T.C
DİCLE UNİVERSİTESİ
Fen Bilimleri Enstitüsü

LOJİSTİK REGRESYONDA BAZI YANLI KESTİRİCİLERİN İNCELENMESİ

Nurkut Nuray URGAN

DOKTORA TEZİ

(MATEMATİK ANABİLİM DALI)

DİYARBAKIR
EKİM-2007

T.C

DİCLE UNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ MÜDÜRLÜĞÜ

DİYARBAKIR

Nurkut Nuray URGAN tarafından yapılan "LOJİSTİK REGRESYONDA BAZI YANLI KESTİRİCİLERİN İNCELENMESİ" konulu bu çalışma , jürimiz tarafından Matematik Anabilim Dalında DOKTORA tezi olarak kabul edilmiştir

Jüri Üyesinin

Ünvanı Adı Soyadı

Başkan:Yrd.Doç.Dr. Pakize TAYLAN(Danışman)..........................

Üye : Prof.Dr. Müjgan TEZ..

Üye : Prof.Dr. Sezai OĞRAŞ..

Üye : Prof.Dr. Ali YILMAZ...

Üye : Yrd.Doç.Dr. Ahmet KAYA...

Tez Savunma Sınavı Tarihi: 19/10/2007

Yukarıdaki bilgilerin doğruluğunu onaylarım.

.../...../2007

Prof. Dr. Necmettin PİRİNÇÇİOĞLU

ENSTİTÜ MÜDÜRÜ

(MÜHÜR)

TEŞEKKÜR

Bu tez çalışmamda katkılarından dolayı tez danışmanım sayın

Yrd. Doç. Dr. Pakize TAYLAN' a,

her konuda bana yardımcı olan, ilgisini ve desteğini hiç esirgemeyen saygıdeğer hocam

Prof. Dr. Müjgan TEZ'e

teşekkürlerimi sunarım.

Bana her konuda emek, güven ve destek veren, sevgileri ve sabırları ile beni güçlendiren aileme ve en önemli yardımcım olan eşim Onur' a teşekkürlerimi sunarım.

İÇİNDEKİLER

AMAÇ

Çalışma alanlarının çoğunda yanