = \frac{1}{a(j)} \sum_{i=1}^n \left(y_i - \frac{\partial b(h_i)}{\partial h_i} \right) x_i = 0 \quad (2.10)$$

elde edilir.

Log-likelihood fonksiyonu, $\mathbf{l}(y, b) = \sum_{i=1}^n \mathbf{l}_i$ olarak yazılabilir. $\mathbf{l}_i$ 'nin h_i 'ye göre kısmi türevi alınır,

$$\frac{\partial \mathbf{l}_i}{\partial h_i} = \frac{y_i - b'(h_i)}{a(j)} \quad (2.11)$$

elde edilir. (2.11)'in beklenen değeri sıfıra eşitlenirse (EK 1),

$$E\left(\frac{\partial \mathbf{l}_i}{\partial h_i}\right) = 0$$

ifadesinde eşitlik (2.11) yerine yazılırsa,

$$E(y_i - b'(h_i))/a(j) = 0$$

olur. $a(j)$ ifadesi bir katsayı olduğu için,

$$E(y_i) - b'(h_i) = 0$$

elde edilir. Buradan,

$$E(y_i) = b'(h_i) \quad (2.12)$$

ile GLM için beklenen değerlerin genel formu bulunmuş olur. Fisher Scoring'le (EK 1) ise,

$$-E\left(\frac{\partial^2 \mathbf{l}_i}{\partial h_i^2}\right) = E\left(\frac{\partial \mathbf{l}_i}{\partial h_i}\right)^2$$

$$\frac{b''(h_i)}{a(j)} = E\left(\frac{y_i - b'(h_i)}{a(j)}\right)^2$$

$$\frac{b''(h_i)}{a(j)} = \frac{\text{var}(y_i)}{a(j)^2}$$

$$\text{var}(y_i) = a(j)b''(h_i) = \frac{\partial E(y_i)}{\partial h_i} = \frac{\partial m_i}{\partial h_i} \quad (2.13)$$

ile GLM için varyansın genel formu elde edilmiş olunur. (2.13)'den anlaşılacağı üzere varyans, beklenen değerin bir fonksiyonu olup, lineer regresyonun varsayımlarından olan sabit varyanslılık durumu sağlanamamaktadır. Özel olarak, Poisson ve Bernoulli dağılımlarına ait beklenen değer ve varyans incelenebilir. Eşitlik (2.12) ve (2.13)'de Poisson dağılımına ait GLM terimleri yerine yazılırsa,

$$E(y_i) = \frac{\partial b(h_i)}{\partial h_i} = \frac{\partial}{\partial h_i}(e^{h_i}) = e^{h_i} = l_i \quad (2.14)$$

$$\text{var}(y_i) = \frac{\partial E(y_i)}{\partial h_i} = l_i \quad (2.15)$$

olarak elde edilir. Bernoulli dağılımı için aynı işlemler yapılırsa,

$$E(y_i) = \frac{\partial b(h_i)}{\partial h_i} = \frac{\partial}{\partial h_i} \left(-\ln(1 - \frac{e^{h_i}}{1 + e^{h_i}}) \right) = \frac{e^{h_i}}{1 + e^{h_i}} = p_i \quad (2.16)$$

$$\text{var}(y_i) = \frac{\partial E(y_i)}{\partial h_i} = \frac{\partial}{\partial h_i} \left(\frac{e^{h_i}}{1 + e^{h_i}} \right) = \frac{e^{h_i}}{(1 + e^{h_i})^2} = p_i(1 - p_i) \quad (2.17)$$

şeklinde bulunur.

Eşitlik (2.10)'un b için çözülmesi gerekmektedir. Fakat $l(b)$, b 'nın lineer bir formunda olmadığından doğrudan çözülemez ve Newton-Raphson metodu yardımıyla çözüm bulmaya çalışılır (EK 2).

Newton-Raphson yöntemi iteratif olarak kökün tahminini geliştirmek için bir $f(q)$ fonksiyonunun türevi, Taylor Serisi ile $q^{(m)}$ etrafında açılırsa aşağıdaki genel formda kullanılır (EK 2):

$$q^{(m+1)} = q^{(m)} - \left[\frac{\partial^2 f(q)}{\partial q \partial q'} \right]^{-1} \cdot f'(q^{(m)}) \quad (2.18)$$

GLM'deki durum için, Newton-Raphson metoduna göre çözüm $\frac{\partial \mathbf{l}}{\partial b}$ ifadesi

b etrafında açılarak elde edilir:

$$b_{MLE} = b - \left[\frac{\partial^2 \mathbf{l}}{\partial b \partial b'} \right]^{-1} \cdot \frac{\partial \mathbf{l}}{\partial b} \quad (2.19)$$

Eşitlik (2.7)'deki eşitlik zincir kuralı ile genişletilirse,

$$\begin{aligned} \frac{\partial \mathbf{l}}{\partial b} &= \sum_{i=1}^n \frac{\partial \mathbf{l}_i}{\partial h_i} \frac{\partial h_i}{\partial m_i} \frac{\partial m_i}{\partial h_i} \frac{\partial h_i}{\partial b} \\ &= \sum_{i=1}^n x_i \frac{\partial m_i}{\partial h_i} \frac{(y_i - m_i)}{\text{var}(y_i)} \end{aligned} \quad (2.20)$$

$$\frac{\partial \mathbf{l}}{\partial b'} = \sum_{i=1}^n \frac{\partial m_i}{\partial h_i} \frac{(y_i - m_i)}{\text{var}(y_i)} x_i' \quad (2.21)$$

elde edilir. Eşitlik (2.19)'daki $\frac{\partial^2 \mathbf{l}}{\partial b \partial b'}$ ifadesi, Fisher Scoring metodu (EK 1) ile,

$$\begin{aligned} E\left[\frac{\partial^2 \mathbf{l}}{\partial b \partial b'}\right] &= -E\left[\frac{\partial \mathbf{l}}{\partial b} \frac{\partial \mathbf{l}}{\partial b'}\right] \\ &= -E\left[\sum_{i=1}^n x_i \frac{\partial m_i}{\partial h_i} \frac{(y_i - m_i)}{\text{var}(y_i)} \frac{(y_i - m_i)}{\text{var}(y_i)} \frac{\partial m_i}{\partial h_i} x_i'\right] \\ &= -\sum_{i=1}^n x_i x_i' \left[\frac{\partial m_i}{\partial h_i}\right]^2 \frac{1}{\text{var}(y_i)} \\ &= -f \end{aligned} \quad (2.22)$$

olarak hesaplanır. Burada, $K^{-1} = \text{diag}\left(\frac{1}{\text{var}(y_i)} \left[\frac{\partial m_i}{\partial h_i}\right]^2\right)$ olmak üzere, $f = X' K^{-1} X$

bilgi matrisi olarak adlandırılmaktadır. Eşitlik (2.20) ve (2.22), eşitlik (2.19)'de yerine yazılırsa, $k_{ii}^{-1}: K^{-1}$ matrisinin köşegen elemanları olmak üzere,

$$\begin{aligned} b^{(m+1)} &= b^{(m)} + \left(f^{(m)}\right)^{-1} \frac{\partial \mathbf{l}}{\partial b} \Big|_{b=b^{(m)}} \\ &= b^{(m)} + \left(X' K_{(m)}^{-1} X\right)^{-1} \left[\sum_{i=1}^n x_i k_{ii}^{-1} (y_i - m_i) \frac{\partial h_i}{\partial m_i}\right]_{(m)} \end{aligned} \quad (2.23)$$

olur. (2.23)'de $\left(X' K^{-1} X\right)^{-1} X' K^{-1}$ ortak parantezine alınarak tekrar düzenlenirse,

$$b^{(m+1)} = \left(X' K_{(m)}^{-1} X\right)^{-1} X' K_{(m)}^{-1} \left(h_i + (y_i - m_i) \left(\frac{\partial h_i}{\partial m_i}\right)\right)_{(m)} \quad (2.24)$$

olarak GLM için maksimum likelihood iteratif tahmin edici elde edilir. Eşitlik (2.24)'de şu anki $b^{(m)}$ tahminleri kullanılarak her iterasyonda k_{ii}^{-1} ağırlıkları

güncellenmektedir. Bu nedenle, bu yönteme “Yeniden Ağırlıklandırılmış İteratif En Küçük Kareler Yöntemi (IRLS)” denir.

2.2. Çoklu İç İlişki Durumu

Regresyon modelindeki bazı ya da bütün açıklayıcı değişkenlerin birbirleri arasında yüksek derecede lineer ilişki olmasına çoklu iç ilişki denir. Bu durumda açıklayıcı değişkenler arasında bağımsızlık varsayımı sağlanmaz.

Lineer regresyon için, Özkale (2007) aşağıdaki durumu incelemiştir: X_j , X matrisinin j – inci kolon vektörü olsun: $X = [X_1, X_2, \dots, X_p]$

$$\sum_{j=1}^p t_j X_j = 0 \quad (2.25)$$

olacak şekilde hepsi sıfır olmayan $t_1, t_2, \dots, t_p$ sabitleri varsa $X_1, X_2, \dots, X_p$ vektörleri lineer bağımlıdır. X in kolonlarının bir kümesi için (2.25) sağlanıyorsa “çoklu iç ilişki” vardır. Çoklu iç ilişki olduğu zaman, tahmin edilen katsayılar istatistiksel anlamlılık, büyüklük ve işaret bakımından kararsız hale gelir.

2.2.1. Çoklu İç İlişkinin Sebepleri

Çoklu iç ilişkinin bir çok sebebi olabilir (Judge ve ark., 1985; Montgomery ve ark, 2001). Bunlardan bir kısmı:

1. Uygulanan veri toplama metodu
2. Modelin ve kitlenin yapısından kaynaklanan zorunluluklar
3. Modelin yanlış belirlenmesi
4. Gözlem sayısından çok açıklayıcı değişkenin modelde olması (overdefined) şeklinde verilmektedir.

2.2.2. Çoklu İç İlişkinin Tespit Edilmesi

Lineer regresyonda,

1. Açıklayıcı değişkenler arasındaki, yani standartlaştırılmış veri ile çalışırken $X'X$ matrisindeki, basit korelasyon katsayıları incelenir, 1'e yakın ise çoklu iç ilişki vardır.

2. VIF (Variance Inflation Factor) değerini kullanarak, $VIF_j = C_{jj}$ ($C_{jj} : C = (X'X)^{-1}$ matrisinin j -inci köşegen elemanı) olarak tanımlanan varyans şişirme faktörü $VIF_j > 10$ ise çoklu iç ilişki vardır.

3. $l_0^*, l_1^*, l_2^*, \dots, l_p^*$ 'ler $X'X$ matrisinin özdeğerleri olmak üzere, $k = \frac{l_{\max}^*}{l_{\min}^*}$ koşul sayısı 100'den büyük ise çoklu iç ilişki olduğu sonucuna varılır.

Benzer düşünceleri Smith ve Marx (1990) GLM'ye taşımışlardır. Lineer regresyonda $X'X$ matrisi çoklu iç ilişkinin etkilerini yansıtmaktadır. GLM'de ise f bilgi matrisi model parametrelerinin tahmini ve sonuç çıkarımında önemlidir. Bu nedenle X matrisinin kolonları arasındaki çoklu iç ilişkiden ziyade $W^* = K^{-1/2}X$ matrisinin kolonları arasındaki çoklu iç ilişki önemli olacaktır. Herhangi bir $d > 0$ sabiti için $K^{-1/2} \cong dI$ iken X ve W^* matrislerinin kolonları arasındaki çoklu iç ilişki denk olur.

1. Λ^* , $X'X$ matrisinin özdeğerlerinden oluşan köşegen matris ve T , $X'X$ matrisinin özvektörlerinden oluşan ortogonal matris olmak üzere, lineer regresyonda özdeğer analizinde, $X'X = T\Lambda T'$ olarak verilmektedir (Montgomery ve ark., 2001). Buradan hareketle, Smith ve Marx (1990) tarafından GLM için, Λ bilgi matrisinin özdeğerlerinden oluşan köşegen matris ve M bu özdeğerlere karşılık gelen özvektörlerden oluşan ortogonal matris olmak üzere $X'K^{-1}X = M\Lambda M'$ verilmiştir. W^{**} , W^* matrisinin standartlaştırılmış formu olmak üzere standartlaştırılmış bilgi matrisi $f^* = W^{**'}W^{**}$ korelasyon formunda olur. VIF ise,

$VIF_j = (f^*)^{-1}$ nin j – inci köşegen elemanı olarak tanımlanır.

2. $x_0, \dots, x_p$, f matrisinin artan sıradaki özdeğerleri olmak üzere bir koşul indeksini

$$CI_j = \left(\frac{x_{\max}}{x_j} \right)^{1/2} \quad j = 0, \dots, p$$

olarak tanımlamışlardır. Bu koşul indeksi çok büyük pozitif değerler (büyüklüğün derecesi belirtilmemiştir) verdiği zaman, çoklu iç ilişkinin varlığı söz konusudur.

2.2.3. Çoklu İç İlişkinin Sonuçları

1. Tahmin edilen parametreler gerçek değerlerinden uzaklaşır. Bunu Schaefer, Roi ve Wolfe(1984);

$$\begin{aligned} E(\hat{b}_{MLE}' \hat{b}_{MLE}) &= E[(\hat{b}_{MLE} - b + b)' (\hat{b}_{MLE} - b + b)] \\ &= E[(\hat{b}_{MLE} - b)' (\hat{b}_{MLE} - b) + 2b' (\hat{b}_{MLE} - b) + b' b] \\ &= E[(\hat{b}_{MLE} - b)' (\hat{b}_{MLE} - b)] + 2b' (E(\hat{b}_{MLE}) - b) + b' b \\ &= \text{trace}[\text{var}(\hat{b}_{MLE})] + [\text{bias}(\hat{b}_{MLE})]^2 + b' b \end{aligned}$$

olmak üzere

$$\begin{aligned} E(\hat{b}_{MLE}' \hat{b}_{MLE}) &= b' b + \text{trace}[\text{var}(\hat{b}_{MLE})] + [\text{bias}(\hat{b}_{MLE})]^2 \\ &\geq b' b + \sum \text{var}(\hat{b}_{(MLE)_j}) \end{aligned} \quad (2.26)$$

eşitliği ile ifade etmiştir. Böylece, gerçek parametre değerinden uzaklaşıldığı görülür. Burada $E[(\hat{b}_{MLE} - b)' (\hat{b}_{MLE} - b)]$, skaler hata kareler ortalamasının (MSE'nin) tanımındır.

2. Tahminlerin varyansları büyüyecektir. Bu durumu Smith ve Marx (1990) $x_j \rightarrow 0$ 'a giderken $\sum \text{var}(\hat{b}_{(MLE)_j}) = \text{trace}[(X'K^{-1}X)^{-1}] \rightarrow \infty$ olduğu şeklinde ifade etmişlerdir.

2.2.4. Çoklu İç İlişkinin Giderilmesi

Özkale (2007) ve Kaşko (2007) tarafından, verideki çoklu iç ilişkinin etkilerinin azaltılmasına yönelik bazı yöntemler şu şekilde özetlenmiştir:

1. Ek verinin toplanması: Ek veri toplamak çoklu iç ilişkiyi ortadan kaldırabilir. Ancak bu, modeldeki veya kitledeki kısıtlamalardan dolayı her zaman mümkün olmayabilir.
2. Modelin yeniden belirlenmesi: Çoklu iç ilişki modelin seçiminden de kaynaklanabilir. Örneğin, açıklayıcı değişkenlerden bazılarının modelden çıkarılması çoklu iç ilişkiyi azaltabilir. Ancak bu modelin etkinliğini de azaltabilir. Diğer bir yöntem olarak, açıklayıcı değişkenlerin birleştirilerek tek bir değişken olarak kullanılması da çoklu iç ilişkiyi azaltabilir. Ya da esas açıklayıcı değişkenin yerine bir oranın kullanılması da faydalı olabilir.
3. Alternatif regresyon analizlerinin kullanılması: Çoklu iç ilişkinin etkilerini azaltmaya yönelik geliştirilmiş tahmin edicilerin kullanılması da en etkin yöntemlerden birisidir. Bu tahmin ediciler minimum yanlılık üretme amacı taşımakta olup, “yanlı tahmin ediciler” olarak isimlendirilirler. GLM için geliştirilmiş ridge, temel bileşenler tahmin edicileri bunların bazılarıdır.

2.3. Genelleştirilmiş Lineer Modellerde Bazı Yanlı Tahmin Ediciler

2.3.1. Ridge GLM Tahmin Edici

Lineer regresyon modeli için Hoerl ve Kennard (1970) tarafından önerilen ridge tahmin edici genelleştirilerek bilgi matrisinin kötü koşulluluk problemi Schaefer (1979) tarafından incelenmiştir. Schaefer (1979) üstel dağılım ailesinden

olan Bernoulli dağılımına sahip yanıt değişkenin kullanıldığı lojistik regresyon modellerinde lojistik ridge tahmin ediciyi önermiştir. Smith ve Marx (1990) tarafından ise, ridge türü bir tahmin edici GLM için verilmiştir. Smith ve Marx (1990) tarafından doğrudan verilen ridge türü tahmin edici elde edilmek istensin.

Smith ve Marx'ın (1990) belirttiği gibi Schaefer (1979) ağırlıklı hata kareler toplamını kullanarak b yı tahmin etmiştir. $\hat{b}$ yerine $\tilde{b}$ kullanılırsa, ağırlıklı hata kareler toplamındaki artış j gibi bir değer olur. $Y' = [y_1, y_2, \dots, y_n]_{1 \times n}$ ve $\hat{Y}(\hat{b})$, $\hat{b}$ 'nın kullanılmasıyla elde edilen Y 'nin tahmini olmak üzere, $(Y - \hat{Y}(\tilde{b}))' K^{-1} (Y - \hat{Y}(\tilde{b}))$ ifadesine $\hat{Y}(\hat{b})$ ekleyip çıkarılırsa,

$$\begin{aligned} (Y - \hat{Y}(\tilde{b}))' K^{-1} (Y - \hat{Y}(\tilde{b})) &= [Y - \hat{Y}(\hat{b}) + \hat{Y}(\hat{b}) - \hat{Y}(\tilde{b})]' K^{-1} [Y - \hat{Y}(\hat{b}) + \hat{Y}(\hat{b}) - \hat{Y}(\tilde{b})] \\ &= (Y - \hat{Y}(\hat{b}))' K^{-1} (Y - \hat{Y}(\hat{b})) + 2(Y - \hat{Y}(\hat{b}))' K^{-1} (\hat{Y}(\hat{b}) - \hat{Y}(\tilde{b})) \\ &\quad + (\hat{Y}(\hat{b}) - \hat{Y}(\tilde{b}))' K^{-1} (\hat{Y}(\hat{b}) - \hat{Y}(\tilde{b})) \end{aligned}$$

olur. $\hat{b}$ 'nın optimal olması nedeni ile

$$j = 2(Y - \hat{Y}(\hat{b}))' K^{-1} (\hat{Y}(\hat{b}) - \hat{Y}(\tilde{b})) + (\hat{Y}(\hat{b}) - \hat{Y}(\tilde{b}))' K^{-1} (\hat{Y}(\hat{b}) - \hat{Y}(\tilde{b}))$$

alınabilir. Eşitlikteki $\hat{Y}$ 'ların lineerizasyonu (lineer regresyon için ridge tahmin ediciye benzetilmek amaçlanıyor) ,

$$j = (\tilde{b} - \hat{b})' X' K^{-1} X (\tilde{b} - \hat{b})$$

elde edilir.

Bilgi matrisi $f = X' K^{-1} X$ kötü koşullu iken, $\hat{b}$ ile b arasındaki uzaklık artmaktadır. Bu nedenle, $\tilde{b}' \tilde{b}$ tahmin edicisinin orijine uzaklığının (boyunun) karesi minimum yapılmak istenir. Fakat bu, keyfi değildir. Rezidü kareler toplamının verilen belli bir seviyesi için uzaklığın karesi minimum yapılmak istenir (aynı rezidü

kareler toplamına sahip pek çok tahmin edici olabilir, fakat bunlar içerisinde en küçük uzaklıkları olan seçilir. $k_0 > 0$ hata kareler toplamı için verilen bir sabit olsun. Bu durumda, $S(\tilde{b}) = S(\hat{b}) + k_0$ koşulunu sağlayan tahmin edicilerin bir $\{\tilde{b}\}$ kümesi vardır. Bu küme içerisinde en küçük uzunluklu $\tilde{b}$ tahmini bulunmak istenir (Özkale, 2007).

Lagrange çarpanı kullanılarak (EK 3),

$$F = \underset{\tilde{b}}{\text{Min}} \left\{ \tilde{b}'\tilde{b} + \frac{1}{k} \left[(\hat{b} - \tilde{b})' X' K^{-1} X (\hat{b} - \tilde{b}) - k_0 \right] \right\} \quad (2.27)$$

ifadesi araştırılır. $\tilde{b}$ 'ya göre kısmi türev alınıp sıfıra eşitlenirse,

$$\frac{\partial F}{\partial \tilde{b}} = 2\tilde{b} + \frac{1}{k} 2X' K^{-1} X (\tilde{b} - \hat{b}) = 0$$

parantez açılırsa,

$$2\tilde{b} + \frac{1}{k} 2X' K^{-1} X \tilde{b} = \frac{1}{k} 2X' K^{-1} X \hat{b}$$

olur. Her iki taraf k ile çarpılıp, eşitliğin sol tarafı $2\tilde{b}$ ortak parantezine alınırsa,

$$2\tilde{b} (kI + X' K^{-1} X) = 2X' K^{-1} X \hat{b} \text{ elde edilir. Son olarak bu eşitlikten } \tilde{b} \text{ çekilirse,}$$

$$\tilde{b} = \hat{b}_{\text{Ridge}} = (X' K^{-1} X + kI)^{-1} X' K^{-1} X \hat{b} \quad (2.28)$$

olarak “ridge GLM tahmin edici” elde edilmiş olunur. (2.28) ile verilen ridge GLM tahmin edici Smith ve Marx (1990) tarafından önerilen ridge tahmin ediciye denktir.

2.3.1.1. Ridge GLM Tahmin Edicide k Parametresinin Seçimi

Ridge GLM tahmin edicide k parametresinin seçimi Smith ve Marx (1990) tarafından Mallows'un (1973) C_p kriterine benzer şekilde

$$C_p^R = \frac{D_{n,p,k}}{\bar{D}_{p,k=0}} - n + 2\text{trace}[(X'K^{-1}X)(X'K^{-1}X + kI)^{-1}]$$

olarak önerilmiştir. Burada $D_{n,p,k}$ ridge parametresi k ve p değişkenli model için sapma (deviance)'dır. $\bar{D}_{p,k=0}$ ise ridge parametresi sıfır iken ölçek parametresinin tahminidir (lineer regresyon için varyansın tahminine denk bir ifade). Farklı k değerleri için C_p^R değerleri elde edilerek uygun olan tercih edilir. Lineer regresyon için C_p^* kriterinin minimum olması tercih edilirken GLM için Smith ve Marx (1990) herhangi bir karar kuralı önermemiştir.

2.3.2. Temel Bileşenler GLM Tahmin Edici

Çoklu iç ilişkinin olması durumunda lojistik regresyonda parametre tahmininde kullanılan diğer bir yöntem ise temel bileşenlerdir. Açıklayıcı değişkenler arasında lineer bağımlılık yoksa veri kümesi “ortogonal (dik)’dir” denir.

Temel bileşenler regresyonu, “her regresyon modelinin bir ortogonal açıklayıcı değişkenler kümesine dayanarak yeniden açıklanabileceği” gerçeği üzerine oluşturulmuştur. Bu yeni ve ortogonal açıklayıcı değişkenler, orijinal açıklayıcı değişkenlerin doğrusal bileşeni olarak elde edilirler ve bağımsız değişkenlerin “temel bileşenleri” olarak adlandırılırlar (Alpar, 1997).

Temel bileşenler analizinde, ortogonal olmayan açıklayıcı değişkenler için yeni açıklayıcı değişkenler kümesi oluşturulur. Oluşturulan bu kümede açıklayıcı değişkenler birbirine ortogonal olur. Bu küme, çok küçük öz değerlere sahip bir takım gözlemlerin atılması ile kalanlardan oluşturulmaktadır.

Smith ve Marx (1990) ve Marx ve Smith (1990) tarafından yayınlanan çalışmalarda, GLM’de çoklu iç ilişkinin mevcut olduğu durumlar için temel bileşenler yöntemini ele almışlar ve örnek bir veri kümesinde uygulama yapmışlardır. Bu çalışmalarda, GLM için temel bileşenler yöntemi genel kullanım için formülize edilerek sunulmuştur.

$y = Xb + e$ modeli kanonik formda yazılsın. $\Lambda = \text{diag}(x_1, x_2, \dots, x_{p+1})$, $f = (X' K^{-1} X)_{(p+1) \times (p+1)}$ bilgi matrisinin özdeğerlerinden oluşan köşegen matrisi ve M kolonları $x_1, x_2, \dots, x_{p+1}$ özdeğerlerine karşılık gelen özvektörleri olan ortogonal bir matrisi göstermek üzere $M' f M = \Lambda$ yazılabilir. $Z = XM$, $a = M' b$ olmak üzere ve $Z = [Z_1, Z_2, \dots, Z_{p+1}]$ ’nin kolonları, “temel bileşenler” olarak adlandırılan yeni bir ortogonal açıklayıcı değişkenlerin kümesini tanımlamak üzere,

$$y = Za + e \quad (2.29)$$

şeklinde yazılır. $p + 1$, tam (full) modeldeki toplam açıklayıcı değişken sayısını; s , indirgenmiş modeldeki açıklayıcı değişken sayısını ve r , tam modelden atılan açıklayıcı değişken sayısını gösterebilir.

Eşitlik (2.29), tam modeli göstermekte olup indirgenmiş model,

$$y_s = Z_s a_s + e \quad (2.30)$$

şeklinde yazılır. Diğer bir deyişle, tam model

$$y = Za + e = \begin{pmatrix} Z_s & Z_r \end{pmatrix} \begin{pmatrix} a_s \\ a_r \end{pmatrix} + e \quad (2.31)$$

olarak gösterilebilir. Ayrıca

$$\begin{bmatrix} Z_s & Z_r \end{bmatrix} \begin{bmatrix} Z_s & Z_r \end{bmatrix} = \begin{pmatrix} \Lambda_s & 0 \\ 0 & \Lambda_r \end{pmatrix}$$

gösterimi geçerlidir.

Smith ve Marx (1990) genelleştirilmiş lineer modeller için bir iteratif temel bileşen tahmin edici tanımlamışlardır.

$$g(m_i) = x_i' b = h_i \quad (2.32)$$

şeklinde yazılan link fonksiyonu, eşitlik (2.29) göz önüne alınarak z_i' , Z matrisinin satır vektörü olmak üzere kanonik formda, tam temel bileşen modelini göstermek üzere yazılırsa

$$h_i = z_i' a \quad (2.33)$$

elde edilir. Tam modeldeki a 'nın maksimum likelihood tahmini, eşitlik (2.23) yardımı ile iteratif Newton-Raphson metoduyla elde edilmek istenirse, $e_i = y_i - m_i$

ve $y_i^* = h_i + e_i \frac{\partial h_i}{\partial m_i}$ olmak üzere,

$$\begin{aligned} a_t &= a_{t-1} + \left[\Lambda^{-1} \sum_{i=1}^n z_i k_{ii}^{-1} e_i \frac{\partial h_i}{\partial m_i} \right]_{t-1} \\ &= \left[\Lambda^{-1} Z' K^{-1} y^* \right]_{t-1} \end{aligned}$$

şeklinde yazılır. Burada, $(t-1)$ -inci iterasyonda a_{t-1} tahminleri kullanılarak her iterasyonda k_{ii}^{-1} ağırlıkları güncellenmektedir. a_{t-1} , a_t 'e yeterince yakın olana kadar iterasyona devam edilir.

Bileşenlerin tam kümesi kullanılarak, tahmin edici $b = Ma$ 'nın hesaplanması, (2.32)'ye direkt olarak maksimum likelihood uygulanmasına denktir. Diğer bir deyişle b 'nın temel bileşenler tahmin edicisi

$$b_t - b_{t-1} = \left[f^{-1} \sum_{i=1}^n x_i k_{ii}^{-1} e_i \frac{\partial h_i}{\partial m_i} \right]_{t-1}$$

olarak elde edilir. Eşitliğin sağ-tarafı $(t-1)$ 'de değer verilerek iterasyona devam edilir. f 'nin sütunları arasında çoklu iç ilişkinin olması kararsız iterasyon adımlarına sebep olacaktır.

İndirgenmiş temel bileşen ise,

$$h_{i,s} = z'_{i,s} a_s$$

olarak tanımlanır. Burada $z'_{i,s}$, Z 'nin bir alt küme matrisi olan Z_s 'nin bir alt küme satır vektörü ve a_s , $f = M\Lambda M' = \sum_{j=0}^p x_j m_j m_j'$ bilgi matrisinde büyük değerlerle ilişkili a 'nın alt küme vektörüdür. Burada m_j , M matrisinin j -inci kolonunu yani x_j ye karşılık gelen özvektörü göstermektedir.

Uygun temel bileşenlerin atılması kararlı katsayı tahminini verecek ve varyansın azalmasını sağlayacaktır. Çok küçük özdeğere sahip gözlemler atılır. Ayrıca, Schaefer (1986) sabit X için maksimum likelihood tahmin edici $\hat{b}$ 'daki küçük değişikliklerin $\hat{K}^{-1}$ de küçük değişikliklere neden olduğunu belirtmiştir. Dolayısıyla maksimum likelihood yönteminde yakınsama olursa bilgi matrisi iyi tahmin edilme eğilimi içerisinde olacaktır yani $f = (X'K^{-1}X) \approx (X'K_{(t-1)}^{-1}X)$. Böylece, bir iteratif temel bileşenler tahmin edici, $\hat{e}_{i,s} = y_{i,s} - m_{i,s}$ olmak üzere,

$$b_{t,s}^{pc} = b_{t-1,s}^{pc} + \sum_{j=0}^{s-1} x_j^{-1} m_j m_j' \left[\sum_{i=1}^n x_i k_{ii}^{-1} e_{i,s} \frac{\partial h_i}{\partial m_i} \right]_{t-1} \quad (2.34)$$

olarak tanımlanır (Smith ve Marx, 1990). $x_s, x_{s+1}, \dots, x_p$, $r = p + 1 - s$ tane çok küçük özdeğer olmak üzere, eşitlik (2.34), a_s^{pc} 'ye göre tekrar düzenlenirse;

$$a_{t,s}^{pc} = a_{t-1,s}^{pc} + \Lambda_s^{-1} \left[\sum_{i=1}^n z_{i,s} k_{ii}^{-1} e_{i,s} \frac{\partial h_i}{\partial m_i} \right]_{t-1}$$

$$= \Lambda_s^{-1} Z_s' K^{-1} y_{t-1}^{*pc}$$

elde edilir. Burada $z_{i,s}$, Z_s' matrisinin sütunlarıdır ve $y_{t-1}^{*pc} = h_{i,s} + e_{i,s} \frac{\partial h_i}{\partial m_i}$ dir ve $(t-1)$ iterasyonunda değer verilir. Doğal bir sonuç olarak, $\hat{a}_s^{pc}$ 'nin yakınsamasına dayalı “temel bileşenler GLM tahmin edici”

$$\hat{b}_s^{pc} = M_s a_s^{pc} \quad (2.35)$$

olur.

Temel bileşenler GLM tahmin edicinin varyansına bakılacak olursa, $M = [M_s \ M_r]$, bilgi matrisinin özvektörünün tam kümesi olmak üzere

$$\text{var}(\hat{b}_s^{pc}) = \text{var}(\hat{b}_{MLE}) - M_r \Lambda_r^{-1} M_r' \quad (2.36)$$

olur. $f_r^{-1} = M_r \Lambda_r^{-1} M_r'$ ifadesi, tam modelden atılan gözlemlerin sebep olduğu

varyansı göstermektedir. $\text{var}(\hat{b}_{MLE}) = (X' K^{-1} X)^{-1}$ ve

$(X' K^{-1} X)^{-1} = M \Lambda^{-1} M' = M_s \Lambda_s^{-1} M_s' + M_r \Lambda_r^{-1} M_r'$ olduğundan (2.36) tekrar düzenlenirse,

$$\text{var}(\hat{b}_s^{PC}) = M_s \Lambda_s^{-1} M_s' \quad (2.37)$$

olarak elde edilir. $\hat{b}_s^{pc}$ 'ye ilişkin yanlılık ise $\text{bias}(\hat{b}_s^{PC}) = E(\hat{b}_s^{PC}) - b = -M_r a_r$ şeklinde verilmektedir. $\hat{b}_s^{PC}$ için MSE ise;

$$MSE(\hat{b}_s^{pc}) = M_s \Lambda_s^{-1} M_s' + (-M_r a_r)(-M_r a_r)'$$

şeklinde matris formunda yazılabilir. $\hat{b}_s^{pc}$ için skaler MSE ise;

$$tr[MSE(\hat{b}_s^{pc})] = \sum_{j=0}^p \sum_{a=0}^{s-1} m_{aj}^2 x_a^{-1} + \sum_{j=0}^p \left(\sum_{k=s}^p a_k m_{jk} \right)^2$$

olarak verilebilir (Marx ve Smith, 1990).

2.3.2.1. Temel Bileşenlerin Atılması

Jolliffe (1986) temel bileşenlerin atılması konusunda çeşitli yaklaşımları özetlemiştir:

Bu yaklaşımlardan biri, bileşenlerin önemliliği üzerine asimptotik dağılımı kullanarak test yapmaya dayalıdır. Temel bileşenlerin tam kümesi $\hat{a} \sim N(a, \Lambda^{-1})$ için C , $q \times (p+1)$ boyutlu sabitlerin matrisi olmak üzere $H_0: Ca = 0$ hipotezi test edilmek istenir. Test istatistiği ise $\hat{a}' C' (C \Lambda^{-1} C')^{-1} C \hat{a} \sim c_q^2$ olarak verilir. Anlamsız olan bileşenler modelden atılır.

Diğer bir yaklaşım, belli bir eşik (cut-off) noktası belirlenerek (örneğin 0.01 gibi) bunun altındaki küçük özdeğere sahip bileşenlerin atılmasıdır. Kaiser (1958) metodu olarak bilinen yöntem, özdeğeri 1.00'in altında olan temel bileşenlerin atılması ile uygulanır.

Bunların dışında, Hill, Fomby ve Johnson'ın (1977) önerdiği kriterlerden birisi b 'ya $\hat{b}_s^{pc}$ ile yaklaşmayı amaçlar. Yani,

$$tr[MSE(\hat{b}_s^{pc})] \leq tr[MSE(\hat{b})] \quad (2.38)$$

ise $\hat{b}_s^{pc}$, $\hat{b}$ ya tercih edilir. Eşitlik (2.38) dekinden daha güçlü bir kriter ise, boş olmayan q tane uygun boyut (modelde olabilecek temel bileşen sayısı) için,

$$MSE(q' \hat{b}_s^{pc}) \leq MSE(q' \hat{b})$$

dir. Burada, $MSE(\hat{b}_s^{pc}) = E[(\hat{b}_s^{pc} - b)(\hat{b}_s^{pc} - b)']$ dir.

2.4. Genelleştirilmiş Lineer Modellerde İstatistiksel Sonuç Çıkarımlar

2.4.1. Sapma (Deviance) ve Uyumun İyiliği

p -tane açıklayıcı değişkenin bulunduğu modelin önemliliğini test etmek için “Likelihood Oran Testi” kullanılır. Bu testi yapmak için, doymuş model (mükemmel uyumun sağlandığı) ile uydurulmuş modelin likelihood fonksiyonları oranlanır. $-2\ln$ ile çarpılarak, sapma (deviance) adı verilen ki-kare dağılımına ve “p-eğim katsayılarının” sayısına eşit serbestlik derecesine sahip bir istatistik elde edilir ve tablo değeri ile karşılaştırılarak H_0 hipotezinin kabul yada red edilmesine karar verilir.

H_0 : Uydurulan model anlamsızdır.

H_1 : Uydurulan model anlamlıdır.

$\hat{h}_i$, h_i 'nin tahmini olmak üzere sapma

$$\begin{aligned} D = \text{Deviance} &= -2 \ln \left[\frac{L(b)}{L(P)} \right] = -2 [L(\hat{m}; y) - L(y; y)] \\ &= -2 \ln [L(\text{uydurulmuş model}) - L(\text{doymuş model})] \\ &= -2 \sum_{i=1}^n \left(\frac{[y_i \hat{h}_i - b(\hat{h}_i)]}{a(j)} - \frac{[y_i h_i - b(h_i)]}{a(j)} \right) \end{aligned}$$

şeklinde yazılır.

Eğer $D > c_{a,p}^2$ ise, a önemlilik düzeyinde H_0 kabul edilir ve modelin yetersiz olduğu karara varılır. Yani, büyük D-değeri modelin anlamsız olduğunu gösterir.

2.4.2. Modeldeki Katsayılar için Önemlilik Testi

Modeldeki katsayıların önemliliğinin test edilmesi için kurulacak hipotez testi aşağıdaki gibidir:

$H_0 : b_i = 0$ (x_i nin modele önemli bir katkısı yoktur.)

$H_1 : b_i \neq 0$ (x_i nin modele önemli bir katkısı vardır.)

Modeldeki katsayıların önemliliğinin test edilmesi için çeşitli yöntemler bulunmaktadır:

- **LR (Likelihood Ratio) Testi:**

$$D(b_i | b_1, b_2, \dots, b_{i-1}, b_{i+1}, \dots, b_k)$$

(İlgili katsayıdan kaynaklanan sapma değeri)

- **Wald Testi:**

$$[W_i]^2 = \left[\frac{\hat{b}_i}{s\hat{e}(\hat{b}_i)} \right]^2$$

- **Score Testi:**

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

Eğer yukarıdaki istatistikler, tablo değeri $c_{a,1}^2$ 'den büyük ise, a önemlilik düzeyinde H_0 reddedilir. Yani, x_i nin modele önemli bir katkısı vardır.

2.4.3. Model Karşılaştırma

Tam (full) model p -tane parametreye sahip olsun. İndirgenmiş (reduced) model ise daha az parametreye sahip olacaktır. Ancak model, tam tersine sapma açısından daha büyük bir değere sahip olacaktır. Buna dayalı olarak, model karşılaştırmaları yapılabilir ve böylelikle modele dahil edilecek değişkenin oluşturulan modele önemli bir katkısının olup olmadığına bakılır.

$$\begin{aligned} h &= Xb \\ &= X_1 b_1 + X_2 b_2 \end{aligned}$$

modeli ele alınacak olursa, tam model p -tane değişkene, b_1 ise $p - r^*$ tane değişkene ve b_2 ise r^* -tane değişkene sahip olur.

$$H_0 : b_2 = 0$$

$$H_1 : b_2 \neq 0$$

Hipotezi test edilmek istenirse, indirgenmiş model $h_{reduced} = X_1 b_1$ olur. Sapma kullanılarak hipotez test edilir:

$$D(b_2 | b_1) = D(b_1) - D(b). \quad (2.39)$$

Eşitlik (2.39)'da indirgenmiş modelin sebep olduğu sapma değerinden tam modelin sahip olduğu sapma değeri çıkarılarak, test edilen modelin sahip olduğu sapma değeri elde edilir. Böylelikle bu sapma değeri r^* serbestlik derecesine sahip ki-kare tablo değeri ile hipotez testi gerçekleştirilir.

Eğer $D(b_2|b_1) \geq c_{a,r}^2$ ise, a önemlilik düzeyinde H_0 reddedilir ve test edilen bir alt küme modelin sahip olduğu değişkenlerin tam modele önemli bir katkısının olduğu kararına varılır.

2.4.4. Rezidüler

Yanıt değişkene ait gözlemlerin gerçekleşmiş ve tahmin edilmiş değerleri arasındaki farka rezidü (artık) denir. y_i , i -inci gözleme ait gerçekleşmiş değeri, $\hat{y}_i$ ise aynı gözleme ait tahmin edilmiş değeri göstermek üzere rezidü,

$$e_i = y_i - \hat{y}_i \quad (2.40)$$

şeklinde gösterilir. Rezidülerin incelenmesi, etkin ve aykırı gözlemlerin belirlenmesinde önemli bir role sahiptir. Eşitlik (2.40) ile verilen ham rezidünün kullanılması yanıltıcı olabilir. Bunun yerine, GLM için sapma rezidü ve Pearson rezidü kullanımı önerilmiştir (Agresti, 2002). Sapma rezidü, her bir gözlemin sebep olduğu $-2\text{Log}L$ değerinin ölçüsüdür. Yani, modelin anlamlı olmasına katkısının ölçüsüdür. Pearson rezidü ise, ham rezidünün varyansına bölümüyle elde edilmekte olduğundan ham rezidüye göre daha kullanışlı olmaktadır.

2.4.4.1. Sapma (Deviance) Rezidü

$D(y; \hat{m}) = \sum d_i$ olmak üzere,

$$d_i = 2w_i [y_i(\hat{h}_i - h_i) - b(\hat{h}_i) - b(h_i)] \quad (2.41)$$

olarak i -inci gözleme ait sapma değeri elde edilebilir. Böylece, i -inci gözleme ait sapma (deviance) rezidü ise,

$$\sqrt{d_i} \text{sign}(y_i - \hat{y}_i)$$

ile verilmektedir. Rezidünün sahip olduğu işaret, $(y_i - \hat{y}_i)$ 'nin işaretidir.

2.4.4.2. Pearson Rezidü

Sapma (deviance) rezidüye alternatif olabilecek diğer bir rezidü ise Pearson rezidüdür. Pearson rezidü,

$$e_i = \frac{y_i - \hat{y}_i}{[\text{var}(\hat{y}_i)]^{1/2}}$$

olarak ifade edilmektedir.

3. POISSON REGRESYON

Herhangi bir olayın belirlenen bir süreç içerisinde yapılan denemeler sonucunda meydana gelme sayısı, sayma verileri olarak ifade edilebilir. Sayma veri modelinde bilinen ilk gelişmeler aktueryal bilimler, biyoistatistik ve demografide gözlenmiştir. Son yıllarda bu modeller iktisat, politik bilimler ve sosyolojide de sıkça kullanılmaya başlanmıştır. Sayma veri modelleri özel bir regresyon türüdür. Bu modeller ekonometricilerin çok fazla dikkatini çekmiş ve mikro ekonomide oldukça fazla kullanılmıştır. Sayma verilerinden oluşan değişkenler için sürekli ve doğrusal regresyon modelinin uygulanabileceği düşünülür. Ancak bu verilere doğrusal regresyon modeli uygulanması halinde sonuçlar, etkisiz ve tutarsız olduğu gibi çelişkili tahminler yapılabilir. Poisson regresyon modeli sayma verileri için en sık kullanılan ve en basit olan yöntemdir. Bu model ile sayımın olasılığı, Poisson dağılımı ile belirlenir (Deniz, 2005).

3.1. Poisson Regresyon Modeli

GLM’de yanıt değişken farklı üstel dağılımlara da sahip olabilmektedir. Bunlardan bir tanesi de, yanıt değişkenin Poisson dağılımına sahip bir sayma (count) verisi olması durumudur.

Eşitlik (2.1) ile verilen model, y_i -yanıt değişkenin Poisson dağılımından olmak üzere ele alınsın. Modelin beklenen değeri incelenecek olursa,

$$E(y_i) = E(x_i' b + e_i) = E(x_i' b) + E(e_i) = E(x_i' b) + 0 = x_i' b \quad (3.1)$$

elde edilir. Bunun yanı sıra, olasılık teorisinden ise, y_i gibi Poisson dağılımına sahip bir yanıt değişkenin olasılık yoğunluk fonksiyonu, $y = 0,1,2,\dots$ ve $l > 0$ olmak üzere,

$$f_i(y_i) = \frac{e^{-l_i} l_i^{y_i}}{(y_i)!} \quad (3.2)$$

şeklinde verilmektedir. Eşitlik (3.2) kullanılarak, Poisson dağılımının beklenen değeri hesaplanacak olursa,

$$E(y_i) = \sum y_i f(y_i) dy_i = l_i \quad (3.3)$$

elde edilir. Eşitlik (3.1) ve (3.3) birlikte ele alınırsa,

$$E(y_i) = x_i' b = l_i \quad (3.4)$$

olur. Poisson dağılımının varyansı incelenecek olursa,

$$\text{var}(y_i) = E(y_i^2) - [E(y_i)]^2 = l_i \quad (3.5)$$

elde edilir. Eşitlik (3.4) kullanılarak, eşitlik (2.1) tekrar düzenlenirse,

$$y_i = E(y_i) + e_i = l_i + e_i$$

şeklinde yazılabilir. Buradan hata terimi, $e_i = y_i - l_i$ şeklinde yazılabilir. Hata teriminin beklenen değeri ise,

$$E(e_i) = E(y_i - l_i) = E(y_i) - E(l_i) = l_i - l_i = 0$$

olur. Ayrıca, hata teriminin varyansı incelenirse,

$$\text{var}(e_i) = \text{var}(y_i - l_i)$$

olur. l_i sabit sayı olduğundan, hata teriminin varyansı,

$$\text{var}(e_i) = \text{var}(y_i)$$

olur. Yani, hata terimlerinin varyansı da beklenen değerin bir fonksiyonudur ve bu da sabit varyanslılık durumunun olmadığını göstermektedir. Bu durumda, lineer forma getirebilmek için yanıt değişkenin Poisson dağılımına sahip olduğu durumlarda kullanılan log-link dönüşümün uygulanması gerekmektedir. Log-link dönüşüm

$$\begin{aligned} h_i &= x_i' b \\ h_i &= \ln(l_i) \\ \ln(l_i) &= x_i' b \\ l_i &= e^{x_i' b} \end{aligned} \tag{3.6}$$

olarak verilir. Bu dönüşümle beraber, yanıt değişkenin tahmin edilen değerlerinin hepsi non-negatif olacaktır (Montgomery ve ark., 2001). Böylelikle, Poisson regresyon modeli,

$$y_i = e^{x_i' b} + e_i \tag{3.7}$$

olarak elde edilmiş olunur.

3.2. Poisson Regresyon Modelinde Maksimum Likelihood Parametre Tahmini

Poisson regresyonda, her bir gözlem Poisson dağılımına sahip bağımsız gözlemler olduğundan olasılık yoğunluk fonksiyonu eşitlik (3.2)'de verildiği gibidir. Likelihood fonksiyonu ise,

$$L(y, b) = \prod_{i=1}^n f(y_i) = \prod_{i=1}^n \frac{e^{-l_i} l_i^{y_i}}{(y_i)!} = \frac{\left(\prod_{i=1}^n l_i^{y_i} \right) e^{-\sum_{i=1}^n l_i}}{\prod_{i=1}^n (y_i)!}$$

şeklindedir. Log-likelihood fonksiyonu ise,

$$l(b) = \sum_{i=1}^n y_i \ln(l_i) - \sum_{i=1}^n l_i - \sum_{i=1}^n \ln(y_i!)$$

olarak yazılmaktadır. Log-link yerine konursa,

$$l(b) = \sum_{i=1}^n y_i x_i' b - \sum_{i=1}^n e^{x_i' b} - \sum_{i=1}^n \ln(y_i!)$$

olur ve maksimum likelihood tahmini bulunmak istendiğinden, b 'ya göre birinci dereceden kısmi türevi alınırsa,

$$\frac{\partial l(b)}{\partial b} = \sum_{i=1}^n y_i x_i - \sum_{i=1}^n x_i e^{x_i' b} = \sum_{i=1}^n x_i (y_i - l_i) = 0 \quad (3.8)$$

olur. Bu denklemin b için çözülmesi gerekmektedir. Fakat $l(b)$, b 'nın lineer bir formunda olmadığından doğrudan çözülemez ve Newton-Raphson metodu yardımıyla çözüm bulmaya çalışılır (EK 2). Eşitlik (2.18) ile verilen Taylor açılımı GLM için eşitlik (2.19) ile verilmişti. Bu eşitlikteki ifadeler, Poisson regresyon için elde edilmelidir:

$$\frac{\partial l(b)}{\partial b} = \sum_{i=1}^n (y_i - l_i) x_i' \quad (3.9)$$

eşitliğinden yararlanılırsa

$$\frac{\partial^2 l(b)}{\partial b \partial b} = - \sum_{i=1}^n l_i x_i x_i' \quad (3.10)$$

elde edilir.

$$x_i' b = b_0 + b_1 x_{i1} + \dots + b_p x_{ip}$$

$$X' = [x_1, x_2, \dots, x_n]_{(p+1) \times n}$$

$$Y' = [y_1, y_2, \dots, y_n]_{1 \times n}$$

$$y' = [l_1, l_2, \dots, l_n]_{1 \times n}$$

$$W = \text{diag}(l_i)_{n \times n}$$

olmak üzere (3.10) ve (3.8) eşitlikleri sırasıyla

$$\frac{\partial^2 \mathbf{l}(b)}{\partial b \partial b'} = -X' W X \quad (3.11)$$

ve

$$\frac{\partial \mathbf{l}(b)}{\partial b} = X' Y - X' y \quad (3.12)$$

olarak yazılır.

(3.11) ve (3.12) eşitlikleri ise (2.19)'de yerine yazılırsa,

$$b^{(m+1)} = b^{(m)} + (X' W_m X)^{-1} X' (Y - y_m) \quad (3.13)$$

elde edilir. Burada, $y_m' = (l_{1,m}, l_{2,m}, \dots, l_{n,m})$ önceki iterasyondan elde edilen uydurulmuş yanıt olasılıklar vektörü ve $W_m = \text{diag}(l_{i,m})$ önceki iterasyonda elde edilen köşegen matristir. (3.13) eşitliği, $(X' W_m X)^{-1} X' W_m$ ortak parantezine alınarak yeniden düzenlenirse,

$$b^{(m+1)} = (X' W_m X)^{-1} X' W_m (X b^{(m)} + W_m^{-1} (Y - y_m))$$

yazılabilir. $Q_a = Xb^{(m)} + W_m^{-1}(Y - y_m)$ düzeltilmiş yanıt (adjusted response) olmak üzere,

$$b^{(m+1)} = (X'W_mX)^{-1}X'W_mQ_a \quad (3.14)$$

olarak elde edilir.

3.3. Poisson Regresyonda Bazı Yanlı Tahmin Ediciler

Çoklu iç ilişkinin olması durumunda ise, yanlı tahmin ediciler kullanılmaktadır. Ridge Poisson tahmin edici ve temel bileşenler Poisson tahmin edici ise bunların başında gelir.

3.3.1. Ridge Poisson Tahmin Edici

GLM için Bölüm 2.3.1’de anlatılanlara paralel olarak, Lagrange çarpanı kullanılarak (EK 3), K^{-1} ifadesi yerine Poisson regresyonun ağırlıkları olan W matrisi kullanılırsa,

$$\hat{b}_{Ridge} = (X'WX + kI)^{-1}X'W\hat{b} \quad (3.15)$$

olarak “ridge Poisson tahmin edici” elde edilmiş olunur.

3.3.1.1. Ridge Poisson Tahmin Edicide k Parametresinin Seçimi

Ridge Poisson tahmin edicide k parametresinin seçimi için Bölüm 2.3.1.1’de verilen yöntemler K^{-1} yerine Poisson regresyonun ağırlıkları olan W matrisi yazılarak kullanılabilir.

3.3.2. Temel Bileşenler Poisson Tahmin Edici

Smith ve Marx (1990) genelleştirilmiş lineer modeller için iteratif temel bileşen tahmin edici tanımlamışlardır. Bu tahmin ediciden Bölüm 2.3.2’de bahsedilmiş olup, bu tahmin edici Poisson regresyona uyarlanabilir. Eşitlik (2.34) ile

verilen GLM için iteratif temel bileşenler tahmin edici, $\frac{\partial h_i}{\partial m_i} = \frac{1}{l_i}$,

$e_{i,s} = y_{i,s} - l_{i,s}$ olmak üzere ve k_{ii}^{-1} ifadesi yerine $w_{ii} = \text{diag}(W)$ kullanılarak,

$$b_{t,s}^{pc} = b_{t-1,s}^{pc} + \sum_{j=0}^{s-1} x_j^{(*)-1} g_j^{(*)} g_j^{(*)'} \left[\sum_{i=1}^n x_i w_{ii} (y_{i,s} - l_{i,s}) \frac{1}{l_i} \right]_{t-1} \quad (3.16)$$

şeklinde Poisson regresyon için güncellenebilir. Burada, $x_j^{(*)}$, Poisson regresyona ait bilgi matrisinin ($f_{\text{Poisson}} = T\Lambda^{(*)}T'$) özdeğerleri ve $g_j^{(*)}$ ise bu özdeğerlere ilişkin özvektörleri göstermektedir. $x_s^{(*)}, x_{s+1}^{(*)}, \dots, x_p^{(*)}$, $r = p + 1 - s$ tane çok küçük özdeğer ve $Z^{(*)} = XT$ olmak üzere, eşitlik (3.16), a_s^{pc} ’ye göre tekrar düzenlenirse;

$$\begin{aligned} a_{t,s}^{pc} &= a_{t-1,s}^{pc} + \Lambda_s^{(*)-1} \left[\sum_{i=1}^n z_{i,s}^{(*)} w_{ii} (y_{i,s} - l_{i,s}) \frac{1}{l_i} \right]_{t-1} \\ &= \Lambda_s^{(*)-1} Z_s^{(*)} W y_{t-1}^{**pc} \end{aligned}$$

elde edilir. Burada $z_{i,s}^{(*)}$, $Z_s^{(*)'}$ matrisinin sütunlarıdır ve

$y_{t-1}^{**pc} = \ln(l_{i,s}) + (y_{i,s} - l_{i,s}) \frac{1}{l_i}$ dir ve $(t-1)$ -inci iterasyonda değer verilerek $a_{t-1,s}^{pc}$

tahminleri kullanılarak her iterasyonda w_{ii} ağırlıkları güncellenmektedir. Doğal bir sonuç olarak, $\hat{a}_s^{pc}$ ’nin yakınsamasına dayalı “temel bileşenler Poisson tahmin edici”

$$\hat{b}_s^{pc} = T_s a_s^{pc} \quad (3.17)$$

olur.

Temel bileşenler Poisson tahmin edicinin varyansına bakılacak olursa, $T = [T_s \ T_r]$, bilgi matrisinin özvektörünün tam kümesi olmak üzere

$$\text{var}(\hat{b}_s^{pc}) = \text{var}(\hat{b}_{MLE}) - T_r \Lambda_r^{(*)-1} T_r' \quad (3.18)$$

olur. $f_{(Poisson)r}^{-1} = T_r \Lambda_r^{(*)-1} T_r'$ ifadesi, tam modelden atılan gözlemlerin sebep olduğu varyansı göstermektedir. $\text{var}(\hat{b}_{MLE}) = (X'WX)^{-1}$ ve $(X'WX)^{-1} = T \Lambda^{(*)-1} T' = T_s \Lambda_s^{(*)-1} T_s' + T_r \Lambda_r^{(*)-1} T_r'$ olduğundan (3.18) tekrar düzenlenirse,

$$\text{var}(\hat{b}_s^{pc}) = T_s \Lambda_s^{(*)-1} T_s' \quad (3.19)$$

olarak elde edilir. $\hat{b}_s^{pc}$ 'ye ilişkin yanlılık ise $\text{bias}(\hat{b}_s^{pc}) = E(\hat{b}_s^{pc}) - b = -T_r a_r$ şeklinde verilmektedir. $\hat{b}_s^{pc}$ için MSE ise;

$$MSE(\hat{b}_s^{pc}) = T_s \Lambda_s^{(*)-1} T_s' + (T_r a_r)(T_r a_r)'$$

şeklinde matris formunda yazılabilir.

3.4. Poisson Regresyonda İstatistiksel Çıkarımlar

3.4.1. Sapma (Deviance) ve Uyumun İyiliği

Genelleştirilmiş lineer modeller için Bölüm 2.4.1'de anlatılanların paralelinde Poisson regresyon için özel hali yazılacak olursa,

$$\begin{aligned}
D = \text{Deviance} &= -2 \ln \left[\frac{L(b)}{L(P)} \right] = -2 [L(\hat{m}; y) - L(y; y)] \\
&= -2 \ln [L(\text{uydurulmuş model}) - L(\text{doymuş model})] \\
&= -2 \sum_{i=1}^n \left(\frac{[y_i \hat{h}_i - b(\hat{h}_i)]}{a(j)} - \frac{[y_i h_i - b(h_i)]}{a(j)} \right) \\
&= -2 \sum_{i=1}^n ([y_i \ln(\hat{l}_i) - \hat{l}_i] - [y_i \ln(l_i) - l_i])
\end{aligned}$$

şeklinde olur. Karar kuralı ise Bölüm 2.4.1’de bahsedildiği gibi olacaktır.

3.4.2. Modeldeki Katsayılar için Önemlilik Testi

Modeldeki katsayıların önemliliğinin test edilmesi için kurulacak hipotez testi, test istatistikleri ve karar kuralı genelleştirilmiş lineer modeller için Bölüm 2.4.2’de anlatıldığı gibidir.

3.4.3. Model Karşılaştırma

Genelleştirilmiş lineer modeller için Bölüm 2.4.3’de anlatılan model karşılaştırma yöntemi Poisson regresyon için de geçerlidir.

3.4.4. Poisson Regresyon İçin Rezidüler

3.4.4.1. Sapma (Deviance) Rezidü

Deviance rezidüye ait genel form eşitlik (2.41) ile verilmektedir. Genel formu oluşturan terimler, Poisson regresyon için, $w_i = 1$, $h_i = \ln(y_i)$, $\hat{h}_i = \ln l_i$, $b(h_i) = y_i$ ve $b(\hat{h}_i) = l_i$ ’dir. Bu ifadeler kullanılarak, i -inci gözleme ait sapma (deviance) rezidü ise,

$$\text{sign}(y_i - l_i) \sqrt{2[y_i(\ln(y_i) - \ln(l_i)) - y_i + l_i]}$$

olarak elde edilir.

3.4.4.2. Pearson Rezidü

Poisson regresyon için, i -inci gözleme ait Pearson rezidü ise,

$$e_i = \frac{y_i - l_i}{\sqrt{l_i}}$$

olarak ifade edilebilir.

4. LOJİSTİK REGRESYON

Lojistik regresyon, yanıt değişkeninin lineer regresyondakinin aksine sürekli değil iki ya da daha fazla değer aldığı kesikli durumlarda kullanılmakta ve diğer regresyon analizleri gibi açıklayıcı değişkenler ile yanıt değişkeni arasında bir model kurmayı amaçlamaktadır. Son yıllarda, veri yapısının getirdiği özellik itibarıyla, lojistik regresyon tıp, biyoloji, ekonomi ve çeşitli birçok alanda kullanılmaya başlanmıştır.

Berkson, 1944 yılında yaptığı bir çalışmada lojistik fonksiyonu biyolojik deneylerin analizinde kullanmıştır. Bu çalışma, lojistik modelin literatürdeki ilk uygulamasıdır. Cox (1970) ise Berkson'nun (1944) önerdiği bu modeli ele alarak farklı uygulamalar yapmıştır.

Ayrıca, lojistik regresyonun ilk kullanımlarından sonra literatürde lojistik regresyonu diğer analiz yöntemleri ile karşılaştıran çalışmalar da bulunmaktadır. Bunların arasında 1971 yılında Halperin ve arkadaşlarının yaptığı ve Finney (1971) tarafından yapılan çalışma gösterilebilir.

Tıp ve epidemiyoloji gibi birçok çalışma alanında binary (ikili) yanıt değişkenini ya da bir olayın ortaya çıkma olasılığını o değişkenle ilişkili olan açıklanabilir değişken kümesiyle tahmin edebilmek oldukça önemlidir. Sağlık alanında yapılan çalışmalara, O'Neill ve Barry (1995) trafik kazalarına ait verileri ele alarak, kazalarda ölüme sebep olabilecek yaş, cinsiyet gibi açıklayıcı değişkenlerinin etkilerini inceledikleri çalışma ve Bircan (2004)'nın çocuklarda doğum ağırlığını etkileyen önemli risk faktörlerini belirlemek için tıp verileri üzerindeki çalışma örnek olarak gösterilebilir. Bunlara ek olarak, Gardside ve Glueck (1995) insanlarda beslenme şekli, sigara ve alkol kullanımı, fiziksel aktivite gibi risk faktörlerinin kalp hastalığı üzerindeki etkilerini inceledikleri çalışma, Kloiber ve ark (1996), Peoples ve ark. (1991), Buescher ve ark. (1993)'nın kadınlarda düşük doğum ağırlığını etkileyen risk faktörlerini, Santos ve ark. (1998) kafein tüketimi ve düşük doğum ağırlığı arasındaki ilişkiyi, Sable ve Herman (1997) erken doğum ve düşük doğum ağırlığı arasındaki ilişkiyi inceledikleri çalışmalar gösterilebilir (Bircan, 2004).

Lojistik regresyon analizi popüler hale geldikçe çok farklı kullanım alanlarında örnekleri çoğalmıştır. Bu alanlardan bir tanesi de sosyal verilerin incelenmesidir. Heath ve ark. (2000) tarafından yapılan çalışmada etnik kimlik farklılıklarının işgücü piyasasındaki etkilerini 1991 UK İşgücü Araştırması verilerini lojistik regresyonla analiz etmişlerdir. Demaris (1995) ise 1993 yılında yapılan Genel Sosyal Araştırma verilerini deneklerin kişisel mutluluklarını lojistik regresyon ile analiz etmiştir ve sonuçlarını tartışmıştır.

Lojistik regresyon bir diğer kullanım alanı ise tarım ve çevre olmuştur. Bu alanda da birçok çalışma yapılmıştır. Örneğin, Önder (2001) tarafından yapılan çalışmada lojistik regresyonda değişken seçimi ele alınmış ve koyunların süten kesilme ağırlıkları ile ilgili veri setinde analiz yapılmıştır. Diğer bir örnek çalışmayı ise, Cankurt ve ark. (2007) Türkiye'nin Avrupa Birliği'ne üyelik potansiyelini tarımsal ve diğer göstergelerle lojistik regresyon ile tahmin edilmesi hakkında yapmışlardır. Bunların dışında, Oğuz ve Budak (2005) Mersin ilindeki pilot geri kazanım programında elde edilen verilerden ailelerin geri kazanım programına katılımlarına etki eden faktörleri lojistik regresyon modeli ile analiz etmiş ve sonuçlarını tartışmışlardır.

Şüphesiz ki, lojistik regresyonun en önemli ve yaygın kullanım alanlarından bir tanesi de ekonomi ve finans alanıdır. Bu alanda da çok sayıda çalışma görmek mümkündür. Örneğin, Akkaya ve Pazarlıoğlu (1998) tarafından ekonomi alanında lojistik regresyon modellerinin kullanımı örneklerle incelenmiştir. Özçomak ve ark. (2006) yaptıkları çalışmada Erzurum ilinde hane halklarına uygulanan anketten elde edilen verilere lojistik regresyon analizi uygulanmış ve doğal gazın talep edilmesinde etkili olan faktörler tahmin edilmiştir. Bunlara ek olarak, Tatlıdil ve Özer (2005) tarafından yapılan çalışmada bankacılıkta firma (kredi) derecelendirme çalışmaları konusunda lojistik regresyon analize dayalı karar destek sistemlerinin kullanımı konusu tartışılmıştır.

Sonuç olarak, lojistik regresyon analizi sağlık, tıp, sosyal, ekonomi, finans, tarım ve çevre gibi birçok farklı alanda kullanılmaktadır ve bu alanlarda yapılmış birçok çalışma bulunmaktadır.

Lojistik regresyonun teorik gelişim süreci ve bu yöntemin teorisine ilişkin yapılmış çalışmalar incelenirse:

Lojistik regresyonda tahmin yöntemlerine ilişkin, Albert ve Anderson (1984) tarafından yapılan çalışmada farklı veri setlerinde lojistik regresyon katsayılarının maksimum likelihood metoduyla tahminlerine ait teorik temelleri anlatmışlardır. Bonney ise 1987 yılında yaptığı çalışmada geniş bir şekilde lojistik regresyonun gelişimini ve kullanımını anlatmayı amaçlamıştır. Bunun yanı sıra, verilerin tahmin edilen lojistik regresyon modeline uyumuna ilişkin çalışmalar da yapılmıştır. Aranda-Ordaz (1981) ve Johnson(1985) tarafından yapılan çalışmalar en önemlileri sayılabilirler. Ayrıca, Pregibon (1981) yanıt değişkenin ikili değer aldığı lojistik modellerde etkin ve aykırı gözlemlerin belirlenmesi hakkında bir çalışma yapmıştır. Hosmer ve Lemeshow (1980) tarafından yapılan çalışmada ise uyumun iyiliği ölçütü geliştirilmiştir. Daha sonra, Hosmer ve Lemeshow (2000) tarafından lojistik regresyon modeli ayrıntılı bir şekilde incelenmiş, maksimum likelihood tahmin yöntemi, uyumun iyiliği ölçütleri, etkin ve aykırı gözlemlerin belirlenmesine dair yapılan çalışmalar özetlenmiş ve çeşitli örnek uygulamalar yapılmıştır. Buna ek olarak, Roy ve Guria (2008) yaptıkları çalışmada gözlemlerin çıkarılmasına dayalı tekniği kullanarak, lojistik regresyon için etkin ve aykırı gözlemleri incelemişlerdir.

Özdamar (2004) ise lojistik regresyon analizini SPSS, Minitab ve NCSS istatistik paket programlarında uygulamalı olarak göstermiştir.

Açıklayıcı değişkenler arasında çoklu iç ilişkinin bulunduğu durumlarda ise maksimum likelihood tahmin edici kararsız parametre tahminleri üretecektir ve kurulan modelden yapılacak çıkarımlarda hataya sebep olacaktır. Bu durum, araştırmacıları oluşan kararsız tahminleri daha kararlı hale getirecek yeni tahmin metotları geliştirmeye sürüklemiştir.

Lojistik regresyonda çoklu iç ilişkiden kaynaklanan problemleri azaltmaya yönelik, ilk olarak bazı basit öneriler, Gordon (1974) tarafından önerilen “değişken atımı” ve Cox (1970) tarafından önerilen “basit değişken dönüşümleri”dir.

Schaefer, Roi ve Wolfe (1984) ise lojistik regresyon için “ridge türü bir tahmin edici” geliştirmişlerdir ve bu ridge tahmin edicinin, maksimum likelihood tahmin ediciye göre daha küçük bir MSE değerine sahip olduğunu göstermişler,

ayrıca ampirik bir örneğin sonuçlarını tartışmışlardır. Daha sonra, Schaefer (1986) tarafından lojistik regresyonda çoklu iç ilişkinin olduğu durumlarda kullanılacak tahmin edicileri ridge tahmin edici, tek basamaklı (yani, en düşük özdeğere sahip gözlemlerin tek seferde atılmasına dayalı) temel bileşen tahmin edici ve Stein tahmin edici olarak önerilmiştir. Ayrıca, bu tahmin ediciler bir simülasyon örneğinde kullanılmış ve sonuçları tartışılmıştır.

Bernoulli dağılımını da kapsayan genelleştirilmiş lineer modellerdeki çoklu iç ilişki hakkında, Mackinnon ve Puterman 1989 yılında bir çalışma yapmışlardır. Bu çalışmada, genelleştirilmiş lineer modellerde çoklu iç ilişkinin tanımı yapılmış, çoklu iç ilişkinin doğurduğu sonuçlar araştırılmış ve teşhis kriterleri önerilmiştir.

Daha sonra, Schaefer (1986)'in önerdiği tek basamaklı temel bileşen tahmin ediciyi, Marx ve Smith (1990) geliştirerek lojistik regresyonu da içeren genelleştirilmiş lineer regresyon modelleri için “iteratif temel bileşenler tahmin edici”yi önermişlerdir. Ayrıca, biyoistatistik alanında bir uygulama yapmışlardır.

Bunun yanı sıra, Lesaffre ve Marx (1993) tarafından yapılan bir çalışmada, genelleştirilmiş lineer modellerde açıklayıcı değişkenlerin arasındaki çoklu iç ilişkiden bahsedilmiştir. Bu problem için tespit yöntemleri sunulmuş ve örnek bir uygulama yapılmıştır.

Bu çalışmaların dışında, Le Cessie ve Van Houwelingen (1992) tarafından kanserli DNA hücreleri ile ilgili elde edilen veri kümesi üzerinde ridge lojistik tahmin edici ile analiz çalışması yapmışlardır. Bu çalışmada, ridge parametresinin seçiminde çapraz geçerlilik yöntemi üzerinde durmuşlardır. Çalışmada ayrıca, ön tahmin hatası için üç farklı yöntem sunmuşlardır. Bunlar, sınıflandırma hatası, hata kareleri ve eksi log-likelihood'dur.

Barker ve Brown (2001) ise, lojistik regresyonda çoklu iç ilişkinin olduğu durumlarda MLE lojistik tahmin edici, ridge lojistik tahmin edici ve temel bileşenler lojistik tahmin ediciyi bir simülasyon örneği ile MSE kriterine göre karşılaştırmış ve sonuçlarını tartışmışlardır.

Ridge türü lojistik regresyon tahmin edicinin kullanımına dair diğer bir çalışmayı ise Vágó ve Kemény (2006) sunmuşlardır. Bu çalışmada, sürekli ve kesikli

açıklayıcı değişkenlerin bulunduğu bir biyomedikal veri kümesine ridge türü lojistik bir tahmin edici uygulanarak analizler yapılmış ve sonuçları tartışılmıştır.

Bunun yanı sıra, Aguilera ve ark. (2006) tarafından yapılan çalışmada yüksek boyutta çoklu iç ilişkinin olması durumunda lojistik regresyon tahmininde temel bileşenler yönteminin kullanılması incelenmiştir. Bir simülasyon örneğinde farklı optimum temel bileşen seçimi metotları karşılaştırmalı olarak uygulanmıştır.

4.1. Lojistik Regresyon Modeli

Yanıt değişkenin sadece 0 ve 1 gibi iki mümkün değerinin olduğu bir regresyon durumunu varsayalım. Bu durumda y_i yanıt değişkeni Bernoulli dağılımına sahip bir değişken olacaktır. $E(e_i) = 0$ olmak üzere, model

$$y_i = x_i' b + e_i \quad (4.1)$$

olarak yazılabilir. Burada $x_i' = [1, x_{i1}, x_{i2}, \dots, x_{ip}]$, $b' = [b_0, b_1, b_2, \dots, b_p]$ ve y_i ise Bernoulli rasgele değişkeni olduğundan Tablo 4.1'deki olasılık dağılımına sahip olacaktır.

Tablo 4.1. Bernoulli olasılık dağılımı

y_i	Olasılık
1	$P(y_i=1) = p_i$
0	$P(y_i=0) = 1 - p_i$

y_i nin beklenen değeri tablodan yazılırsa,

$$E(y_i) = 1.(p_i) + 0.(1-p_i) = p_i \quad (4.2)$$

olur. (4.1) modelinin beklenen değerine bakılırsa,

$$E(y_i) = E(x_i' b + e_i) = E(x_i' b) + E(e_i) = E(x_i' b) + 0 = x_i' b \quad (4.3)$$

dir. (4.2) ve (4.3) birlikte göz önünde bulundurulursa,

$$E(y_i) = x_i' b = p_i \quad (4.4)$$

olur. Bu da yanıt değişkenin 1 olduğu zaman aldığı olasılığa eşit olur (Tablo 4.1). Modelin varyansı incelenecek olursa, $E(y_i^2) = 1^2 p_i + 0^2 (1 - p_i) = p_i$ olmak üzere,

$$\begin{aligned} s_{y_i}^2 &= E(y_i^2) - [E(y_i)]^2 = p_i - p_i^2 \\ &= p_i(1 - p_i) \\ &= E(y_i)(1 - E(y_i)) \\ &= (x_i' b)(1 - x_i' b) \end{aligned}$$

olduğu görülür.

Eşitlik (4.4) kullanılarak, eşitlik (4.1) tekrar düzenlenirse, $y_i = E(y_i) + e_i = p_i + e_i$ şeklinde yazılabilir. Buradan hata terimi,

$$e_i = y_i - p_i$$

şeklinde yazılabilir. Hata teriminin beklenen değeri ise,

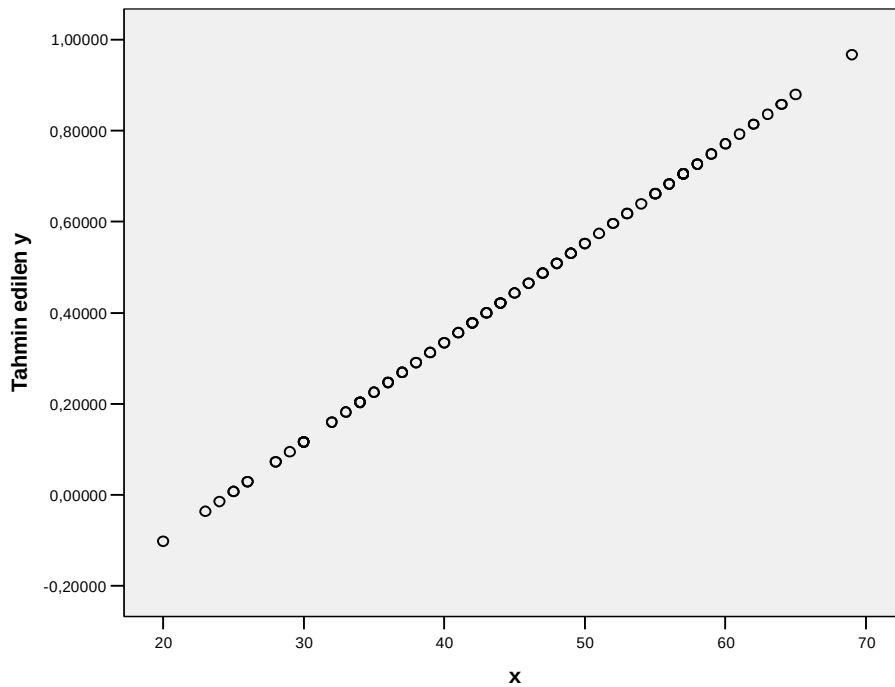
$$E(e_i) = E(y_i - p_i) = E(y_i) - E(p_i) = p_i - p_i = 0$$

olur. Ayrıca, p_i sabit bir değer olduğundan, hata teriminin varyansı,

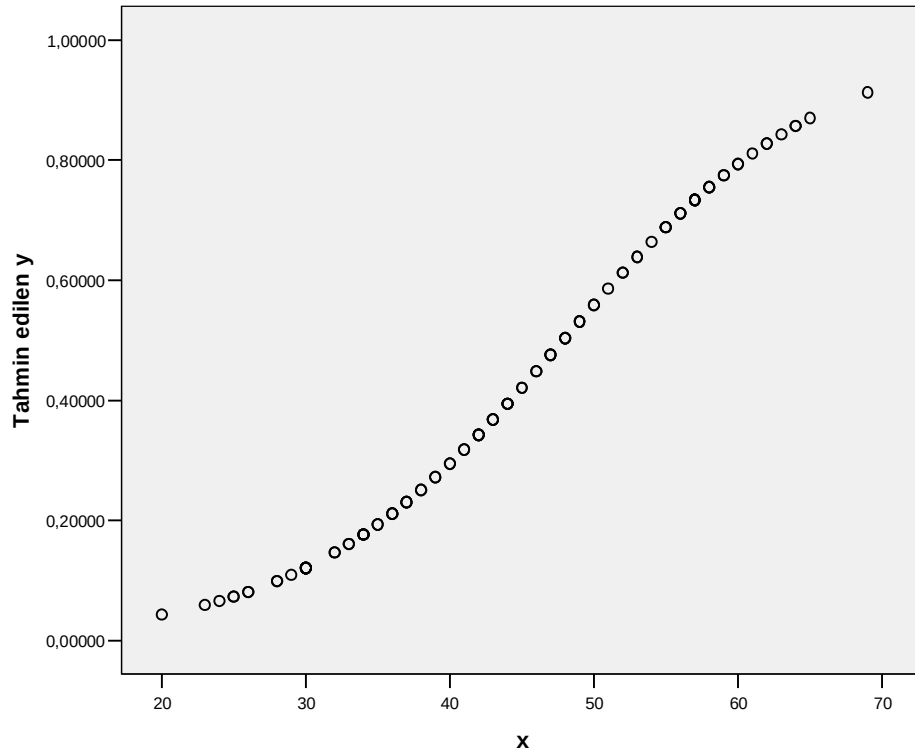
$$\text{var}(e_i) = \text{var}(y_i - p_i) = \text{var}(y_i) = E(y_i)(1 - E(y_i))$$

olur. Yani, hata terimlerinin varyansı da beklenen değerin bir fonksiyonudur ve bu da sabit varyanslılık durumunun olmadığını göstermektedir. Diğer bir deyişle $E(e) = 0$ ve $\text{var}(e) = s^2 V$ olduğundan ($V = \text{diag}(p_i(1-p_i))$), sabit varyanslılık varyasyonu sağlanamadığından doğrusal regresyon kullanılamamakta ve lineer forma getirilerek sabit varyanslılık durumu sağlanabilmektedir.

Bunun yanı sıra, yanıt değişken ikili (Bernoulli) verilerden oluşuyorken lineer regresyon kullanılırsa, çeşitli problemler ortaya çıkmaktadır. Bunlardan birisi de uydurulan yanıt değişkeninin değeri 0 ile 1 arasında olması zorunluluğunun olmamasıdır. Şekil 4.1’de de görüleceği gibi eğer ikili veri tipi için lineer regresyon uygulanırsa, uydurulan yanıt değişkeni negatif değer alabilmektedir. Halbuki bir olasılık değerinin negatif olabilmesi imkansızdır. Uydurulan yanıt değişkeni (0,1) aralığında olması gerektiğinden, tahminleri bu aralığa çekmek için lojistik yanıt fonksiyonu kullanılarak (Şekil 4.2) lineer form dönüşümü yapılmalıdır. Lojistik yanıt fonksiyonu monoton artan ya da azalan S-şeklinindedir (S-Shape) (Şekil 4.2).



Şekil 4.1. Lineer regresyonda, uydurulmuş yanıt değişkeni ve x değişkeni



Şekil 4.2. Lojistik regresyon yanıt fonksiyonu (S-Shape)

Lojistik regresyonda modeli sabit varyanslı hale getirebilmek ve uydurulan yanıt değişkeni (0,1) aralığına çekebilmek için logit dönüşüm yapılarak lineer forma getirilir:

Logit dönüşüm: $h_i = x_i' b$

$$h_i = \ln\left(\frac{p_i}{1-p_i}\right)$$

$$x_i' b = \ln\left(\frac{p_i}{1-p_i}\right) \quad (4.5)$$

Eşitlik (4.5)'in her iki tarafı üstel paranteze alınır,

$$\exp(x_i' b) = \exp \left[\ln \left(\frac{p_i}{1-p_i} \right) \right]$$

ifadesi buradan da

$$\exp(x_i' b) [1 - p_i] = p_i$$

elde edilir. $\exp(x_i' b)$ ifadesi parantez içine dağıtılırsa,

$$\exp(x_i' b) = p_i + p_i (\exp(x_i' b)) = p_i (1 + \exp(x_i' b))$$

olur. Son olarak, p_i yalnız bırakılırsa,

$$p_i = \frac{\exp(x_i' b)}{1 + \exp(x_i' b)} \quad (4.6)$$

elde edilir. Diğer bir ifadeyle,

$$E(y_i) = p_i = \frac{e^{x_i' b}}{1 + e^{x_i' b}} = \frac{1}{1 + e^{-x_i' b}} \quad (4.7)$$

olur. Böylece lojistik regresyon modeli elde edilmiş olur.

Görüldüğü gibi lojistik regresyonda da, doğrusal regresyon analizinde olduğu gibi bazı açıklayıcı değişken değerlerine dayanarak tahmin yapılmaya çalışılır. Ancak bu iki yöntem arasında üç önemli fark vardır:

1. Doğrusal regresyon analizinde tahmin edilecek olan yanıt değişken sürekli iken, lojistik regresyon analizinde yanıt değişken kesikli bir değer almaktadır.
2. Doğrusal regresyon analizinde yanıt değişkenin değeri tahmin edilirken lojistik regresyon analizinde ise yanıt değişkenin alabileceği değerlerin gerçekleşme olasılığı tahmin edilir.

3. Doğrusal regresyon analizinde sonuç çıkarım için yanıt değişkeninin normal dağılım göstermesi şartı aranırken, lojistik regresyon analizinde böyle bir şart yoktur (Elhan, 1997).

4.2. Lojistik Regresyon Modelinde Maksimum Likelihood Parametre Tahmini

4.2.1. Gruplandırılmamış Veri İçin Lojistik Regresyon Modeli Parametre Tahmini

Lojistik regresyonun genel formu eşitlik (4.7) ile verilmiştir. Her bir gözlem Bernoulli dağılımından bağımsız birer gözlem olduğundan, olasılık fonksiyonu

$$f(y_i) = p_i^{y_i} (1 - p_i)^{1-y_i} \quad i = 1, 2, \dots, n \text{ ve } y_i = 0, 1$$

olarak yazılabilir. Likelihood fonksiyonu ise

$$L(y, b) = \prod_{i=1}^n f(y_i) = \prod_{i=1}^n p_i^{y_i} (1 - p_i)^{1-y_i}$$

şeklindedir. Buradan Log-likelihood fonksiyonu,

$$\begin{aligned} \mathbf{l}(b) &= \sum_{i=1}^n [y_i \ln(p_i) + (1 - y_i) \ln(1 - p_i)] \\ &= \sum_{i=1}^n y_i \ln\left(\frac{p_i}{1 - p_i}\right) + \sum_{i=1}^n \ln(1 - p_i) \\ &= \sum_{i=1}^n y_i x_i' b - \sum_{i=1}^n \ln(1 + e^{x_i' b}). \end{aligned} \tag{4.8}$$

elde edilir. Maksimum likelihood tahmin bulunmak istendiğinden

$$\frac{\partial \mathbf{l}(b)}{\partial b} = \sum_{i=1}^n (y_i - \frac{e^{x_i' b}}{1 + e^{x_i' b}}) x_i = \sum_{i=1}^n (y_i - p_i) x_i = 0 \quad (4.9)$$

denkleminin b için çözülmesi gerekmektedir. Fakat $\mathbf{l}(b)$, b 'nin lineer bir formunda olmadığından doğrudan çözülemez ve Newton-Raphson metodu yardımıyla çözüm bulmaya çalışılır (EK 2). Eşitlik (2.18) ile verilen Taylor açılımı GLM için eşitlik (2.19) ile verilmişti. Bu eşitlikteki ifadeler, lojistik regresyon için yazılacak olursa,

$$\frac{\partial \mathbf{l}(b)}{\partial b'} = \sum_{i=1}^n (y_i - \frac{e^{x_i' b}}{1 + e^{x_i' b}}) x_i' \quad (4.10)$$

eşitliğinden yararlanılırsa

$$\frac{\partial^2 \mathbf{l}(b)}{\partial b \partial b'} = - \sum_{i=1}^n \frac{e^{x_i' b}}{(1 + e^{x_i' b})^2} \cdot x_i x_i' = - \sum_{i=1}^n p_i (1 - p_i) x_i x_i' \quad (4.11)$$

elde edilir.

$$x_i' b = b_0 + b_1 x_{i1} + \dots + b_p x_{ip}$$

$$X' = [x_1, x_2, \dots, x_n]_{(p+1) \times n}$$

$$Y' = [y_1, y_2, \dots, y_n]_{1 \times n}$$

$$P' = [p_1, p_2, \dots, p_n]_{1 \times n}$$

$$V = \text{diag}(p_i (1 - p_i))_{n \times n}$$

olmak üzere (4.11) ve (4.9) eşitlikleri aşağıdaki gibi yazılır:

$$\frac{\partial^2 \mathbf{l}(b)}{\partial b \partial b'} = -X' V X \quad (4.12)$$

ve

$$\frac{\partial \mathbf{l}(b)}{\partial b} = X'Y - X'P \quad (4.13)$$

(4.12) ve (4.13) eşitlikleri ise (4.10) de yerine yazılırsa,

$$b_{MLE} = b + (X'VX)^{-1} X' (Y - P) \quad (4.14)$$

bulunur. Newton-Raphson metodu iteratif bir yöntem olup, gerçek köke yaklaşılmak istendiğinden iteratif yöntem

$$b^{(m+1)} = b^{(m)} - \left[\frac{\partial^2 \mathbf{l}(b)}{\partial b \partial b'} \right]^{-1} \cdot \frac{\partial \mathbf{l}(b)}{\partial b}$$

eşitliğinin uygulaması ile elde edilir:

$$b^{(m+1)} = b^{(m)} + (X'V_m X)^{-1} X' (Y - P_m) \quad (4.15)$$

Burada, $P_m' = (p_{1,m}, p_{2,m}, \dots, p_{n,m})$ önceki iterasyondan elde edilen uydurulmuş yanıt olasılıklar vektörü ve $V_m = \text{diag}(p_{i,m}(1-p_{i,m}))$ önceki iterasyonda elde edilen köşegen matristir. (4.15) eşitliği, $(X'V_m X)^{-1} X'V_m$ ortak parantezine alınarak yeniden düzenlenirse,

$$b^{(m+1)} = (X'V_m X)^{-1} X'V_m (Xb^{(m)} + V_m^{-1}(Y - P_m))$$

yazılabilir. $Q_a = Xb^{(m)} + V_m^{-1}(Y - P_m)$ düzeltilmiş yanıt (adjusted response) olmak üzere,

$$b^{(m+1)} = (X'V_m X)^{-1} X'V_m Q_a \quad (4.16)$$

olarak elde edilir. (4.16) eşitliği ağırlıklandırılmış en küçük kareler formülasyonuna benzemektedir. Şu anki $b^{(m)}$ tahminleri kullanılarak her iterasyonda $p_i(1-p_i)$ ağırlıkları güncellenmektedir. Böylelikle, Yeniden Ağırlıklandırılmış İteratif En Küçük Kareler Yöntemi (IRLS) ile $b^{(m)}$, $b^{(m+1)}$ 'e yeterince yakın olana kadar iterasyona devam edilir. Yakınsama ölçüsü olarak, istatistik analiz paket programlarında $b^{(m+1)}$ ve $b^{(m)}$ arasındaki farkın 1×10^{-8} , 1×10^{-7} , 1×10^{-6} gibi değerler alması kullanılmaktadır. Örneğin, SAS paket programında 1×10^{-8} (1E-8) olarak kullanılmaktadır (SAS Institute Inc., 2008).

4.2.2. Gruplandırılmış Veri İçin Lojistik Regresyon Modeli Parametre Tahmini

Lojistik regresyon modellerinin bazı uygulamalarında x değişkenlerinin her seviyesinde tekrarlı gözlemler veya denemeler olabilir (Tablo 4.2).

n -birimlik gözlem ($i = 1, 2, \dots, n$) x 'in her bir düzeyini göstermektedir.

$y_{ji} \sim \text{Bernoulli}(p_i)$ dağılımına sahip olmak üzere $y_i = \sum_{j=1}^{n_i} y_{ji}$ tanımlanırsa

$y_i \sim \text{Binom}(p_i, n_i)$ olur. i -inci gruptaki n_i gözlemden y_i tanesinin sonucu olumlu

yani $y_{ji} = 1$ ise başarı oranı $p_i = \frac{y_i}{n_i}$ olacaktır.

Tablo 4.2. Gruplandırılmış veri yapısı

x_i	y_i		Başarı Sayısı
1	y_{11}	Gözlenen 1'lerin, yani başarıların sayısı	y_1
	y_{21}		
	M		
	$y_{n_1 1}$		
2	y_{12}	Gözlenen 1'lerin, yani başarıların sayısı	y_2
	y_{22}		
	M		
	$y_{n_2 2}$		
M	M	M	M
N	y_{1n}	Gözlenen 1'lerin, yani başarıların sayısı	y_n
	y_{2n}		
	M		
	$y_{n_n n}$		

Bu durumda, log-likelihood fonksiyonu

$$\begin{aligned}
 \mathbf{l}(b) &= \ln L(y, b) = \ln \prod_{i=1}^n p_i^{y_i} (1-p_i)^{n_i-y_i} \\
 &= \sum_{i=1}^n y_i \ln p_i + \sum_{i=1}^n n_i \ln(1-p_i) - \sum_{i=1}^n y_i \ln(1-p_i)
 \end{aligned}$$

şeklinde olur. Newton-Raphson metodu kullanarak tahmin edici bulunursa zincir kuralından

$$\begin{aligned}
\frac{\partial \mathbf{l}(b)}{\partial b} &= \frac{\partial \mathbf{l}(b)}{\partial p_i} \cdot \frac{\partial p_i}{\partial b} \\
&= \left[\sum_{i=1}^n \frac{y_i}{p_i} - \sum_{i=1}^n \frac{n_i}{1-p_i} + \sum_{i=1}^n \frac{y_i}{1-p_i} \right] p_i (1-p_i) x_i \\
&= \left[\sum_{i=1}^n \frac{y_i - n_i p_i}{p_i (1-p_i)} \right] [p_i (1-p_i) x_i]
\end{aligned}$$

bulunur. $m' = [n_1 p_1, n_2 p_2, \dots, n_n p_n]$ olmak üzere

$$\frac{\partial \mathbf{l}(b)}{\partial b} = \sum_{i=1}^n (y_i - m_i) x_i$$

$$\frac{\partial \mathbf{l}(b)}{\partial b'} = \sum_{i=1}^n (y_i - m_i) x_i'$$

yazılabilir. $\frac{\partial p_i}{\partial b} = p_i (1-p_i) x_i$ eşitliğinin kullanılması ile

$$\frac{\partial^2 \mathbf{l}(b)}{\partial b \partial b'} = \frac{\partial}{\partial b} \left[\sum_{i=1}^n (y_i - n_i p_i) x_i' \right] = \sum_{i=1}^n \left[-n_i \left(\frac{\partial p_i}{\partial b} \right) x_i' \right] = - \sum_{i=1}^n n_i p_i (1-p_i) x_i x_i'$$

elde edilir. $V^* = \text{diag}[n_i p_i (1-p_i)]$ alınır, $\frac{\partial^2 \mathbf{l}(b)}{\partial b \partial b'} = -X' V^* X$ sonucuna ulaşılır ve

$\frac{\partial \mathbf{l}(b)}{\partial b} = X' (Y - m)$ ile birlikte Newton-Raphson genel formunda yerine koyulursa,

$$b^{(m+1)} = b^{(m)} + (X' V_m^* X)^{-1} X' (Y - m_m) \quad (4.17)$$

olur. Burada, $m'_m = [n_{1,m}p_{1,m}, n_{2,m}p_{2,m}, \dots, n_{n,m}p_{n,m}]$ önceki iterasyondan elde edilen uydurulmuş yanıt olasılıklar vektörü ve $V_m^* = \text{diag}[n_{i,m}p_{i,m}(1-p_{i,m})]$ önceki iterasyonda elde edilen köşegen matristir. (4.17) eşitliği yeniden düzenlenirse,

$$b^{(m+1)} = (X'V_m^*X)^{-1}X'V_m^*(Xb^{(m)} + V_m^{*-1}(Y - m_m))$$

şeklinde yazılabilir. $Q_b = Xb^{(m)} + V_m^{*-1}(Y - m_m)$ düzeltilmiş yanıt olmak üzere,

$$b^{(m+1)} = (X'V_m^*X)^{-1}X'V_m^*Q_b \quad (4.18)$$

olarak bulunur. (4.18) eşitliği ağırlıklandırılmış en küçük kareler formülasyonuna benzemektedir. Şu anki, $b^{(m)}$ tahminleri kullanılarak her iterasyonda $n_i p_i (1 - p_i)$ ağırlıkları güncellenmektedir. Bu yöntemle, yani “yeniden ağırlıklandırılmış iteratif en küçük kareler yöntemi (IRLS)” ile $b^{(m)}$, $b^{(m+1)}$ ’e yeterinde yakın olana kadar iterasyona devam edilir.

4.3. Lojistik Regresyonda Bazı Yanlı Tahmin Ediciler

Lojistik regresyonda da çoklu iç ilişki problemi ortaya çıkabilmektedir. Schaefer, Roi ve Wolfe (1984) çoklu iç ilişkinin varlığının aşağıdaki şekilde belirlenebileceğini göstermiştir:

1. R_j^2 , j-inci gözlemin diğer gözlemler üzerinde regresyon uydurulması ile elde edilen belirleyicilik katsayısı
 2. $(d_j' d_j)$, (1). maddede uydurulan regresyonun rezidü kareler toplamı
 3. $X'VX$ in en küçük özdeğeri
- incelenir. R_j^2 bazı j’ler için 1’e yaklaşırsa; $(d_j' d_j)$ bazı j’ler için 0’a yaklaşırsa ve $X'VX$ in en küçük özdeğeri sıfıra yaklaşırsa çoklu iç ilişkinin var olduğu sonucuna

varılır. Bu nedenle lojistik regresyonda MLE yöntemine alternatif tahmin ediciler önerilmiştir.

4.3.1. Ridge Lojistik Tahmin Edici

Hata kareler toplamını belli bir sabite eşit olan tahmin ediciler içerisinde boyu en küçük olan tahmin ediciye, “ridge lojistik tahmin edici” denir. Lojistik regresyonda parametre tahmini genellikle $\hat{b}_{MLE} = (X'VX)^{-1} X'VQ$ ((4.16) ve (4.18) e dikkat edilirse Q yerine Q_a ve Q_b gelmiştir) ile bulunur. Çoklu iç ilişki olması durumunda $X'VX$ matrisinin tersi alınamayabilir. Bu durumda, bu matrisin köşegen elemanlarına kI eklenerek tersi bulunur. Böylece lojistik regresyon için ridge tahmin edici elde edilir.

Schaefer, Roi ve Wolfe (1984) ridge lojistik tahmin ediciyi şu şekilde önermiştir:

$\hat{b}$ yeniden ağırlıklandırılmış iteratif en küçük kareler (IRLS) olduğundan ağırlıklı hata kareler toplamını yaklaşık olarak minimum yapar. $\hat{b}$ yerine $\tilde{b}$ kullanılırsa, ağırlıklı hata kareler toplamındaki artış j gibi bir değer olur.

$\hat{p}(\hat{b})$, $\hat{b}$ 'nin kullanılmasıyla elde edilen p 'nin tahmini olmak üzere, $(Y - \hat{p}(\tilde{b}))'V(Y - \hat{p}(\tilde{b}))$ ifadesine $\hat{p}(\hat{b})$ ekleyip çıkarılırsa,

$$\begin{aligned} (Y - \hat{p}(\tilde{b}))'V(Y - \hat{p}(\tilde{b})) &= [Y - \hat{p}(\hat{b}) + \hat{p}(\hat{b}) - \hat{p}(\tilde{b})]'V[Y - \hat{p}(\hat{b}) + \hat{p}(\hat{b}) - \hat{p}(\tilde{b})] \\ &= (Y - \hat{p}(\tilde{b}))'V(Y - \hat{p}(\tilde{b})) + 2(Y - \hat{p}(\tilde{b}))'V(\hat{p}(\hat{b}) - \hat{p}(\tilde{b})) \\ &\quad + (\hat{p}(\hat{b}) - \hat{p}(\tilde{b}))'V(\hat{p}(\hat{b}) - \hat{p}(\tilde{b})) \end{aligned}$$

olur. $\hat{b}$ 'nin optimal olması nedeni ile

$$j = 2(Y - \hat{p}(\tilde{b}))'V(\hat{p}(\hat{b}) - \hat{p}(\tilde{b})) + (\hat{p}(\hat{b}) - \hat{p}(\tilde{b}))'V(\hat{p}(\hat{b}) - \hat{p}(\tilde{b}))$$

alınabilir. Eşitlikteki $\hat{p}$ 'ların lineerizasyonuyla,

$$j = (\tilde{b} - \hat{b})' X' V X (\tilde{b} - \hat{b})$$

elde edilir. j ile verilen kısıtlamaya bağlı olarak $\tilde{b}$ 'nın boyu minimum yapılmak istenirse Lagrange çarpanının kullanılmasıyla (EK 3),

$$F = \underset{\tilde{b}}{\text{Min}} \left\{ \tilde{b}' \tilde{b} + \frac{1}{k} \left[(\tilde{b} - \hat{b})' X' V X (\tilde{b} - \hat{b}) - c_0 \right] \right\} \quad (4.19)$$

ifadesi araştırılır. Burada c_0 herhangi bir sabittir. $\tilde{b}$ 'ya göre kısmi türev alınıp sıfıra eşitlenirse,

$$\frac{\partial F}{\partial \tilde{b}} = 2\tilde{b} + \frac{1}{k} 2X' V X (\tilde{b} - \hat{b}) = 0$$

parantez açılırsa,

$$2\tilde{b} + \frac{1}{k} 2X' V X \tilde{b} = \frac{1}{k} 2X' V X \hat{b}$$

olur. Her iki taraf k ile çarpılıp, eşitliğin sol tarafı $2\tilde{b}$ ortak parantezine alınırsa,

$2\tilde{b}(kI + X' V X) = 2X' V X \hat{b}$ elde edilir. Son olarak bu eşitlikten $\tilde{b}$ çekilirse,

$$\tilde{b} = \hat{b}_{\text{Ridge}} = (X' V X + kI)^{-1} X' V X \hat{b} \quad (4.20)$$

olarak “ridge lojistik tahmin edici” elde edilmiş olunur.

Le Cessie ve Van Houwelingen (1992), tarafından yapılan çalışmada ise Dufy ve Santner (1989)'in lojistik regresyon için kısıtlanmış maksimum likelihood kullanarak elde ettikleri ridge lojistik regresyon tahmin edicinin, Schaefer ve ark. (1984) tarafından sunulan ridge lojistik regresyon tahmin ediciye asimptotik olarak denk olduğunu göstermişlerdir:

Lojistik regresyon log-likelihood fonksiyonu (eşitlik (4.8)) ele alınsın. Dufy ve Santner (1989) tarafından bu log-likelihood fonksiyonundan minimum yanlılık üretecek şekilde b nin normu ($\|b\|^2 = b'b = \sum_{i=1}^n b_i^2$), d gibi bir sabit sayı ile çarpılarak, çıkarılmıştır:

$$\begin{aligned} \mathbf{l}'(b) &= \mathbf{l}(b) - d\|b\|^2 \\ &= \sum_{i=1}^n (y_i(x_i'b) - \ln(1 + e^{x_i'b})) - db'b \end{aligned} \quad (4.21)$$

Daha sonra, $\sum_{i=1}^n x_i'(y_i - p_i) = \frac{\partial \mathbf{l}(b)}{\partial b}$ olmak üzere, eşitlik (4.21) in b 'ya göre kısmi türevi alınarak:

$$\begin{aligned} \frac{\partial \mathbf{l}'(b)}{\partial b} &= \frac{\partial \mathbf{l}(b)}{\partial b} - \frac{\partial (db'b)}{\partial b} \\ &= \sum_{i=1}^n \left(x_i'y_i - x_i' \frac{e^{x_i'b}}{1 + e^{x_i'b}} \right) - 2db \\ &= \sum_{i=1}^n x_i'(y_i - p_i) - 2db \end{aligned} \quad (4.22)$$

elde edilir. I : Birim matris olmak üzere, eşitlik (4.22) ün b 'ya göre kısmi türevi alınarak,

$$\begin{aligned}
\frac{\partial^2 \mathbf{l}^l(b)}{\partial b^2} &= \left[\sum_{i=1}^n \left(x_i' y_i - x_i' \frac{e^{x_i' b}}{1 + e^{x_i' b}} \right) \right]' - (2db)' \\
&= 0 - \sum_{i=1}^n \frac{\partial \left(\frac{e^{x_i' b}}{1 + e^{x_i' b}} \right)}{\partial b} x_i - (2db)' \\
&= - \sum_{i=1}^n \frac{x_i' e^{x_i' b}}{(1 + e^{x_i' b})^2} x_i - 2dI \\
&= - \sum_{i=1}^n x_i' p_i (1 - p_i) x_i - 2dI
\end{aligned} \tag{4.23}$$

ifadesi eşitlik (4.21) nin b 'ya göre 2. dereceden türevi elde edilmiş olunur. Eşitlik (4.23) ifadesi tekrar düzenlenerek matris formunda yazılırsa

$$- \frac{\partial^2 \mathbf{l}^l(b)}{\partial b^2} = X' V(b) X + 2dI \tag{4.24}$$

elde edilir. Burada tanımlanan $V(b)$ matrisi Bölüm 4.2.1 de tanımlanan V matrisi ile aynı olup b ya bağlı olduğunu göstermek için $V(b)$ gösterimi kullanılmıştır. Eşitlik (4.21)'nin birinci dereceden b 'ya göre kısmi türevinin, Taylor serisi ile b parametresinin gerçek bir değeri olan b_0 etrafında birinci dereceden açılımı

$$\left. \frac{\partial \mathbf{l}^l(b_{Ridge})}{\partial b} \right|_{\hat{b}_{Ridge}} = \left. \frac{\partial \mathbf{l}^l(b)}{\partial b} \right|_{b_0} - (b_{Ridge} - b_0)' (X' V(b_0) X + 2dI) + o(\|b_{Ridge} - b_0\|)$$

şeklinde elde edilir. Burada hata terimi atılırsa ve (4.22) dikkate alındığında

$$\left. \frac{\partial \mathbf{l}'(b)}{\partial b} \right|_{b_0} = U'(b_0) = U(b_0) - 2db_0 \text{ olmak üzere,}$$

$$\left. \frac{\partial \mathbf{l}'(b_{Ridge})}{\partial b} \right|_{\hat{b}_{Ridge}} = [U(b_0) - 2db_0] - (b_{Ridge} - b_0)' [X'V(b_0)X + 2dI] \quad (4.25)$$

elde edilir. Burada $U(b_0) = \sum_{i=1}^n x_i'(y_i - \frac{e^{x_i'b_0}}{1 + e^{x_i'b_0}})$ dır. $\left. \frac{\partial \mathbf{l}'(b_{Ridge})}{\partial b} \right|_{\hat{b}_{Ridge}} = 0$ alınarak,

$\hat{b}_{Ridge}$ için birinci-sıra yaklaşımı yapılır. Eşitlik (4.25) sıfıra eşitlendikten sonra,

$$0 = [U(b_0) - 2db_0] - (\hat{b}_{Ridge} - b_0)' [X'V(b_0)X + 2dI]$$

elde edilir. Sağdan $[X'V(b_0)X + 2dI]^{-1}$ ile çarpıldıktan sonra eşitliğin her iki tarafının transpozu alınır

$$\hat{b}_{Ridge} = b_0 + [X'V(b_0)X + 2dI]^{-1} [U(b_0) - 2db_0]$$

elde edilir. Düzenlemelerden sonra

$$\hat{b}_{Ridge} = (X'V(b_0)X + 2dI)^{-1} (U(b_0) + X'V(b_0)Xb_0) \quad (4.26)$$

olarak $\hat{b}_{Ridge}$ için birinci-sıra tahmin edicisi elde edilmiş olunur. Eşitlik (4.26)'te $d = 0$ alınarak parantez dağıtılırsa,

$$\begin{aligned} \hat{b}_{MLE} &= [X'V(b_0)X]^{-1} [U(b_0) + X'V(b_0)Xb_0] \\ &= (X'V(b_0)X)^{-1} U(b_0) + b_0 \end{aligned}$$

elde edilir. b_0 eşitliğin karşı tarafına geçirilirse,

$$\hat{b}_{MLE} - b_0 = (X'V(b_0)X)^{-1}U(b_0)$$

olur ve $X'V(b_0)X$ ifadesi ile eşitliğin her iki tarafı çarpılırsa,

$$U(b_0) = X'V(b_0)X(\hat{b}_{MLE} - b_0) \quad (4.27)$$

elde edilir. Eşitlik (4.27) ve (4.26) birlikte ele alınırsa ve $2d = k$ olmak üzere,

$$\begin{aligned} \hat{b}_{Ridge} &= [X'V(b_0)X + kI]^{-1} [X'V(b_0)X(\hat{b}_{MLE} - b_0) + X'V(b_0)Xb_0] \\ &= [X'V(b_0)X + kI]^{-1} X'V(b_0)X\hat{b}_{MLE} \end{aligned} \quad (4.28)$$

şeklinde “ridge lojistik tahmin edici” elde edilmiş olunur. Buradan, $k \rightarrow 0$ ise $\hat{b}_{Ridge} \rightarrow \hat{b}_{MLE}$ ve $k \rightarrow \infty$ ise $\hat{b}_{Ridge} \rightarrow 0$ olduğu görülebilir. Bu ridge tahmin edicide $X'V(b_0)X$ yerine, $X'V(b_0)X$ ’nin tahmini olan $X'VX$ kullanılırsa, Schaefer ve ark. (1984)’nın sunmuş olduğu eşitlik (4.20) de verilen ridge lojistik tahmin edici elde edilir. Ancak, unutulmamalıdır ki, eşitlik (4.8)’de yer alan

$$\sum_{i=1}^n y_i \ln \left(\frac{p_i}{1-p_i} \right) + \sum_{i=1}^n \ln(1-p_i)$$

durumlarda Schaefer (1984)’in sunduğu tahmin edici tanımsız olmaktadır.

Eşitlik (4.14), $(X'VX)^{-1}X'V$ ortak parantezine alınarak yeniden düzenlenip beklenen değerine bakılırsa,

$$E(\hat{b}_{MLE}) = (X'VX)^{-1}X'V[Xb_0 + V^{-1}E(Y - P)]$$

yazılabilir. (4.7) den dolayı $E(Y - P) = 0$ olur yani $E(\hat{b}_{MLE}) = b_0$ dır. $\hat{b}_{MLE}$, b_0 için yansız tahmin edici olmak üzere, ridge lojistik tahmin edicinin beklenen değerine bakılacak olursa, (4.28) den

$$E(\hat{b}_{Ridge}) = [X'V(b_0)X + kI]^{-1} X'V(b_0)XE(\hat{b}_{MLE})$$

elde edilir. Ridge lojistik tahmin edicinin yanlılığı ise

$$\begin{aligned} \text{Bias}(\hat{b}_{Ridge}) &= E(\hat{b}_{Ridge} - b_0) \\ &= [X'V(b_0)X + kI]^{-1} [X'V(b_0)X - (X'V(b_0)X + kI)]b_0 \\ &= -k[X'V(b_0)X + kI]^{-1} b_0 \end{aligned} \quad (4.29)$$

olarak bulunur. Varyansına bakılacak olursa,

$$\text{var}(\hat{b}_{Ridge}) = \text{var}\left\{[X'V(b_0)X + kI]^{-1} X'V(b_0)X\hat{b}_{MLE}\right\}$$

$\text{var}(Ax) = A \text{var}(x) A'$ kuralı yardımıyla tekrar düzenlenirse,

$$\text{var}(\hat{b}_{Ridge}) = [X'V(b_0)X + kI]^{-1} (X'V(b_0)X) \text{var}(\hat{b}_{MLE}) (X'V(b_0)X) [X'V(b_0)X + kI]^{-1}$$

elde edilir. $\text{var}(\hat{b}_{MLE})$ nin yazılabilmesi için, (4.14) ile MLE lojistik tahmin edici $\hat{b}_{MLE} = (X'V(b_0)X)^{-1} X'V(b_0)[Xb_0 + V^{-1}(b_0)(Y - P)]$ ile verilmektedir. Köşeli parantez içerisindeki ifadelerden sadece Y nin bir rasgele değişken olduğu ve $\text{var}(Y) = V(b_0)$ olduğu dikkate alınır bu tahmin edicinin varyansı,

$$\begin{aligned} \text{var}(\hat{b}_{MLE}) &= (X'V(b_0)X)^{-1} X'V(b_0) \text{var}[Xb_0 + V^{-1}(b_0)(Y - P)] V(b_0) X' (X'V(b_0)X)^{-1} \\ &= (X'V(b_0)X)^{-1} \end{aligned}$$

olarak elde edilir. Böylelikle, $\text{var}(\hat{b}_{MLE}) = (X'V(b_0)X)^{-1}$ olduğundan, ridge lojistik tahmin edicinin varyansı,

$$\text{var}(\hat{b}_{Ridge}) = [X'V(b_0)X + kI]^{-1} (X'V(b_0)X) [X'V(b_0)X + kI]^{-1} \quad (4.30)$$

şeklinde elde edilir. Varyans ve yanlılık ifadeleri kullanılarak,

$$MSE(\hat{b}_{Ridge}) = \text{var}(\hat{b}_{Ridge}) + [bias(\hat{b}_{Ridge})][bias(\hat{b}_{Ridge})']$$

hata kareler ortalaması yazılabilir. Eşitlik (4.29) ve (4.30), $MSE(\hat{b}_{Ridge})$ 'de yerine yazılırsa,

$$\begin{aligned} MSE(\hat{b}_{Ridge}) &= (X'V(b_0)X + kI)^{-1} X'V(b_0)X (X'V(b_0)X + kI)^{-1} \\ &\quad + 4k^2 (X'V(b_0)X + kI)^{-1} b_0 b_0' (X'V(b_0)X + kI)^{-1} \end{aligned} \quad (4.31)$$

elde edilir.

4.3.1.1. Ridge Lojistik Tahmin Edicide k Parametresinin Seçimi

Ridge parametresi k 'nın seçimi için önerilen yöntemler çoklu lineer regresyonda önerilen yöntemlere benzerdir. Aşağıda Kural 1 ve 2 ile verilen yöntemler sırasıyla Hoerl ve Kennard (1970a,b) ve Hoerl, Kennard ve Baldwin (1975)'e benzer şekilde önerilmiştir.

Kural 1: $k = \frac{1}{\max_j (\hat{g}_j' \hat{b}_{MLE})^2}$, ($\hat{g}_j$: $(X'VX)$ in özvektörüdür) (Schaefer, Roi ve

Wolfe, 1984)

$$\text{Kural 2: } k = \frac{(p+1)}{\hat{b}_{MLE}' \hat{b}_{MLE}}, \text{ (Schaefer, Roi ve Wolfe, 1984)}$$

$$\text{Kural 3: } k = \frac{1}{\hat{b}_{MLE}' \hat{b}_{MLE}}, \text{ (Schaefer, Roi ve Wolfe, 1984)}$$

$$\text{Kural 4: } \hat{k} = \frac{[\text{trace}(X' \hat{V} X)]}{\hat{b}_{MLE}' (X' \hat{V} X) \hat{b}_{MLE}}, \text{ (Lee ve Silvapulle, 1988)}$$

Bu yöntemlerin dışında, ön tahmin amaçlı kullanılan yöntem de bulunmaktadır (Le Cessie ve Van Houwelingen, 1992). Ön tahmin amaçlı kullanılan yöntemler ise modelin ön tahmin hatasının (prediction error) tahminini minimum yapmaya dayalıdır. Öntahmin hatasını değerlendirmek için üç kriter dikkate alınır. x_{new} ve y_{new} yeni bir gözleme ait değerler olmak üzere, bu kriterler;

a) Sınıflandırma hatası

$$CE = \begin{cases} 1, & y_{new} = 1 \text{ ve } \hat{p}(x_{new}) < 0.5 \text{ ya da } y_{new} = 0 \text{ ve } \hat{p}(x_{new}) > 0.5 \text{ ise} \\ 0.5, & \hat{p}(x_{new}) = 0.5 \text{ ise} \\ 0, & \text{diger} \end{cases}$$

b) Hata karesi

$$SE = (y_{new} - \hat{p}(x_{new}))^2$$

c) Eksi log-likelihood hatası

$$ML = -[y_{new} \ln(\hat{p}(x_{new})) + (1 - y_{new}) \ln(1 - \hat{p}(x_{new}))]$$

şeklinde verilmektedir (Le Cessie ve Van Houwelingen, 1992). $\hat{b}_{(-i)}$, (x_i, y_i) gözlemi hariç diğer bütün gözlemlere dayalı tahmin; $\hat{p}_{(-i)}(x)$, $\hat{b}_{(-i)}$ 'ye dayalı $p(x)$ 'in tahminini gösterebilir. Böylece ortalama ön tahmin hatası için bir tahmin

bulmak mümkün olur ve ortalama ön tahmin hatası minimum yapılarak ridge parametresi bulunur. CE, SE ve ML ile verilen ön tahmin hatalarının ortalamasının çapraz geçerlilik tahminleri aşağıdaki gibidir:

a) Ortalama sınıflandırma hatası

$$MCE_{CV} = n^{-1} \sum_i \left\{ y_i \left[\hat{p}_{(-i)}(x_i) < \frac{1}{2} \right] + (1 - y_i) \left[\hat{p}_{(-i)}(x_i) > \frac{1}{2} \right] + \frac{1}{2} \left[\hat{p}_{(-i)}(x_i) = \frac{1}{2} \right] \right\}$$

Burada, $[]$ indikatör fonksiyonunu göstermektedir. Parantez içi doğruysa 1, yanlışsa 0'dır.

b) Ortalama hata kareler

$$MSE_{CV} = n^{-1} \sum_i (y_i - \hat{p}_{(-i)}(x_i))^2$$

c) Ortalama eksi log-likelihood

$$MML_{CV} = -n^{-1} \sum_i [y_i \log \hat{p}_{(-i)}(x_i) + (1 - y_i) \log(1 - \hat{p}_{(-i)}(x_i))]$$

Çapraz geçerlilik, her bir gözlem dışarı atılıp tekrar tekrar parametre tahmini yapıldığı için zaman alıcı bir yöntemdir.

Cook ve Weisberg (1982), kısıtlandırılmamış MLE'leri göz önünde bulundurarak $\hat{b}_{Ridge(-i)}$ için yaklaşımlar yapmışlardır.

Tüm gözlemleri içeren ridge lojistik tahmin edici (4.28) de verilmişti. i-inci gözlem çıkarılarak elde edilen ridge lojistik tahmin edici ise

$$X'_{(-i)} V_{(-i)}(b_0) X_{(-i)} = X' V(b_0) X - X'_i v_{ii} X_i, \quad (4.32)$$

$$\hat{b}_{MLE(-i)} = \hat{b}_{MLE} - \frac{(X' V(b_0) X)^{-1} X'_i [y_i - \hat{p}(x_i)]}{1 - t_{ii}}, \quad (4.33)$$

ve

$$\left[X'_{(-i)} V_{(-i)}(b_0) X_{(-i)} + kI \right]^{-1} = \left[X' V(b_0) X + kI - X'_i v_{ii} X_i \right]^{-1} \quad (4.34)$$

olmak üzere,

$$\hat{b}_{Ridge(-i)} = \left[X'_{(-i)} V_{(-i)}(b_0) X_{(-i)} + kI \right]^{-1} X'_{(-i)} V_{(-i)}(b_0) X_{(-i)} \hat{b}_{MLE(-i)} \quad (4.35)$$

şeklinde ifade edilir. Burada $t_{ii} = v_{ii} X_i (X' V(b_0) X)^{-1} X'_i$ şeklindedir.

$$\left[B - CC' \right]^{-1} = B^{-1} + \frac{B^{-1} CC' B^{-1}}{1 - C' B^{-1} C} \quad \text{olarak tanımlanan Householder dönüşümü}$$

kullanılarak $B = X' V(b_0) X + kI$, $C = X'_i \sqrt{v_{ii}}$ ve $CC' = X'_i v_{ii} X_i$ olmak üzere, (4.34) ile verilen eşitlik yeniden düzenlenirse,

$$\begin{aligned} \left[X'_{(-i)} V_{(-i)}(b_0) X_{(-i)} + kI \right]^{-1} &= \left[X' V(b_0) X + kI \right]^{-1} \\ &+ \frac{\left[X' V(b_0) X + kI \right]^{-1} X'_i v_{ii} X_i \left[X' V(b_0) X + kI \right]^{-1}}{1 - v_{ii} X_i \left[X' V(b_0) X + kI \right]^{-1} X_i} \end{aligned} \quad (4.36)$$

elde edilir. $h_{ii} = v_{ii} X_i \left[X' V(b_0) X + kI \right]^{-1} X'_i$ olmak üzere, eşitlik (4.32), (4.33) ve (4.36) birlikte ele alınarak, eşitlik (4.35) tekrar düzenlenirse,

$$\begin{aligned} \hat{b}_{Ridge(-i)} &= \left\{ \left[X' V(b_0) X + kI \right]^{-1} + \frac{\left[X' V(b_0) X + kI \right]^{-1} X'_i v_{ii} X_i \left[X' V(b_0) X + kI \right]^{-1}}{1 - h_{ii}} \right\} \\ &\times \left\{ X' V(b_0) X - X'_i v_{ii} X_i \right\} \hat{b}_{MLE} - \frac{(X' V(b_0) X)^{-1} X'_i [y_i - \hat{p}(x_i)]}{1 - t_{ii}} \end{aligned}$$

elde edilir. $X'V(b_0)X - X_i'v_{ii}X_i$ ifadesi parantez içine dağıtılırsa,

$$\begin{aligned}\hat{b}_{Ridge(-i)} = & \{[X'V(b_0)X + kI]^{-1} X'V(b_0)X \\ & + \frac{[X'V(b_0)X + kI]^{-1} X_i'v_{ii}X_i [X'V(b_0)X + kI]^{-1}}{1 - h_{ii}} X'V(b_0)X \\ & - [X'V(b_0)X + kI]^{-1} X_i'v_{ii}X_i \\ & - \frac{[X'V(b_0)X + kI]^{-1} X_i'v_{ii}X_i [X'V(b_0)X + kI]^{-1}}{1 - h_{ii}} X_i'v_{ii}X_i \} \\ & \times \left\{ \hat{b}_{MLE} - \frac{(X'V(b_0)X)^{-1} X_i'[y_i - \hat{p}(x_i)]}{1 - t_{ii}} \right\}\end{aligned}$$

ve sadeleştirmelerden sonra

$$\begin{aligned}\hat{b}_{Ridge(-i)} = & \{[X'V(b_0)X + kI]^{-1} X'V(b_0)X \\ & + \frac{[X'V(b_0)X + kI]^{-1} X_i'v_{ii}X_i [X'V(b_0)X + kI]^{-1}}{1 - h_{ii}} X'V(b_0)X \\ & - \frac{[X'V(b_0)X + kI]^{-1} X_i'v_{ii}X_i}{1 - h_{ii}} \} \times \left\{ \hat{b}_{MLE} - \frac{(X'V(b_0)X)^{-1} X_i'[y_i - \hat{p}(x_i)]}{1 - t_{ii}} \right\}\end{aligned}\quad (4.37)$$

olur. Eşitlik (4.37) tekrar düzenlenirse,

$$\hat{b}_{Ridge(-i)} = \hat{b}_{Ridge} - \frac{[X'V(b_0)X + kI]^{-1} X_i'[y_i - \hat{p}(x_i)]}{1 - h_{ii}}\quad (4.38)$$

olarak elde edilir. Burada dikkat edilmesi gereken nokta, eşitlik (4.33)'teki $\hat{p}(x_i)$ MLE kullanılarak elde edilen tahmin edici iken eşitlik (4.38)'dekinin ridge kullanılarak elde edilen tahmin edici olmasıdır. Bu tahmin edici ile birlikte bu zaman alıcı prosedürden kurtulmuş olunur. Böylece MSE_{CV} ye

$$MSE_{ACV} = n^{-1} \sum_i \frac{[y_i - \hat{p}(x_i)]^2}{(1 - h_{ii})^2} \quad (4.39)$$

eşitliği ile yaklaşılır.

4.3.2. Temel Bileşenler Lojistik Tahmin Edici

Lojistik regresyonda çoklu iç ilişki bulunması durumunda ridge tahmin ediciye alternatif olarak temel bileşenler tahmin edici de kullanılmaktadır.

Schaefer (1986), maksimum likelihood tahmin ediciye tek basamaklı, yani sadece en küçük öz değerlere sahip gözlemlerin tek seferde atılmasına dayalı bir düzeltme olacak şekilde lojistik regresyon içetin bir temel bileşen tahmin edici sunmuştur. Tek seferde uygulanan bu tahmin edici, daha sonraları Marx ve Smith (1990)'in sunacağı iteratif yöntemle hemen hemen aynı sonuçları vermektedir.

$x_j^{(**)}$, lojistik regresyona ait bilgi matrisinin ($f_{Lojistik} = T^{(L)} \Lambda^{(**)} T^{(L)'}$) özdeğerleri, $g_j^{(**)}$ ise bu özdeğerlere ilişkin özvektörleri ve $f_{Lojistik} = X' V X = T^{(L)} \Lambda^{(**)} T^{(L)'} = \sum_{j=0}^p x_j^{(**)} g_j^{(**)} g_j^{(**)'}$ bilgi matrisini gösterebilir. Schaefer (1986) lojistik regresyon programlarının MLE tahmini için genellikle başlangıç noktasını EKK tahmin edici olarak aldığını belirtmiştir. L , yakınsamanın sağlandığı iterasyon ve $\hat{b}_0$, b 'nın EKK tahmin edicisi olmak üzere,

$$\begin{aligned} \hat{b} &= \hat{b}_0 + \sum_{l=0}^L (X' \hat{V}_l X)^{-1} X' (y - p_l) \\ &= (X' X)^{-1} X' y + \sum_{l=0}^L (X' \hat{V}_l X)^{-1} X' (y - p_l) \end{aligned} \quad (4.40)$$

olarak verilmektedir. Schaefer (1986) tek bir iç ilişkinin var olduğunu varsayarak eşitlik (4.40)'ı,

$$\hat{b}_{PC} = (X'X)^+ X'y + \sum_{l=0}^L (X'\hat{V}_l X)^+ X'(y - p_l) \quad (4.41)$$

şeklinde yazmıştır. Burada, $(X'X)^+ = \sum_{j=2}^p l_j^{*-1} g_j g_j'$ ve $(X'\hat{V}_l X)^+ = \sum_{j=2}^p x_j^{(**)-1} g_j^{(**)} g_j^{(**)'}'$

'dir. Ancak (4.41) ile verilen tahmin ediciyi elde etmek, $(X'\hat{V}_l X)^+$ ifadesinin her iterasyonda güncellenmesi gerektiğinden zordur.

Lojistik fonksiyonun yapısı gereği sabit bir x değeri için, $\hat{b}$ 'daki küçük değişiklikler p_l 'de ve dolayısıyla $\hat{V}$ 'de çok daha küçük değişikliğe yol açar. Böylelikle, $(X'\hat{V}X) \approx (X'\hat{V}_l X)$ ve $(X'\hat{V}X)^+ \approx (X'\hat{V}_l X)^+$ olur.

$(X'VX)_s^+ = \sum_{j=0}^{s-1} x_j^{(**)-1} g_j^{(**)} g_j^{(**)'}'$ tanımlayarak, Schaefer (1986) tek basamaklı temel bileşenler lojistik tahmin ediciyi,

$$\hat{b}_s^{pc} = (X'VX)_s^+ X'VX\hat{b} \quad (4.42)$$

şeklinde ifade etmiştir.

Eşitlik (4.42) ile verilen Schaefer (1986)'ın sunduğu tahmin ediciye alternatif olarak, eşitlik (2.34) ile verilen Smith ve Marx (1990)'ın genelleştirilmiş lineer modeller için sunduğu iteratif temel bileşen tahmin edici,

$$\frac{\partial h_i}{\partial m_i} = \frac{1}{p_i(1-p_i)},$$

$e_{i,s} = y_{i,s} - p_{i,s}$, $k_{ii}^{-1} = v_{ii} = \text{diag}(V) = p_i(1-p_i)$ olmak üzere,

$$b_{t,s}^{pc} = b_{t-1,s}^{pc} + \sum_{j=0}^{s-1} x_j^{(**)-1} g_j^{(**)} g_j^{(**)'}' \left[\sum_{i=1}^n x_i (y_{i,s} - p_{i,s}) \right]_{t-1} \quad (4.43)$$

lojistik regresyon için güncellenebilir. $x_s^{(**)}, x_{s+1}^{(**)}, \dots, x_p^{(**)}$, $r = p + 1 - s$ tane çok küçük özdeğer ve $Z^{(**)} = XT^{(L)}$ olmak üzere, eşitlik (3.16), a_s^{pc} 'ye göre tekrar düzenlenirse;

$$\begin{aligned} a_{t,s}^{pc} &= a_{t-1,s}^{pc} + \Lambda_s^{(**)-1} \left[\sum_{i=1}^n z_{i,s}^{(**)} (y_{i,s} - p_{i,s}) \right]_{t-1} \\ &= \Lambda_s^{(**)-1} Z_s^{(**)} V y_{t-1}^{***pc} \end{aligned}$$

elde edilir. Burada $z_{i,s}^{(**)}$, $Z_s^{(**)}$ matrisinin sütunlarıdır ve

$$y_{t-1}^{***pc} = \ln\left(\frac{p_{i,s}}{1-p_{i,s}}\right) + (y_{i,s} - p_{i,s}) \frac{1}{p_{i,s}(1-p_{i,s})} \text{ dir ve } (t-1) \text{ iterasyonunda değeri}$$

verilir. Doğal bir sonuç olarak, $\hat{a}_s^{pc}$ 'nin yakınsamasına dayalı “temel bileşenler lojistik tahmin edici”

$$\hat{b}_s^{pc} = T_s^{(L)} a_s^{pc} \quad (4.44)$$

olur.

Temel bileşenler lojistik tahmin edicinin varyansına bakılacak olursa, $T^{(L)} = \begin{bmatrix} T_s^{(L)} & T_r^{(L)} \end{bmatrix}$, bilgi matrisinin özvektörünün tam kümesi olmak üzere

$$\text{var}(\hat{b}_s^{pc}) = \text{var}(\hat{b}_{MLE}) - T_r^{(L)} \Lambda_r^{(**)-1} T_r^{(L)'} \quad (4.45)$$

olur. $f_{(Lojistik)r}^{-1} = T_r^{(L)} \Lambda_r^{(**)-1} T_r^{(L)'}$ ifadesi, tam modelden atılan gözlemlerin sebep

olduğu varyansı göstermektedir. $\text{var}(\hat{b}_{MLE}) = (X'VX)^{-1}$ ve

$(X'VX)^{-1} = T^{(L)} \Lambda^{(**)-1} T^{(L)'} = T_s^{(L)} \Lambda_s^{(**)-1} T_s^{(L)'} + T_r^{(L)} \Lambda_r^{(**)-1} T_r^{(L)'}$ olduğundan (4.45) tekrar düzenlenirse,

$$\text{var}(\hat{b}_s^{PC}) = T_s^{(L)} \Lambda_s^{(**)-1} T_s^{(L)'} \quad (4.46)$$

olarak elde edilir. $\hat{b}_s^{PC}$ 'ye ilişkin yanlışlık ise $\text{bias}(\hat{b}_s^{PC}) = E(\hat{b}_s^{PC}) - b = -T_r^{(L)} a_r$ şeklinde verilmektedir. $\hat{b}_s^{PC}$ için MSE ise;

$$\text{MSE}(\hat{b}_s^{PC}) = T_s^{(L)} \Lambda_s^{(**)-1} T_s^{(L)'} + \left(-T_r^{(L)} a_r \right) \left(-T_r^{(L)} a_r \right)'$$

şeklinde matris formunda yazılabilir.

4.3.3. Stein Lojistik Tahmin Edici

Stein (1960) tarafından lineer regresyon modelleri için önerilen tahmin ediciyi Schaefer (1986) lojistik regresyona uyarlamıştır. Stein lojistik tahmin edici,

$$\hat{b}_{Stein} = c \hat{b}_{MLE} \quad (4.47)$$

şeklinde verilmektedir. Burada, $0 < c < 1$ 'dir. c 'nin seçimi için aşağıdaki iki kural kullanılmaktadır:

$$\text{Kural 1: } c = \frac{\hat{b}_{MLE}' \hat{b}_{MLE}}{\hat{b}_{MLE}' \hat{b}_{MLE} + \text{trace}[(X' V X)^{-1}]} \quad (\text{Stein, 1960})$$

$$\text{Kural 2: } c = \frac{\sum_{i=1}^{p+1} a_i^2 x_i^{(**)}}{(p+1) + \sum_{i=1}^{p+1} a_i^2 x_i^{(**)}}, \quad a = T^{(L)'} b \quad (\text{Marx, 1988})$$

4.4. Lojistik Regresyonda İstatistiksel Çıkarımlar

4.4.1. Sapma (Deviance)

Genelleştirilmiş lineer modeller için Bölüm 2.4.1’de anlatılanların paralelinde lojistik regresyon modeli için sapmanın (deviance) özel hali yazılacak olursa,

$$\begin{aligned}
 D = \text{Deviance} &= -2 \ln \left[\frac{L(b)}{L(P)} \right] = -2 [L(\hat{m}; y) - L(y; y)] \\
 &= -2 \ln \left[\frac{L(\text{Uydurulmuş Model})}{L(\text{Doymuş Model})} \right] = -2 \ln \left[\frac{\prod_{i=1}^n \hat{p}_i^{y_i} (1 - \hat{p}_i)^{1-y_i}}{\prod_{i=1}^n y_i^{y_i} (1 - y_i)^{1-y_i}} \right] \\
 &= -2 \ln \left[\prod_{i=1}^n \left(\frac{\hat{p}_i}{y_i} \right)^{y_i} \cdot \left(\frac{1 - \hat{p}_i}{1 - y_i} \right)^{1-y_i} \right] \\
 &= -2 \sum_{i=1}^n \left(y_i \ln \left(\frac{\hat{p}_i}{y_i} \right) + (1 - y_i) \ln \left(\frac{1 - \hat{p}_i}{1 - y_i} \right) \right)
 \end{aligned}$$

şeklinde olur. Karar kuralı ise Bölüm 2.4.1’de bahsedildiği gibi olacaktır.

4.4.2. Uyumun İyiliği

Uyumun iyiliği, kurulan modelin yanıt değişkeni ne kadar iyi açıklayabildiğinin ifadesidir. Bu da uydurulmuş y değerleri ile orijinal y değerleri arasında ki farkla ilgilidir. Lojistik regresyon modelinde uyumun iyiliği ölçütlerine geçmeden önce “covariate pattern” kavramının incelenmesi gerekir.

Öncelikle, lojistik regresyonda, covariate, sürekli tipteki açıklayıcı değişkenlere denir. Kesikli tipteki açıklayıcı değişkenlere ise faktör değişken adı verilir. Covariate pattern ise, modeldeki covariate değişkenlerin tekli kümelerine denir (Hosmer ve Lemeshow, 2000).

p : açıklayıcı değişken sayısı

n : örneklem hacmi

$x' = (x_1, x_2, \dots, x_p)$

m_j : x 'in belli bir değeri aldığı durumdaki x 'lerin gözlem sayısı

J : x_j nin farklı (distinct) değerleri

$$\sum m_j = n \text{ dir.}$$

4.4.2.1. Pearson Ki-Kare İstatistiği ve Sapma (Deviance)

$\hat{g}(x_j)$: tahmin edilmiş logit olmak üzere, j – inci covariate pattern için uydurulmuş yanıt

$$\hat{y}_j = m_j \hat{p}_j = m_j \frac{e^{\hat{g}(x_j)}}{1 + e^{\hat{g}(x_j)}} \quad (4.48)$$

şeklinde olur. Lojistik regresyon için, Bölüm 2.4.4.2'den i -inci gözleme ait Pearson rezidü ise,

$$e_i = \frac{y_i - p_i}{\sqrt{p_i(1-p_i)}} \quad (4.49)$$

olarak ifade edilebilir. Her bir tek covariate pattern için Pearson rezidü ise

$$r(y_j, \hat{p}_j) = \frac{(y_j - m_j \hat{p}_j)}{\sqrt{m_j \hat{p}_j (1 - \hat{p}_j)}} \quad (4.50)$$

şeklindedir. Buna bağlı olarak geliştirilen Pearson Ki-Kare istatistiği ise:

$$X^2 = \sum_{j=1}^J r(y_j, \hat{p}_j)^2 \quad (4.51)$$

olarak verilir. Sapma (deviance) rezidü ise

$$d(y_j, \hat{p}_j) = \pm \left\{ 2 \left[y_j \ln \left(\frac{y_j}{m_j \hat{p}_j} \right) + (m_j - y_j) \ln \left(\frac{(m_j - y_j)}{m_j (1 - \hat{p}_j)} \right) \right] \right\}^{1/2} \quad (4.52)$$

olarak tanımlanır. Burada işaret, $(y_j - m_j \hat{p}_j)$ 'nin işareti ile aynıdır.

$y_j = 0$ 'lı covariate pattern ler için deviance rezidü;

$$d(y_j, \hat{p}_j) = -\sqrt{2m_j |\ln(1 - \hat{p}_j)|} \quad (4.53)$$

ve $y_j = m_j$ olduğu zaman deviance rezidü ise;

$$d(y_j, \hat{p}_j) = -\sqrt{2m_j |\ln(\hat{p}_j)|} \quad (4.54)$$

olur.

Deviance rezidüye ait genel form eşitlik (2.41) ile verilmektedir. Genel formu oluşturan terimler, lojistik regresyon için, $w_i = 1$, $h_i = \ln\left(\frac{y_i}{1 - y_i}\right)$, $\hat{h}_i = \ln\left(\frac{p_i}{1 - p_i}\right)$, $b(h_i) = -\ln(1 - y_i)$ ve $b(\hat{h}_i) = -\ln(1 - p_i)$ 'dir. Bu ifadeler kullanılarak, i -inci gözleme ait deviance rezidü ise,

$$\text{sign}(y_i - p_i) \sqrt{2 \left[y_i \ln\left(\frac{y_i}{p_i}\right) + (1 - y_i) \ln\left(\frac{1 - y_i}{1 - p_i}\right) \right]} \quad (4.55)$$

olarak elde edilir. Hosmer ve Lemeshow'da (2000) eşitlik (4.52) ile verilen deviance rezidüye denk bir eşitlik elde edilir. Deviance rezidülere dayalı oluşturulan deviance istatistiği,

$$D = \sum_{j=1}^J d(y_j, \hat{p}_j)^2 \quad (4.56)$$

olarak verilir. Deviance (D) ve Pearson Ki-kare istatistikleri (X^2) $J - (p + 1)$ serbestlik dereceli ki-kare dağılımına sahiptir.

4.4.2.2. Hosmer-Lemeshow Testi

Hosmer-Lemeshow uyumun iyiliği testi

$$\hat{C} = \sum_{k=1}^g \frac{(o_k - n'_k \bar{p}_k)^2}{n'_k \bar{p}_k (1 - \hat{p}_k)} \quad (4.57)$$

ile verilir. Burada, g yüzdeler dilim (decile) sayısını (aksi söylenmediği sürece paket programlar dahil $g=10$ olarak alınmaktadır). Örneğin, $g = 10$ ise $n'_1 = \frac{n}{10}$ olacaktır.

Bu bize, en düşük olasılığa sahip $\frac{n}{10}$ tane gözlem 1. gruba alınır.

n'_k , k -ıncı gruptaki gözlemlerin sayısını; c_k , k -ıncı dilimdeki (decile)

covariate pattern'lerin sayısını göstermek üzere $o_k = \sum_{j=1}^{c_k} y_j$, c_k covariate

pattern'lerin arasındaki yanıtların sayısını ve $\bar{p}_k = \sum_{j=1}^{c_k} \frac{m_j \hat{p}_j}{n'_k}$, ortalama tahmin

edilmiş olasılığı göstermektedir. Hosmer-Lemeshow uyumun iyiliği testi, $\hat{C}$, $g - 2$ serbestlik dereceli ki-kare dağılımına sahiptir.

4.4.3. Modeldeki Katsayılar İçin Önemlilik Testi

Modeldeki katsayıların önemliliğinin test edilmesi için kurulacak hipotez testi, test istatistikleri ve karar kuralı lojistik regresyon için Bölüm 2.4.2’de anlatıldığı gibidir.

4.4.4. Model Karşılaştırma

Genelleştirilmiş lineer modeller için Bölüm 2.4.3’de anlatılan model karşılaştırma yöntemi lojistik regresyon için de geçerlidir.

4.4.5. Lojistik Regresyonda Tanılama (Diagnostics)

Lineer regresyonda, şapka (hat) matrisi $H = X(X'X)^{-1}X'$ olarak verilmektedir. Benzer şekilde, ağırlıklandırılmış en küçük kareler yöntemini kullanarak, Pregibon (1981) uydurulmuş değerler için bir lineer yaklaşım türetti ki bu lojistik regresyon için $H = V^{1/2}X(X'VX)^{-1}X'V^{1/2}$ olarak tanımlanan şapka matrisini verir. Burada, V matrisinin elemanları $v_j = m_j\hat{p}(x_j)[1-\hat{p}(x_j)]$ olan $J \times J$ ’lik köşegen bir matristir. Lineer regresyonda, şapka matrisin köşegen elemanları “leverage” yani etkin değerlerdir. Bu düşünce lojistik regresyona uyarlandığında $b_j = x_j'(X'VX)^{-1}x_j$ ve $x_j' = (1, x_{1j}, x_{2j}, \dots, x_{pj})$ olmak üzere,

$$h_j = m_j\hat{p}(x_j)[1-\hat{p}(x_j)]x_j'(X'VX)^{-1}x_j = v_j \times b_j \quad (4.58)$$

olur ve H – şapka matrisinin j – inci köşegen elemanınıdır. Burada önemli olan, v_j nin h_j üzerindeki etkisini, leverage’ın büyüklüğünü yorumlamada kullanılmasıdır.

x_j covariate pattern için standartlaştırılmış Pearson rezidü;

$$r_{sj} = \frac{r_j}{\sqrt{1-h_j}} \quad (4.59)$$

ile verilir. Diğer bir faydalı tanılama ise, lineer regresyondaki Cook değerleri benzeri bir yaklaşımdır:

$$\begin{aligned}\Delta \hat{b}_j &= (\hat{b} - \hat{b}_{(-j)})' (X'VX) (\hat{b} - \hat{b}_{(-j)}) \\ &= \frac{r_j^2 h_j}{(1-h_j)^2} \\ &= \frac{r_{sj}^2 h_j}{(1-h_j)}\end{aligned}\tag{4.60}$$

Pearson Ki-kare istatistiğinin azalması ise şu şekilde yazabilir:

$$\Delta X_j^2 = \frac{r_j^2}{(1-h_j)} = r_{sj}^2.\tag{4.61}$$

Aynı şekilde, sapmadaki değişim ise;

$$\Delta D_j = d_j^2 + \frac{r_j^2 h_j}{(1-h_j)}\tag{4.62}$$

ile verilir. Eğer r_j^2 nin yerine d_j^2 yazılırsa,

$$\Delta D_j = \frac{d_j^2}{(1-h_j)}\tag{4.63}$$

olur. Bu tanılama istatistikleri, covariate pattern'lerin kötü uyup uymadığının tespit edilmesinde kullanışlıdır.

Aşağıdaki gibi çeşitli grafikler oluşturulup bu grafikler üzerinde kötü uyum sağlayan gözlemler belirlenebilir (Hosmer and Lemeshow, 2000):

- ΔX_j^2 ye karşın $\hat{p}_j$ grafiği

- ΔD_j ye karşın $\hat{p}_j$ grafiği
- $\Delta \hat{b}_j$ ye karşın $\hat{p}_j$ grafiği
- ΔX_j^2 ye karşın h_j grafiği
- ΔD_j ye karşın h_j grafiği
- $\Delta \hat{b}_j$ ye karşın h_j grafiği

5. ÖRNEK UYGULAMA

Maksimum likelihood tahmin edicinin kullanılması ile yapılan lojistik regresyon örnek uygulaması için veri seti TÜİK'ten elde edilmiştir. TÜİK tarafından tüketici fiyat indekslerinde kullanılacak maddelerin seçimi ve temel yıl ağırlıklarının elde edilmesi, hanelerin tüketim yapısı ve tüketim kalıplarında zaman içinde meydana gelecek değişikliklerin izlenmesi, ülkede yaratılan kullanılabilir gelirin, hanehalkı ve fertler arasındaki bölüşümünün ortaya konulması, milli gelir hesaplamalarında özel nihai tüketim harcamaları tahminlerine yardımcı olacak verilerin derlenmesi, asgari ücret tespit çalışmaları için ihtiyaç duyulan verilerin elde edilmesi ve yoksulluk sınırının belirlenmesi, hanehalklarının yaşam seviyeleri, beslenme sorunları vb. diğer sosyo-ekonomik analizler için gerekli verilerin elde edilmesi amacıyla gerçekleştirilen Hanehalkı Bütçe Anketi verileri analiz için kullanılmıştır. Veri seti 2006 yılı Türkiye geneli aylık olarak yapılan anket verilerini içermekte olup 8556 haneye ait gözlem bulunmaktadır.

Hanehalkı Bütçe Anketi verileri, FERT, HANE ve TÜKETİM tablolarından oluşmaktadır. FERT tablosu, hanede yaşayan fertlere ait eğitim durumu, cinsiyet, yaş ve gelir durumu gibi değişkenleri içermektedir. HANE tablosunda ise, yaşanan haneye ait kolaylıklar, hanede kullanılan ısıtma türü, yakıt türü ve sahip olunan gayrimenkuller gibi değişkenler sorgulanmaktadır. TÜKETİM tablosu ise, hanehalkına ait aylık COICOP 5'li kod sistemine göre ürün bazında tüketim harcaması değerlerini içermektedir.

Yapılan örnek çalışmada ise, HANE ve FERT tablolarından bazı değişkenler sezgisel olarak seçilmiştir. TÜKETİM tablosundan ise hane bazında COICOP 3. Ana harcama grubuna ait giyim ve ayakkabı harcama değerleri hesaplanmıştır. Yanıt değişkeni olarak, hanenin giyim ve ayakkabı harcaması olup olmadığı seçilmiştir. Yani,

$$y = \begin{cases} 1, & \text{giyim ve ayakkabı harcama değeri} > 0TL \\ 0, & \text{giyim ve ayakkabı harcama değeri} = 0TL \end{cases}$$

olarak belirlenmiştir. Açıklayıcı değişkenler olarak ise, konutta oturulan süre (OTUR_SUR), konutta ısınma amaçlı kullanılan birinci yakıt türü (YAKIT), hanedeki referans kişinin eğitim durumu (EGITIM) ve eşdeğer hanehalkı büyüklüğü (Hanedeki ilk yetişkin için 1, 14 ve daha yukarı yaştaki fertler için 0.5, 14 yaşından küçük fertler için 0.3 değerleri dikkate alınarak hesaplanan hanehalkı büyüklüğüdür) (EFB_OECD) alınmıştır. Açıklayıcı değişkenler,

$$x_1 = OTUR_SUR$$

$$x_2 = YAKIT$$

$$x_3 = EGITIM$$

$$x_4 = EFB_OECD$$

olarak seçilmiştir. Bu açıklayıcı değişkenlerden bazıları kategorik olduğundan dolayı kukla değişkenler kullanılmıştır. Tablo 5.1, YAKIT değişkeni için tanımlanan kukla değişkenin 9 seviyesini göstermektedir.

Tablo 5.1. YAKIT değişkeni için kukla değişken seviyeleri

YAKIT	AÇIKLAMA	1	2	3	4	5	6	7	8	9
YAKIT(1)	Odun	0	0	0	0	0	0	0	0	0
YAKIT(2)	Kömür	1	0	0	0	0	0	0	0	0
YAKIT(3)	Doğalgaz	0	1	0	0	0	0	0	0	0
YAKIT(4)	Fuel-Öil	0	0	1	0	0	0	0	0	0
YAKIT(5)	Motorin-Mazot	0	0	0	1	0	0	0	0	0
YAKIT(6)	Gazyağı	0	0	0	0	1	0	0	0	0
YAKIT(7)	Lpg	0	0	0	0	0	1	0	0	0
YAKIT(8)	Elektrik	0	0	0	0	0	0	1	0	0
YAKIT(9)	Tezek	0	0	0	0	0	0	0	1	0
YAKIT(10)	Diğer	0	0	0	0	0	0	0	0	1

Tablo 5.2. EGITIM değişkeni için kukla değişken seviyeleri

EGITIM	AÇIKLAMA	1	2	3	4	5	6	7	8	9	10
EGITIM(1)	Okur-Yazar Değil	0	0	0	0	0	0	0	0	0	0
EGITIM(2)	Okur-Yazar, Bir Okul Bitirmedi	1	0	0	0	0	0	0	0	0	0
EGITIM(3)	İlkokul	0	1	0	0	0	0	0	0	0	0
EGITIM(4)	İlköğretim	0	0	1	0	0	0	0	0	0	0
EGITIM(5)	Ortaokul	0	0	0	1	0	0	0	0	0	0
EGITIM(6)	Orta Dengi Meslek	0	0	0	0	1	0	0	0	0	0
EGITIM(7)	Lise	0	0	0	0	0	1	0	0	0	0
EGITIM(8)	Lise Dengi Meslek	0	0	0	0	0	0	1	0	0	0
EGITIM(9)	2 Yıllık Yüksekokul	0	0	0	0	0	0	0	1	0	0
EGITIM(10)	4 Yıllık Yüksekokul, Fakulte	0	0	0	0	0	0	0	0	1	0
EGITIM(11)	Yüksek Lisans, Doktora	0	0	0	0	0	0	0	0	0	1

Tablo 5.2 ise, EGITIM değişkeni için tanımlanan kukla değişkenin 10 seviyesini göstermektedir. Bu çerçevede X matrisi,

$$\begin{aligned}
X_1 &= \begin{bmatrix} x_{11} \\ x_{21} \\ M \\ M \\ x_{85561} \end{bmatrix}_{8556 \times 1}, \quad X_2 = \begin{bmatrix} x_{12}^A & x_{12}^B & x_{12}^C & L & L & L & L & L & x_{12}^I \\ x_{22}^A & x_{22}^B & x_{22}^C & L & L & L & L & L & x_{22}^I \\ M & M & M & M & M & M & M & M & M \\ M & M & M & M & M & M & M & M & M \\ x_{85562}^A & x_{85562}^B & x_{85562}^C & L & L & L & L & L & x_{85562}^I \end{bmatrix}_{8556 \times 9}, \\
X_3 &= \begin{bmatrix} x_{13}^A & x_{13}^B & x_{13}^C & L & L & L & L & L & x_{13}^J \\ x_{23}^A & x_{23}^B & x_{23}^C & L & L & L & L & L & x_{23}^J \\ M & M & M & M & M & M & M & M & M \\ M & M & M & M & M & M & M & M & M \\ x_{85563}^A & x_{85563}^B & x_{85563}^C & L & L & L & L & L & x_{85563}^J \end{bmatrix}_{8556 \times 10} \quad \text{ve} \\
X_4 &= \begin{bmatrix} x_{14} \\ x_{24} \\ M \\ M \\ x_{85564} \end{bmatrix}_{8556 \times 1} \quad \text{olmak üzere } X = [1 \quad X_1 \quad X_2 \quad X_3 \quad X_4]_{8556 \times 22}
\end{aligned}$$

şeklinde yazılır. Yanıt değişken 1 ve 0 gibi farklı iki değer aldığından dolayı model tahmini için lojistik regresyon kullanılması gerekmektedir. Maksimum likelihood tahmin edici kullanılarak model kurulmadan önce veri setinde çoklu iç ilişkinin bulunup bulunmadığı incelenmelidir. Bu bağlamda, veri setinden $f_{Lojistik} = X'VX$ matrisinin öz değerleri MATLAB 7.1 programı kullanılarak hesaplanmıştır. Bu özdeğerler sırasıyla 4.8234, 0.0445, 0.0037, 0.0017, 2 tane 0.0016, 0.0015, 0.0013, 0.0012, 0.0011, 0.001, 2 tane 0.0008, 0.0007, 0.0005, 0.0003, 0.0002, 2 tane 0.0001 ve 3 tane de sıfırdır. Özdeğerler incelendiği zaman sıfıra yakınsayan değerler görülmektedir. Bu da demek oluyor ki, veride çoklu iç ilişki bulunmaktadır. Ancak, lojistik regresyonda kukla değişken kullanmak, başlı başına çoklu iç ilişkiye sebep olabilecek bir konudur. Dolayısıyla bu çoklu iç ilişki göz ardı edilerek veri analizine devam edilecektir. Böylelikle maksimum likelihood tahmin edici kullanılabilir. SPSS 13.0 paket programının Regression menüsü kullanılarak veri seti analiz edilmiştir.

MLE lojistik tahmin edici, ilk olarak sabit katsayı ile iterasyona başlamaktadır. Tablo 5.3'den anlaşılabacağı üzere, ilk iterasyon adımına 1.207 olarak

başlamış ve bu katsayının sebep olduğu sapma 8574.392 olarak hesaplanmıştır. 3 iterasyon sonucunda sapma değeri 8523.368 olarak bulunur. Sonraki iterasyon adımlarında ise sapma değerinde 0.001 den daha küçük değişim olduğu için iterasyona son verilir. Ve 4. iterasyonun sonunda sabit katsayı 1.397 olarak hesaplanmıştır.

Tablo 5.3. Sadece sabit terimi içeren başlangıç modeli

İterasyon		-2 Log L	Sabit Katsayı
Step 0	1	8574.392	1.207
	2	8523.498	1.387
	3	8523.368	1.397
	4	8523.368	1.397

Tablo 5.4 ise sadece sabit terimi içeren modele ait katsayı değeri, standart hata, Wald istatistiği ve serbestlik derecesini göstermektedir. Başlangıç bloğunda tahmin edilen logit, $E(x_i) = g(x_i) = b_0 = 1.397$ olarak elde edilir. İlgili katsayının standart hatası 0.027 ve Wald istatistiği 2653.860 olarak bir serbestlik derecesinde elde edilir. Wald istatistiği ile katsayının önemliliği test edilmiş olur. $c_{0.05,1}^2 = 3.84$ olarak ki-kare tablosundan bulunur ve $2653.860 > 3.84$ olduğundan katsayının modele önemli bir katkısı vardır.

Tablo 5.4. Sadece sabit terimi içeren başlangıç modeli katsayı değeri

		b	S.E.	Wald	df
Step 0	Constant	1.397	0.027	2653.860	1

Daha sonra iteratif yöntem Newton-Raphson metodu bütün değişkenleri modele alarak katsayılarını belirlemek için minimum sapma değeri sağlayacak şekilde iterasyona devam eder. Bu iterasyon adımları detaylı bir şekilde EK-5 de görülebilmektedir. Her bir iterasyonda hesaplanan katsayı tahminleri kullanılarak eşitlik (4.15)'teki uydurulmuş yanıt olasılıklar vektörü ve V matrisi hesaplanarak bir sonraki iterasyonda kullanılmaktadır. Örneğin, ilk iterasyon adımında OTUR_SUR

değişkeni için belirlenen katsayı tahmini -0.005'dir. İkinci adımda ise -0.007 olarak hesaplanmıştır ve daha sonraki iterasyon adımlarında hep -0.007 kullanılmıştır. Bunun sebebi modelin sapma değerinde, sıfıra çok yakın bir katkısı olduğundandır. OTUR_SUR değişkeninin katsayı tahmini 2 iterasyon adımında belirlenebilmişken EFB_OECD değişkeninin katsayı tahmini 3 iterasyonda belirlenebilmiştir. SPSS 13.0 paket programında maksimum yapılacak iterasyon adım sayısı belirlenebilmektedir. Bu çalışmada 20 iterasyon adımı seçilmiştir. Ancak, 20 adıma gerek kalmadan modelin sapma değerindeki yakınsamaya 10 iterasyon adımında ulaşılmıştır. Bu adımdan sonraki değişim miktarları 0.001 den daha az azalmakta olduğundan büyük bir değişim sağlamamaktadır. Modelin sapma değerinin yakınsaması sonucunda elde edilen katsayı tahminlerinin analiz sonucu Tablo 5.5'teki gibidir.

Tablo 5.5'te katsayı değerleri, standart hataları ve Wald istatistik değerleri yer almaktadır. Wald istatistikleri kullanılarak hangi açıklayıcı değişkenlerin istatistiksel açıdan modele önemli bir katkıda bulunduğu belirlenebilir.

MLE tahmin edici ile kurulan modeli yazabilmek için, her bir açıklayıcı değişken için ayrı ayrı hipotez testi kullanarak modele önemli bir katkı sağlayıp sağlamadığına bakılır:

$H_0 : b_i = 0$ (x_i nin modele önemli bir katkısı yoktur.)

$H_1 : b_i \neq 0$ (x_i nin modele önemli bir katkısı vardır.)

$\alpha = 0.05$ önem düzeyinde, OTUR_SUR için Wald istatistiği, ki-kare tablo değeri ile karşılaştırılırsa, $8.601 > c_{0.05,1}^2 = 3.84$ olduğundan kurulan modele istatistiksel açıdan önemli bir katkı sağladığı görülür ve böylece modele dahil edilir.

Tablo 5.5. Açıklayıcı değişkenler ve Wald istatistikleri-1

Açıklayıcı Değişkenler	<i>b</i>	S.E.	Wald	df
OTUR_SUR	-0.007	0.002	8.601	1
YAKIT			13.802	9
YAKIT(1)	0.192	0.075	6.477	1
YAKIT(2)	0.121	0.088	1.893	1
YAKIT(3)	-0.022	0.376	0.004	1
YAKIT(4)	0.175	0.396	0.196	1
YAKIT(5)	19.866	20048.627	0.000	1
YAKIT(6)	-0.727	0.369	3.881	1
YAKIT(7)	-0.058	0.142	0.168	1
YAKIT(8)	-0.092	0.157	0.347	1
YAKIT(9)	-0.123	0.236	0.269	1
EFB_OECD	0.418	0.040	111.636	1
EGITIM			107.290	10
EGITIM(1)	0.146	0.142	1.045	1
EGITIM(2)	0.506	0.100	25.508	1
EGITIM(3)	-1.940	1.180	2.704	1
EGITIM(4)	0.669	0.130	26.677	1
EGITIM(5)	2.020	1.033	3.821	1
EGITIM(6)	0.599	0.130	21.111	1
EGITIM(7)	1.050	0.163	41.438	1
EGITIM(8)	1.211	0.228	28.175	1
EGITIM(9)	1.401	0.177	62.795	1
EGITIM(10)	1.630	0.614	7.043	1
Constant	-0.069	0.139	0.244	1

YAKIT için ise $13.802 < c_{0.05,9}^2 = 16.92$ 'dir. Yani YAKIT değişkeni modele önemli bir katkı sağlamamaktadır. Ayrıca, sabit terim de $0.244 < c_{0.05,1}^2 = 3.84$ olduğundan modele önemli bir katkı sağlamamaktadır. Bu değişkenler modele dahil edilmeyecektir. EFB_OECD ve EGITIM için ise sırasıyla Wald istatistikleri 111.636 ve 107.290 olarak hesaplanmıştır. Bu değerler de ki-kare tablo değerlerinden (EK 4) büyüktür. Böylelikle EFB_OECD ve EGITIM değişkenleri de modele önemli katkı sağladıklarından modele dahil edilecektir.

YAKIT değişkeni ve sabit terim modele önemli bir katkı sağlamadığından modelden çıkarılmış ve kalan değişkenlerle tekrar model uydurulmuştur. Uydurulan

bu modelde de öncelikle çoklu iç ilişkinin mevcudiyeti incelenmiştir. MATLAB programı yardımıyla $f_{Lojistik} = X'VX$ matrisinin öz değerleri hesaplanmış ve özdeğerler büyükten küçüğe sıralı olarak 4.8210, 0.0387, 0.0037, 0.0013, 0.001, 0.0007, 0.0005, 0.0003, 0.0002 ve 3 adet de sıfır elde edilmiştir. Ve böylelikle yine bazı gözlemlere ait öz değerlerin sıfıra yakınsadığı görülmüştür. Bunun sebebi daha önce bahsedildiği gibi kukla değişken kullanılmasıdır. Lojistik regresyon çok seviyeli kukla değişken kullanımına çok hasas olduğundan çoklu iç ilişkiye sebep olabilmektedir. Tekrardan bu çoklu iç ilişki göz ardı edilmiştir.

İteratif yöntem Newton-Raphson metodu kalan bütün değişkenleri modele alarak katsayılarını belirlemek için minimum sapma değeri sağlayacak şekilde iterasyona devam eder. Bu iterasyon adımları detaylı bir şekilde EK-6 da görülebilmektedir. Her bir iterasyonda hesaplanan katsayı tahminleri kullanılarak eşitlik (4.15)'teki uydurulmuş yanıt olasılıklar vektörü ve V matrisi hesaplanarak bir sonraki iterasyonda kullanılmaktadır. Örneğin, ilk iterasyon adımında OTUR_SUR değişkeni için belirlenen katsayı tahmini -0.004'dir. İkinci adımda ise -0.007 olarak hesaplanmıştır ve daha sonraki iterasyon adımlarında hep -0.007 kullanılmıştır. Bunun sebebi modelin sapma değerinde sıfıra çok yakın bir katkısı olduğundandır. OTUR_SUR değişkeninin katsayı tahmini 2 iterasyon adımında belirlenebilmişken EFB_OECD değişkeninin katsayı tahmini 3 iterasyonda belirlenebilmiştir. Modelin sapma değerindeki yakınsamaya 6 iterasyon adımında ulaşılmıştır. Bu adımdan sonraki değişim miktarları 0.001 den daha az azalmakta olduğundan büyük bir değişim sağlamamaktadır. Modelin sapma değerinin yakınsaması sonucunda elde edilen katsayı tahminlerinin analiz sonucu Tablo 5.6'daki gibidir.

Tablo 5.6. Açıklayıcı değişkenler ve Wald istatistikleri-2

Açıklayıcı Değişkenler	<i>b</i>	S.E.	Wald	df
OTUR_SUR	-0.007	0.002	13.138	1
EFB_OECD	0.407	0.029	193.861	1
EGITIM			178.106	10
EGITIM(1)	0.147	0.128	1.305	1
EGITIM(2)	0.511	0.073	48.499	1
EGITIM(3)	-1.921	1.168	2.703	1
EGITIM(4)	0.683	0.107	40.897	1
EGITIM(5)	2.042	1.030	3.927	1
EGITIM(6)	0.621	0.105	35.211	1
EGITIM(7)	1.085	0.142	58.255	1
EGITIM(8)	1.266	0.213	35.415	1
EGITIM(9)	1.441	0.154	87.541	1
EGITIM(10)	1.685	0.605	7.750	1

Her bir değişken için tekrar hipotez testleri kurulup incelenirse, OTUR_SUR, EFB_OECD ve EGITIM değişkenlerinin modele önemli katkılarının olduğu görülür. $E(y_i) = g(x_i)$, tahmin edilen logit fonksiyonu, Tablo 5.6'teki katsayıların tahminleri kullanılarak yazılabilir:

$$g(x_i) = -0.007x_1 + 0.147x_{31} + 0.511x_{32} - 1.921x_{33} + 0.683x_{34} + 2.042x_{35} \\ + 0.621x_{36} + 1.085x_{37} + 1.266x_{38} + 1.441x_{39} + 1.685x_{310} + 0.407x_4$$

Olasılıkların tahmini ise eşitlik (4.7) kullanılarak,

$$\hat{p}_i = \frac{e^{g(x_i)}}{1 + e^{g(x_i)}} = \frac{e^{-0.007x_1 + 0.147x_{31} + 0.511x_{32} - 1.921x_{33} + 0.683x_{34} + 2.042x_{35} + 0.621x_{36} + 1.085x_{37} + 1.266x_{38} + 1.441x_{39} + 1.685x_{310} + 0.407x_4}}{1 + e^{-0.007x_1 + 0.147x_{31} + 0.511x_{32} - 1.921x_{33} + 0.683x_{34} + 2.042x_{35} + 0.621x_{36} + 1.085x_{37} + 1.266x_{38} + 1.441x_{39} + 1.685x_{310} + 0.407x_4}}$$

şeklinde elde edilebilir.

Tablo 5.7, 10 farklı senaryo uygulamasını göstermektedir. Çeşitli özelliklere sahip hanelerin giyim ve ayakkabı harcama grubuna harcama yapma olasılıkları tahmin edilmiştir. Örneğin, 7 nolu senaryoda, hanehalkı ikamet ettikleri hanede 8 yıldır bulunmaktadır. Hanedeki referans ferдин eğitim durumu Orta Dengi Meslek

okuludur ve hanehalkı büyüklüğü eş değer fert cinsinden 2.3 tür. Bu özelliklere sahip 7 nolu senaryonun giyim ve ayakkabı harcama grubuna harcama yapma olasılığı % 94.89'dir.

Tablo 5.7. Senaryolar

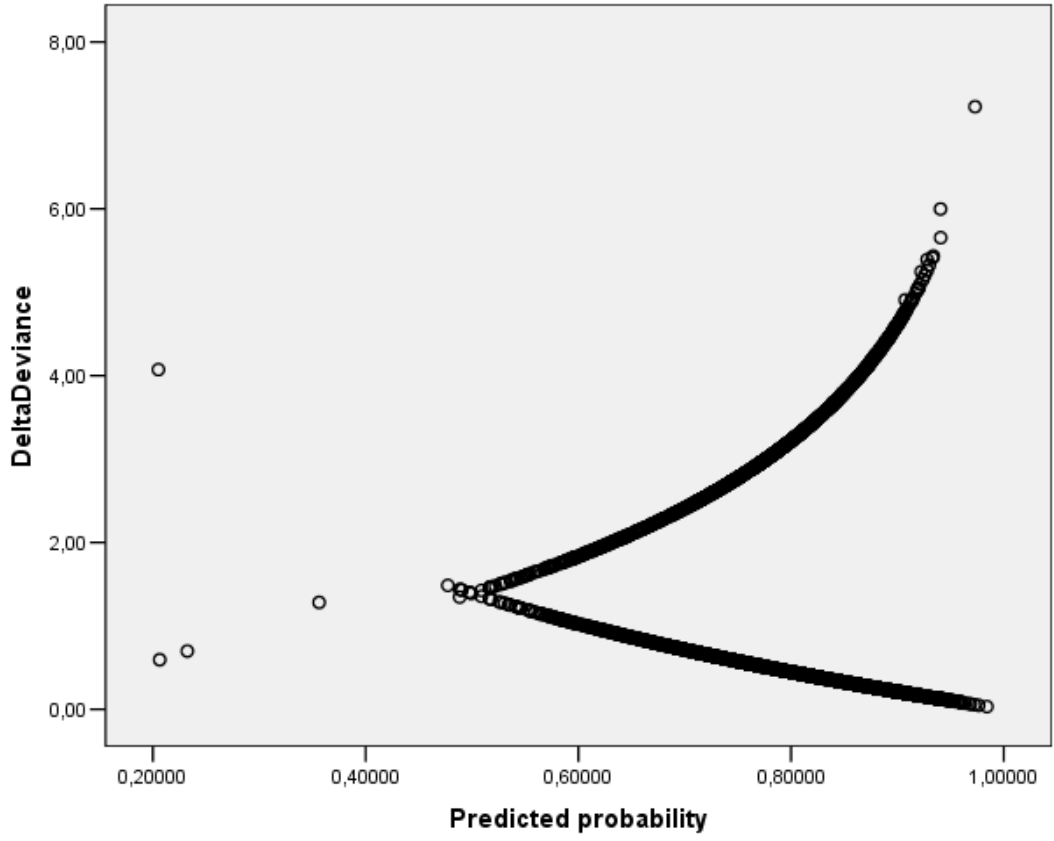
Senaryo	OTUR_SUR	EGITIM	EFB_OECD	$g(x_i)$	$\hat{p}_i$
1	10	Lise	1.5	1.161500	0.761605
2	12	Lise Dengi Meslek	2	1.815000	0.859965
3	5	Ortaokul	1.8	1.380600	0.799087
4	1	İlkokul	3.4	1.887800	0.868504
5	26	Doktora	1	1.910000	0.871019
6	7	Okur-Yazar Değil	3.1	1.212700	0.770776
7	8	Orta Dengi Meslek	2.3	2.922100	0.948928
8	2	Lise	2	1.421000	0.805495
9	14	Okur-Yazar Değil	1.8	0.634600	0.653532
10	5	Yüksek Lisans	3.4	3.033800	0.954078

Tablo 5.8'de eşitlik (4.57) ile hesaplanan Hosmer-Lemeshow testine ait değerler bulunmaktadır. Ki-kare tablosundan, $c_{0.05,8} = 15.51$ 'dir ve 15.178'ten çok az farkla büyüktür. Bu da uydurulan modelin yeterli olduğunu yani istatistiksel açıdan önemli olduğunu gösterir. Ancak tablo değerine çok yakın olduğu da unutulmamalıdır.

Tablo 5.8. Hosmer-Lemeshow testi

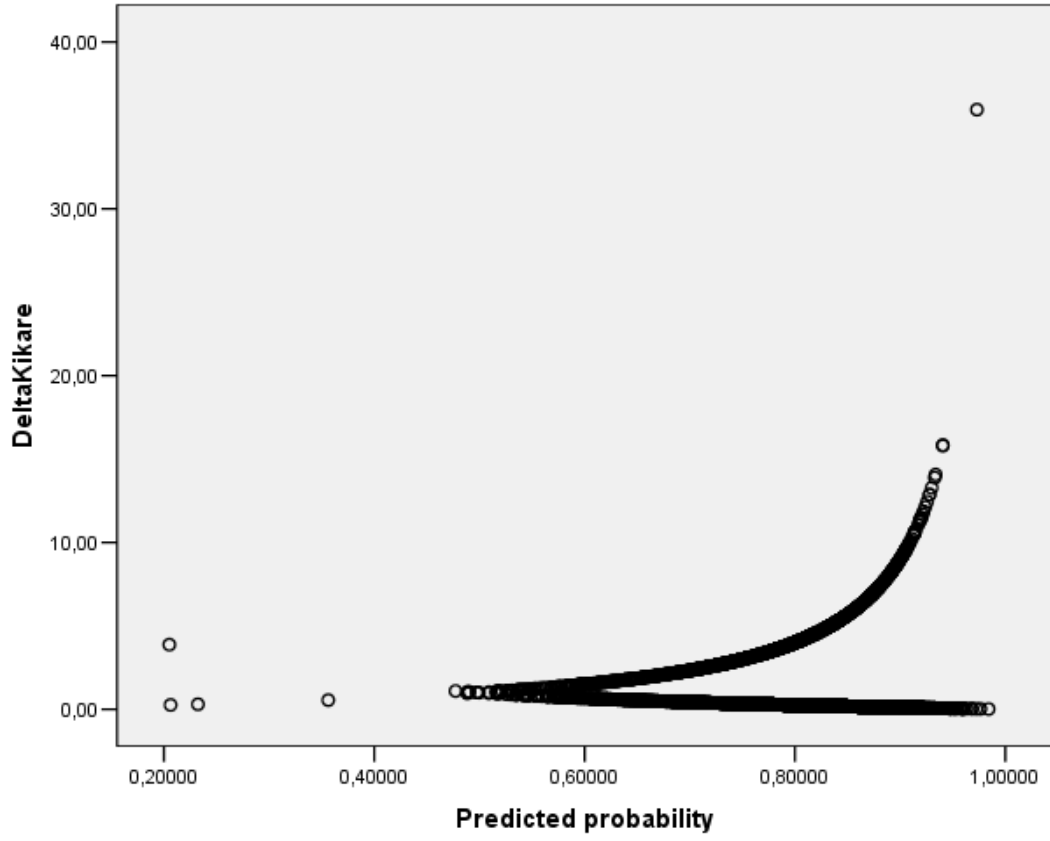
Chi-square	df	Sig.
15.178	8	0.056

İyi uyum sağlamayan gözlemleri belirlemek için ΔX_j^2 , ΔD_j ve $\Delta \hat{b}_j$ ye karşın $\hat{p}_j$ grafiklerinin ve ΔX_j^2 , ΔD_j ve $\Delta \hat{b}_j$ ye karşın h_j grafiklerinin incelemesi faydalı olacaktır:



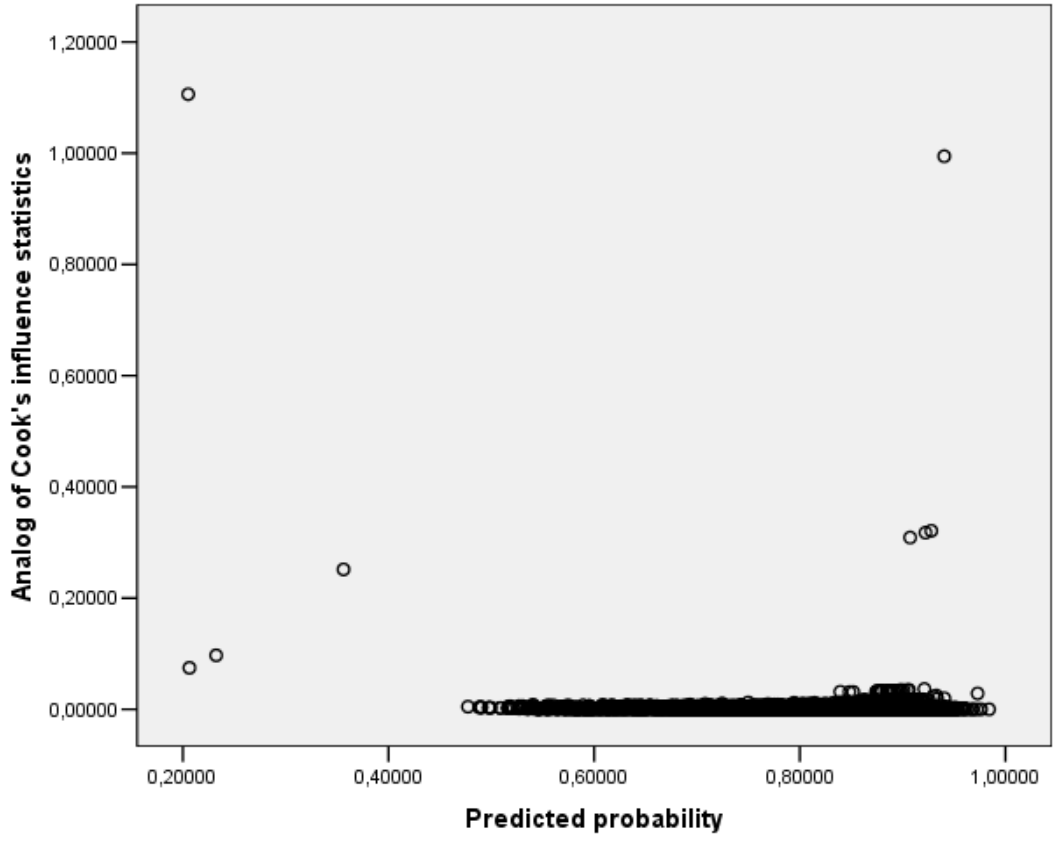
Şekil 5.1. ΔD_j ye karşın $\hat{p}_j$ grafiği

Şekil 5.1'e bakılırsa, $\hat{p}_j < 0.40$ olan 4 gözlem ve ΔD_j değeri en büyük olan gözlem ayırık durmaktadır. Bu gözlemler, 606301, 605282, 601463, 605060 ve 607852 bülten numaralı hanelere ait gözlemlerdir (Bülten numarası her bir haneye has verilen kimlik numaralarıdır).



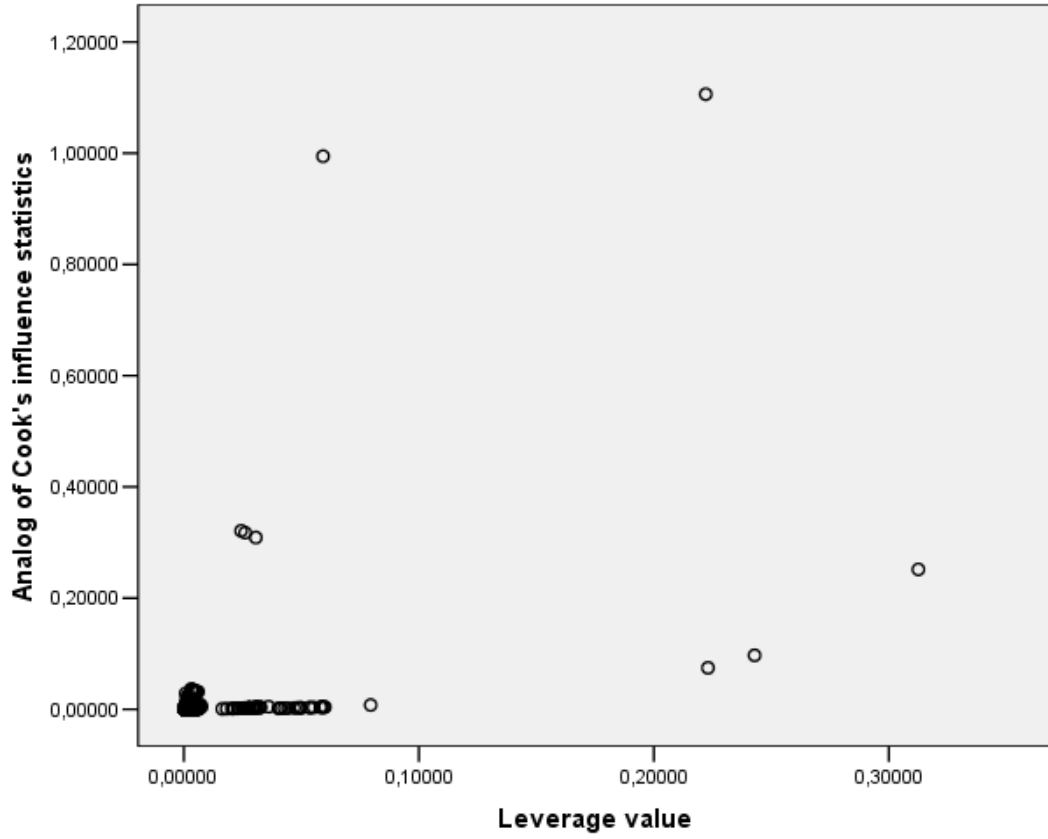
Şekil 5.2. ΔX_j^2 ye karşın $\hat{p}_j$ grafiği

ΔX_j^2 ye karşın $\hat{p}_j$ grafiği (Şekil 5.2) incelenirse, $\hat{p}_j < 0.40$ olan 4 gözlem ve $\Delta X_j^2 > 30$ olan bir gözlem ayırık durmaktadır. Bu gözlemler ΔD_j ye karşın p_j grafiğinden elde edilen gözlemlerle aynıdır.



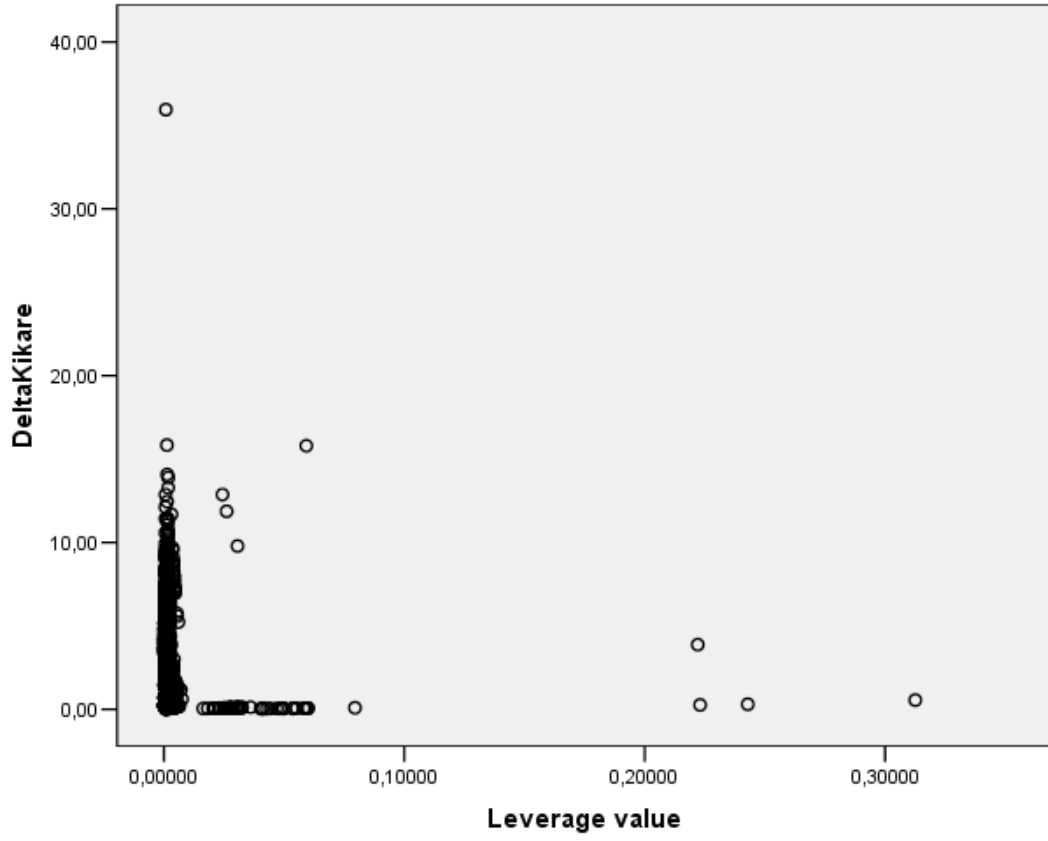
Şekil 5.3. $\Delta \hat{b}_j$ ye karşın $\hat{p}_j$ grafiği

Şekil 5.3’de ise, $\hat{p}_j < 0.40$ olan 4 gözlem ve Cook değeri en büyük olan gözlem ayırık durmaktadır. Bu gözlemler 606301, 605282, 601463 ve 605060 bülten numaralı hanelere ait gözlemlerdir.



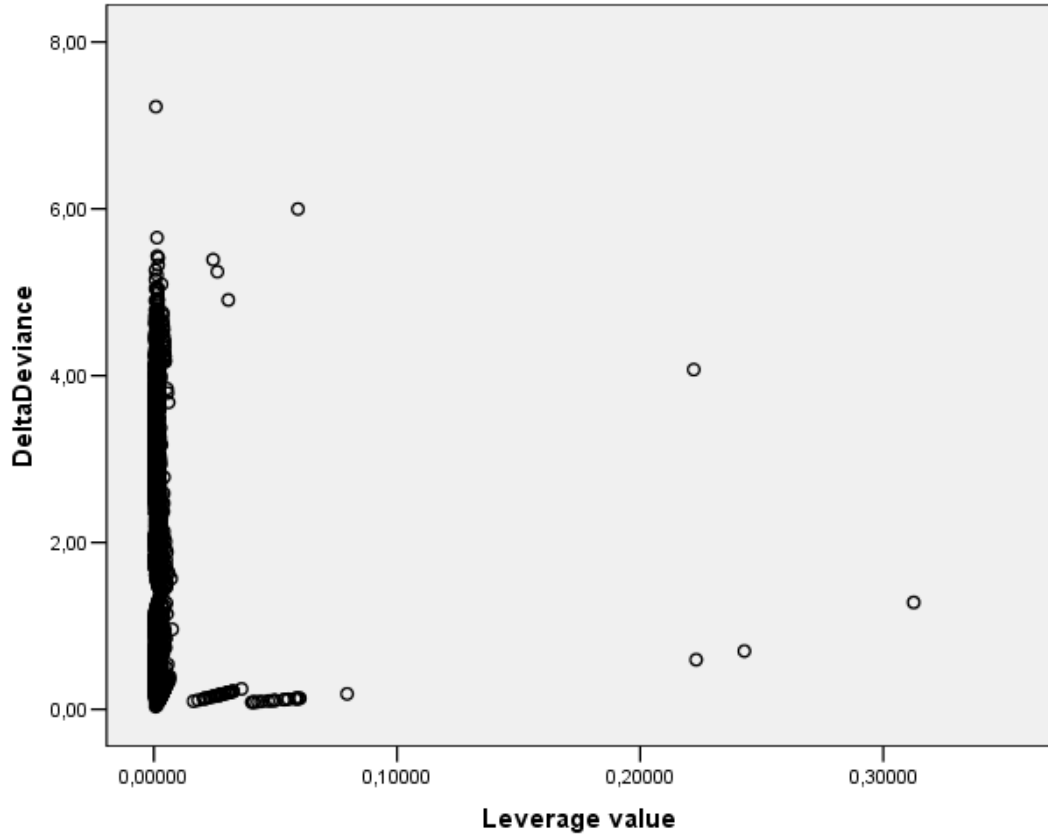
Şekil 5.4. $\Delta \hat{b}_j$ ye karşın h_j grafiği

Şekil 5.4'e bakılırsa, $h_j > 0.1$ olan gözlemler ve Cook değeri en büyük olan gözlem ayırık durmaktadır. Bu gözlemler 605060, 605282, 601463 ve 606301 bülten numaralı hanelere ait gözlemlerdir.



Şekil 5.5. ΔX_j^2 ye karşın h_j grafiği

ΔX_j^2 ye karşın h_j grafiğine (Şekil 5.5) bakılırsa, $h_j > 0.1$ olan gözlemler ve ΔX_j^2 değeri en büyük olan gözlem ayırık durmaktadır. Bu gözlemler, 605060, 605282, 601463, 606301 ve 607852 bülten numaralı hanelere ait gözlemlerdir.

Şekil 5.6. ΔD_j ye karşın h_j grafiği

Şekil 5.6 incelenirse, $h_j > 0.1$ olan gözlemler ve ΔD_j değeri en büyük olan gözlem ayırık durmaktadır. Bu gözlemler, ΔX_j^2 ye karşın h_j grafiğinden elde edilen gözlemlerle aynıdır.

606301, 605282, 601463 ve 605060 bülten numaralı hanelere ait gözlemler 6 grafikte de ortak olarak bulunmuş uydurulan modele iyi uyum sağlayamamış gözlemlerdir. Tablo 5.9’de ise bu gözlemlere ait değerler bulunmaktadır. Örneğin, 605282 bülten numaralı hanede 6 yıldır oturulmaktadır. Eş değer hane büyüklüğü 1.5’dir. Ve hanedeki referans kişi, ilköğretim mezunudur. Bu gözleme ait giyim ve ayakkabıya harcama olasılığı ise %20,5 olarak tahmin edilmiştir. Gözleme ait $\Delta \hat{b}_j$ değeri ise 1.10628 olarak hesaplanmıştır. Bu rakam en büyük $\Delta \hat{b}_j$ değeridir. Yani bu gözlem, tahmin edilen katsayılar üzerinde en çok etkisi bulunan gözlemdir.

Tablo 5.9. İyi uyum sağlamayan gözlemlere ait değerler

	606301	605282	601463	605060
OTUR_SUR	2	6	5	1
EGITIM	İlköğretim	İlköğretim	İlköğretim	İlköğretim
EFB_OECD	3.3	1.5	1.5	1.8
y_j	0	1	0	0
p_j	0.35625	0.20511	0.20632	0.23232
$\Delta \hat{b}_j$	0.25157	1.10618	0.07462	0.09706
ΔX_j^2	0.55	3.88	0.26	0.30
ΔD_j	1.28	4.07	0.59	0.7
h_j	0.31252	0.22205	0.22302	0.24284

Tablo 5.10’da ise gözlemler atılmadan önce ve atıldıktan sonra uyumun iyiliği ölçütleri görülebilmektedir. Gözlemlerin tek tek yada toplu bir şekilde veri setinden çıkarılması $-2\text{Log}L$ ve c^2 değerleri açısından çok bariz farklılıklara yol açmamıştır. Ancak Hosmer-Lemeshow test istatistiğinde en fazla azalmayı sağlayan 605282 bülten numaralı haneye ait gözlem olmuştur. Ayrıca, gözlemlerin toplu bir şekilde çıkarılması bu test istatistiğinde daha da azalma sağlayarak 13.64 değerine ulaşmıştır. Gözlemler atılmadan önceki değerine göre azalma sağlaması modeldeki uyumun iyiliğinin daha da attığını göstermektedir.

Tablo 5.10. Gözlemler atılmadan önce ve sonra uyumun iyiliği ölçütleri

	Gözlemler Atılmadan Önce	Sadece 606301 Atılırsa	Sadece 605282 Atılırsa	Sadece 601463 Atılırsa	Sadece 605060 Atılırsa	4 Gözlemin Tamamı Atıldıktan Sonra
$-2\text{Log}L$	8235.068	8233.961	8230.027	8234.535	8234.477	8230.027
c^2	3626.067	3625.787	3629.721	3625.214	3625.301	3625.562
$\hat{C}$	15.178	14.478	14.022	15.182	15.180	13.640

6. SONUÇ VE ÖNERİLER

Bu çalışmada, ilk olarak, ağırlıklandırılmış en küçük kareler yöntemine alternatif olarak sunulmuş GLM'ye ait parametre tahmini incelenmiştir. GLM'de çoklu iç ilişkinin bulunmadığı durumlarda parametre tahmini maksimum likelihood yöntemi ile yapılmaktadır. Buna karşın, çoklu iç ilişkinin bulunduğu durumlarda ise çoklu iç ilişkinin etkilerini en aza indirmek için kullanılan ridge GLM tahmin edici ve temel bileşenler GLM tahmin edici anlatılmıştır. Ayrıca bu tahmin edicilere ait hata kareler ortalamaları verilmiştir. Bunların yanı sıra, GLM'de istatistiksel sonuç çıkarımlarından bahsedilmiştir.

Ridge GLM tahmin edicide, ridge parametresi k 'nın seçiminde sadece Smith ve Marx (1990) tarafından Mallows'un (1973) C_p kriterine benzer şekilde önerilen C_p^R kriteri bulunmaktadır. Literatürde k 'nın seçimi konusunda başka bir yöntem bulunmamaktadır. Bu, üzerinde çalışılabilecek bir konu olarak araştırmacıların dikkatini beklemektedir.

Poisson regresyon ise GLM'de yanıt değişkeninin bir sayım verisi olduğu durumlarda kullanılmaktadır. GLM'de çoklu iç ilişkinin bulunduğu durumlarda parametre tahmini için kullanılan yöntemlerin özel hali Poisson regresyon için sunulmuştur. Ve hata kareler ortalamaları verilmiştir.

GLM'de yanıt değişkeninin Bernoulli dağılımından olduğu durumlarda ise lojistik regresyon kullanılmaktadır. İlk olarak, lojistik regresyonun kullanım alanlarına ait çalışmalar ve teorik açıdan gelişim süreci anlatılmıştır. Lojistik regresyon modelinde maksimum likelihood parametre tahmini anlatıldıktan sonra ise çoklu iç ilişkinin bulunduğu durumlarda ridge lojistik tahmin edici, temel bileşenler lojistik tahmin edici ve bunlara ek olarak Stein lojistik tahmin edici anlatılmış ve hata kareler ortalamaları verilmiştir. Ayrıca, lojistik regresyonda istatistiksel çıkarımlar incelenmiş ve uyumun iyiliği ölçütleri verilmiştir.

Son olarak, TÜİK'ten elde edilen 2006 yılı HBA (Hanehalkı Bütçe Anketi) verileri lojistik regresyon kullanılarak analiz edilmiştir. Anket kapsamında çalışma yapılan hanelerin COICOP sınıflamasının 3. harcama grubu olan giyim ve ayakkabı

harcamaları hesaplanmıştır. Ve yanıt değişken olarak hanenin giyim ve ayakkabı harcaması yapıp yapmadığı alınmıştır. Veri setinde çoklu iç ilişkinin olup olmadığı araştırılmış ve çoklu iç ilişkinin bulunduğu tespit edilmiş ancak açıklayıcı değişkenlerden bazıları için çok seviyeli kukla değişken kullanıldığı için çıkan çoklu iç ilişki göz ardı edilerek maksimum likelihood yöntemi kullanılarak analize devam edilmiştir. Modele önemli katkısı olan katsayı tahminleri ile lojistik model yazılmıştır. Daha sonra iyi uyum sağlamayan gözlemler veri setinden çıkarılıp tekrar model uydurulmuş ve uyumun iyiliği ölçütleri kullanılarak karşılaştırma yapılmıştır. Örnek uygulamada iyi uyum sağlamayan gözlemlerin veri setinden çıkarılması uyumun iyiliğini arttırmıştır. Ayrıca, uydurulan lojistik model kullanılarak farklı senaryolar oluşturulmuş ve sonuçlar yorumlanmıştır.

KAYNAKLAR

- AGRESTI, A., 2002. Categorical Data Analysis. John Wiley & Sons, Canada, pp.710.
- AGUILERA, A. M., ESCABIAS, M., and VALDERRAMA, M. J., 2006. Using Principal Components for Estimating Logistic Regression with High-Dimensional Multicollinear Data. Computational Statistics & Data Analysis, 50: 1905-1924.
- AKKAYA, S., ve PAZARLIOĞLU, M. V., 1998. Ekonometri. Erkan Matbaacılık, İzmir.
- ALBERT, A., and ANDERSON, J. A., 1984. On the Existence of Maximum Likelihood Estimates in Logistic Regression Models. Biometrika, 71(1), 1-10.
- ALPAR, R., 1997. Uygulamalı Çok Değişkenli İstatistiksel Yöntemlere Giriş 1. Bağırğan Yayınevi, Ankara.
- ARANDA-ORDAZ, F. J., 1981. On Two Families of Transformations to Additivity for Binary Response Data. Biometrika, 68, 357-363.
- BARKER, L., and BROWN, C., 2001. Logistic Regression When Binary Predictor Variables Are Highly Correlated. Statistics in Medicine, 20: 1431-1442.
- BERKSON, J., 1944. Application of the Logistic Function to Bio-Assay. Journal of the American Statistical Association, 9, 357-365.
- BIRCAN, H., 2004. Lojistik Regresyon Analizi: Tıp Verileri Üzerine Bir Uygulama. Kocaeli Üniversitesi Sosyal Bilimler Enstitüsü Dergisi, 2004 / 2 : 185-208.
- BONNEY, G. E., 1987. Logistic Regression for Dependent Binary Observations. Biometrics (43), p.951-973.
- BUESCHER, P. A., LARSON, L. C., NELSON, M. D., and LENIHAN, A. J., 1993. Prenatal WIC Participation Can Reduce Low Birth Weight and Newborn Medical Costs: A Cost Benefit Analysis of Wic Participation in North Carolina. Journal of the American Dietetic Association, 93:163-166.

- CANKURT, M., GÜNDEN, C., ve MİRAN B., 2007. Türkiye'nin AB Sürecinde Üyelik Potansiyelinin Tarımsal ve Diğer Bazı Önemli Kriterler Açısından Belirlenmesi. Finans Politik & Ekonomik Yorumlar, Cilt: 44 Sayı:513, 35-45.
- COOK, R. D., and WEISBERG, S., 1982. Residuals and Influence in Regression. sect. 5.4, London, Chapman and Hall.
- COX, D. R., 1970. The Analysis of Binary Data. Methuen, London.
- DEMARIS, A., 1995. A Tutorial in Logistic Regression. Journal of the Marriage and the Family, 57, 956-968.
- DENİZ, Ö., 2005. Poisson Regresyon Analizi. İstanbul Ticaret Üniversitesi Fen Bilimleri Dergisi, 7, 59-72.
- DUFFY, D. E., and SANTNER, T. J., 1989. On the Small Sample Properties of Norm-Restricted Maximum Likelihood Estimators for Logistic Regression Models. Commun Statist. Theory Meth., 18, 959-980.
- ELHAN, A. H., 1997. Lojistik Regresyon Analizinin İncelenmesi ve Tıpta Bir Uygulaması. Yüksek Lisans Tezi, Ankara Üniversitesi, Fen Bilimleri Enstitüsü.
- FINNEY, D. J., 1971. Probit Analysis. 3rd edn, Cambridge University, London.
- GARSIDE, P. S., and GLUECK, C. J., 1995. The Important Role of Modifiable Dietary and Behaviour Characteristic in the Causation and Prevention of Coronary Heart Disease Hospitalization and Mortality. Journal of American College of Nutrition, 14: 71-79.
- GORDON, T., 1974. Hazards in the Use of Logistic Function with Special Reference to Data from Prospective Cardiovascular Studies. Journal of Chronic Diseases, 27, 97-102.
- HALPERIN, M., BLACKWELDER, W. L., and VERTER, J. I., 1971. Estimation of the Multivariate Logistic Risk Function: A Comparison of the Discriminant Function and Maximum Likelihood Approaches. Journal of the Chronic Disease, 24, 125-158.
- HEATH, A., MCMAHON, D., and ROBERTS, J., 2000. Ethnic Differences in the Labour Market: A Comparison of the Samples of Anonymized Records and

- Labour Force Survey. *Journal of the Royal Statistical Society, Series A*, 163, part 3, 341-361.
- HILL, R. C., FOMBY, T. B., and JOHNSON, S. R., 1977. Component Selection Norms for Principal Component Regression. *Communications in Statistics A: Theory and Methods*, 6, 309-334.
- HOERL, A. E., and KENNARD, R. W., 1970a. Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, 12, 1, 55-67.
- HOERL, A. E., and KENNARD, R. W., 1970b. Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, 12, 1, 69-82.
- HOERL, A. E., KENNARD, R. W., and BALDWIN, K. F., 1975. Ridge Regression: Some Simulation. *Communication in Statistics*, 4, 105-123.
- HOSMER, D. W., and LEMESHOW, S., 1980. A Goodness-of-Fit Test For the Multiple Logistic Regression Model. *Communications in Statistics*, A10, 1043-1069.
- HOSMER, D. W., and LEMESHOW, S., 2000. *Applied Logistic Regression*. John Wiley and Sons, Canada.
- JOHNSON, W., 1985. Influence Measures for Logistic Regression. Another Point of View, *Biometrika*, 72 (1), 59-65.
- JOLLIFFE, I. T., 1986. *Principal Component Analysis*. New York, Springer-Verlag.
- JUDGE, G. G., GRIFFITHS, W. E., HILL, R. C., LÜTKEPOHL, H., and LEE, T. C., 1985. *The Theory and Practice of Econometrics*. John Wiley and Sons, New York, 1018p.
- KAISER, H. F., 1958. The Varimax Criterion for Analytic Rotation in Factor Analysis. *Psychometrika*, 23: 187-200.
- KAŞKO, Y., 2007. Çoklu Bağlantı Durumunda İkili (Binary) Lojistik Regresyon Modelinde Gerçekleşen I. Tip Hata ve Testin Gücü. Yüksek Lisans Tezi, Ankara Üniversitesi, Fen Bilimleri Enstitüsü.
- KLOIBER, L. L., WINN, N. J., SHAFFER, S. G., and HASSANEIN, R. S., 1996. Late Hyponatremia in Very Low Birth Weight Infants: Incidence and Associated Risk Factors. *Journal of the American Dietetic Association*, 96: 880-884.

- LECESSIE, S., and VAN HOUWELINGEN J. C., 1992. Ridge Estimators in Logistic Regression. *Applied Statistics*, 41, No. 1, 191-201.
- LEE, A. H., and SILVAPULLE, M. J., 1988. Ridge Estimation in Logistic Regression. *Communications in Statistics, part B-Simulations and Computations* 4: 1231-1257.
- LESAFFRE, E., and MARX, B. D., 1993. Collinearity in Generalized Linear Regression. *Communications in Statistics Theory and Methods*, 22, 1933 - 1952.
- MACKINNON, M. J., and PUTERMAN, M. L., 1989. Collinearity in Generalized Linear Models. *Communications in Statistics Theory and Methods*, 18, 3463-3472.
- MALLOWS, C. L., 1973. Some Comments on C_p . *Technometrics*, 15, 661-675.
- MARX, B. D., 1988. Ill-Conditioned Information Matrices and the Generalized Linear Models: An Asymptotically Biased Estimation Approach. Doctorial Dissertation, Virginia: Virginia Polytechnic Institute and State University.
- MARX, B. D. And SMITH, E. P., 1990. Ill-Conditioned Information Matrices, Generalized Linear Models and Estimation of the Effects of Acid Rain. *Environmetrics Press, Canada*, 1 (1), 57-71.
- MONTGOMERY, D. C., PECK, E. A., and VINING, G. G., 2001. Introduction to Linear Regression Analysis. John Wiley and Sons, Canada
- MYERS, R. H., 1990. Classical and Modern Regression with Applications. Duxbury Press: California.
- O'NEILL, T. J., and BARRY, S. C., 1995. Truncated Logistic Regression. *Biometrics*, 51, 533-541.
- OĞUZ, B., ve BUDAK, F., 2005. Evsel Katı Atıkların Geri Kazanımı Programlarına Katılımı Etkileyen Faktörlerin Belirlenmesi. 6. Ulusal Çevre Mühendisliği Kongresi, İstanbul.
- ÖNDER, H., 2001. Lojistik Regresyonlarda Değişken Seçimi. Yüksek Lisans Tezi, Zootekni Anabilim Dalı, Çukurova Üniversitesi.

- ÖZÇOMAK, M. S., OKTAY, E., ve ÖZER, H., 2006. Erzurum İlinde Potansiyel Doğal Gaz Talebini Etkileyen Faktörlerin Tespiti. EKEV Akademi Dergisi, 10(27).
- ÖZDAMAR, K., 2004. Paket Programlar ile İstatistiksel Veri Analizi. Kaan Kitabevi.
- ÖZKALE, M. R., 2007. Çoklu İç İlişki İle İlgili Problemler. Doktora Tezi, Çukurova Üniversitesi, Fen Bilimleri Enstitüsü.
- PEOPLES, M. D., SIEGEL, E., SUCHINDRAN, C. M., ORIGASA, H., WARE, A., and BARAKAT, A., 1991. Characteristics of Maternal Employment During Pregnancy: Effects on Low Birthweight. American Journal of Public Health, 81: 1007-1012.
- PREGIBON, D., 1981. Logistic Regression Diagnostics. Ann. Statist., 9, 705-724.
- ROY, S. S., and GRUIA, S., 2008. Diagnostics in Logistic Regression Models. Journal of the Korean Statistical Society, 37, 89-94.
- SABLE, M. R., and HERMAN, A. A., 1997. The Relationship Between Prenatal Health Behaviour Advice and Low Birthweight. Public Health Reports. 112: 332-339.
- SANTOS, I. S., VICTORIA, C. G., HUTTLY, S., and CARVALHAL, J. B., 1998. Caffeine Intake and Low Birth Weight: A Population Based Case Control Study. American Journal of M. (1988), The Retreat From Class: A New True Socialism, London: Verso.
- SAS INSTITUTE INC., 2008. SAS/STAT 9.2 User's Guide. Cary, NC: SAS Institute Inc.
- SCHAEFER, R. L., 1979. Multicollinearity and Logistic Regression. Ph. D. Dissertation, University of Michigan.
- SCHAEFER, R. L., 1986. Alternative Estimators in Logistic Regression When the Data are Collinear. J. Statist. Comput. Simul., 25, 75-91.
- SCHAEFER, R. L., ROÍ, L. D., WOLFE, R. A., 1984. A Ridge Logistic Estimator. Communications in Statistics Theory and Methods, 13, 99-113.
- SMITH, E. P., and MARX, B. D., 1990. Principal Component Estimation for Generalized Linear Regression. Biometrika, Great Britain, 77, 1, 23-31.

- STEIN, C. M., 1960. Multiple Regression, Contributions to Probability and Statistics. Stanford University Press.
- TATLIDİL, H., ve ÖZER, M., 2005. Firma Derecelendirme Çalışmaları Konusunda Çok Değişkenli İstatistiksel Analize Dayalı Karar Destek Sistemlerinin Kullanımı. Bankacılar Dergisi, Sayı 54, 46-58.
- VAGO, E., and KEMENY, S., 2006. Logistic Ridge Regression for Clinical Data Analysis (A Case Study). Applied ecology and environmental research 4 (2): 171-179.

ÖZGEÇMİŞ

Engin Arıcan, 1982 yılında Adana’da dünyaya geldi. İlkokulu, ortaokulu ve liseyi sırasıyla Gaziantep Şahinbey İlkokulu’nda, F.N.T. Anadolu Lisesi’nde (ortaokul kısmı) ve V.D. Fen Lisesi’nde tamamladı. 2000 yılında lisans eğitimine Marmara Üniversitesi Endüstri Mühendisliği bölümünde başladı ve 2005 yılında bu bölümden mezun oldu. Çok iyi seviyede İngilizce bilen, Engin Arıcan halen, Türkiye İstatistik Kurumunda uzman yardımcısı olarak çalışmaktadır. Ayrıca, evli ve bir çocuk babasıdır.

EKLER

1. Fisher Scoring

Fisher Scoring, gözlemlenen rastgele değişken X 'leri içeren, bilinmeyen parametre q 'nın likelihood fonksiyonuna bağlı bilginin miktarını ölçmenin bir yoludur. Bu metodla,

$$E\left(\frac{\partial \mathbf{l}}{\partial q}\right) = 0 \text{ ve } -E\left(\frac{\partial^2 \mathbf{l}}{\partial q^2}\right) = \left[E\left(\frac{\partial \mathbf{l}}{\partial q}\right)\right]^2$$

olduğu verilir (Agresti, 2002).

2. Newton-Raphson Metodu

Newton-Raphson metodu bir kök bulma algoritmasıdır. $\frac{\partial f(q)}{\partial q} = 0$ ifadesinin bir

kökü bulunmak istenir. Böylece maksimumu bulmak amaçlanır. $\frac{\partial f(q)}{\partial q}$ ifadesi

Taylor Serisi ile q_0 etrafında açılacak olursa,

$$\frac{\partial f(q)}{\partial q} = f'(q) = f'(q_0) + \frac{\partial^2 f(q)}{\partial q \partial q'} (q - q_0)$$

olur. Bu ifade sıfıra eşitlenerek, kök için çözüm bulunur.

$$f'(q_0) + \frac{\partial^2 f(q)}{\partial q \partial q'} (q - q_0) = 0$$

$$q = q_0 - \left[\frac{\partial^2 f(q)}{\partial q \partial q'} \right]^{-1} \cdot f'(q)$$

Kökün tahminini geliřtirmek için yukarıdaki eřitlik iteratif olarak gerek köke yakınsamak için kullanılabilir:

$$q^{(m+1)} = q^{(m)} - \left[\frac{\partial^2 f(q)}{\partial q \partial q'} \right]^{-1} \cdot f'(q^{(m)})$$

3. Lagrange arpanı

Matematiksel optimizasyonda, “Lagrange arpanı” metodu bazı kısıtları göz önünde bulundurarak bir fonksiyonun maksimum ya da minimumunu bulmaya yarayan bir yöntemdir. En temel haliyle, $g(x, y) = c$ kısıtı altında $f(x, y)$ fonksiyonu minimum yapılmak istensin. l Lagrange arpanı olmak üzere

$$F(x, y, l) = f(x, y) + l(g(x, y) - c)$$

Langrange fonksiyonu tanımlanır ve x, y, l deęiřkenlerine göre fonksiyonun kısmi türevi alınır. Bu kısmi türevler sıfıra eřitlenip elde edilen denklemler özölür.

4. Ki-kare Tablo Deęerleri

Önem Düzeylerine Göre Bazı Ki-Kare Tablo Deęerleri

Serbestlik Derecesi	$\alpha = 0.05$	$\alpha = 0.10$
1	3.84	2.71
3	7.81	6.25
8	15.51	13.36
9	16.92	14.68
10	18.31	15.99
14	23.68	21.06
59	77.92	73.28
60	79.08	74.40

5. Maksimum Likelihood Tahmin Edici İterasyonları-1

Ite.	EGIT(1)	EGIT(2)	EGIT(3)	EGIT(4)	EGIT(5)	EGIT(6)	EGIT(7)	EGIT(8)	EGIT(9)	EGIT(10)
1	0.118	0.394	-1.749	0.496	0.993	0.455	0.692	0.757	0.841	0.917
2	0.144	0.501	-1.923	0.656	1.658	0.590	0.997	1.128	1.283	1.449
3	0.146	0.506	-1.940	0.669	1.967	0.599	1.049	1.208	1.395	1.617
4	0.146	0.506	-1.940	0.669	2.019	0.599	1.050	1.211	1.401	1.630
5	0.146	0.506	-1.940	0.669	2.020	0.599	1.050	1.211	1.401	1.630
6	0.146	0.506	-1.940	0.669	2.020	0.599	1.050	1.211	1.401	1.630
7	0.146	0.506	-1.940	0.669	2.020	0.599	1.050	1.211	1.401	1.630
8	0.146	0.506	-1.940	0.669	2.020	0.599	1.050	1.211	1.401	1.630
9	0.146	0.506	-1.940	0.669	2.020	0.599	1.050	1.211	1.401	1.630
10	0.146	0.506	-1.940	0.669	2.020	0.599	1.050	1.211	1.401	1.630
11	0.146	0.506	-1.940	0.669	2.020	0.599	1.050	1.211	1.401	1.630
12	0.146	0.506	-1.940	0.669	2.020	0.599	1.050	1.211	1.401	1.630
13	0.146	0.506	-1.940	0.669	2.020	0.599	1.050	1.211	1.401	1.630
14	0.146	0.506	-1.940	0.669	2.020	0.599	1.050	1.211	1.401	1.630
15	0.146	0.506	-1.940	0.669	2.020	0.599	1.050	1.211	1.401	1.630
16	0.146	0.506	-1.940	0.669	2.020	0.599	1.050	1.211	1.401	1.630
17	0.146	0.506	-1.940	0.669	2.020	0.599	1.050	1.211	1.401	1.630
18	0.146	0.506	-1.940	0.669	2.020	0.599	1.050	1.211	1.401	1.630
19	0.146	0.506	-1.940	0.669	2.020	0.599	1.050	1.211	1.401	1.630
20	0.146	0.506	-1.940	0.669	2.020	0.599	1.050	1.211	1.401	1.630

Ite.	YAKIT(1)	YAKIT(2)	YAKIT(3)	YAKIT(4)	YAKIT(5)	YAKIT(6)	YAKIT(7)	YAKIT(8)	YAKIT(9)
1	0.115	0.076	-0.004	0.105	0.868	-0.562	-0.029	-0.050	-0.067
2	0.180	0.115	-0.017	0.164	1.819	-0.713	-0.052	-0.084	-0.113
3	0.192	0.121	-0.022	0.175	2.842	-0.727	-0.058	-0.092	-0.122
4	0.192	0.121	-0.022	0.175	3.857	-0.727	-0.058	-0.092	-0.123
5	0.192	0.121	-0.022	0.175	4.863	-0.727	-0.058	-0.092	-0.123
6	0.192	0.121	-0.022	0.175	5.865	-0.727	-0.058	-0.092	-0.123
7	0.192	0.121	-0.022	0.175	6.866	-0.727	-0.058	-0.092	-0.123
8	0.192	0.121	-0.022	0.175	7.866	-0.727	-0.058	-0.092	-0.123
9	0.192	0.121	-0.022	0.175	8.866	-0.727	-0.058	-0.092	-0.123
10	0.192	0.121	-0.022	0.175	9.866	-0.727	-0.058	-0.092	-0.123
11	0.192	0.121	-0.022	0.175	10.866	-0.727	-0.058	-0.092	-0.123
12	0.192	0.121	-0.022	0.175	11.866	-0.727	-0.058	-0.092	-0.123
13	0.192	0.121	-0.022	0.175	12.866	-0.727	-0.058	-0.092	-0.123
14	0.192	0.121	-0.022	0.175	13.866	-0.727	-0.058	-0.092	-0.123
15	0.192	0.121	-0.022	0.175	14.866	-0.727	-0.058	-0.092	-0.123
16	0.192	0.121	-0.022	0.175	15.866	-0.727	-0.058	-0.092	-0.123
17	0.192	0.121	-0.022	0.175	16.866	-0.727	-0.058	-0.092	-0.123
18	0.192	0.121	-0.022	0.175	17.866	-0.727	-0.058	-0.092	-0.123

19	0.192	0.121	-0.022	0.175	18.866	-0.727	-0.058	-0.092	-0.123
20	0.192	0.121	-0.022	0.175	19.866	-0.727	-0.058	-0.092	-0.123

Ite.	-2 Log L	Constant	OTUR_SUR	EFB_OECD
1	8329.032	0.229	-0.005	0.251
2	8222.187	-0.016	-0.007	0.392
3	8219.833	-0.068	-0.007	0.418
4	8219.750	-0.069	-0.007	0.418
5	8219.722	-0.069	-0.007	0.418
6	8219.711	-0.069	-0.007	0.418
7	8219.708	-0.069	-0.007	0.418
8	8219.706	-0.069	-0.007	0.418
9	8219.706	-0.069	-0.007	0.418
10	8219.705	-0.069	-0.007	0.418
11	8219.705	-0.069	-0.007	0.418
12	8219.705	-0.069	-0.007	0.418
13	8219.705	-0.069	-0.007	0.418
14	8219.705	-0.069	-0.007	0.418
15	8219.705	-0.069	-0.007	0.418
16	8219.705	-0.069	-0.007	0.418
17	8219.705	-0.069	-0.007	0.418
18	8219.705	-0.069	-0.007	0.418
19	8219.705	-0.069	-0.007	0.418
20	8219.705	-0.069	-0.007	0.418

6. Maksimum Likelihood Tahmin Edici İterasyonları-2

Ite.	-2LogL	OTUR_SUR	EFB_OECD	EGITIM(1)	EGITIM(2)	EGITIM(3)	EGITIM(4)
1	8340.699	-0.004	0.285	0.259	0.553	-1.562	0.669
2	8236.902	-0.007	0.392	0.165	0.527	-1.884	0.693
3	8235.073	-0.007	0.407	0.147	0.511	-1.921	0.683
4	8235.068	-0.007	0.407	0.147	0.511	-1.921	0.683
5	8235.068	-0.007	0.407	0.147	0.511	-1.921	0.683
6	8235.068	-0.007	0.407	0.147	0.511	-1.921	0.683
Ite.	-2LogL	EGITIM(5)	EGITIM(6)	EGITIM(7)	EGITIM(8)	EGITIM(9)	EGITIM(10)
1	8340.699	1.174	0.640	0.887	0.965	1.045	1.139
2	8236.902	1.703	0.636	1.056	1.206	1.349	1.532
3	8235.073	1.989	0.621	1.084	1.263	1.436	1.673
4	8235.068	2.040	0.621	1.085	1.266	1.441	1.685
5	8235.068	2.042	0.621	1.085	1.266	1.441	1.685
6	8235.068	2.042	0.621	1.085	1.266	1.441	1.685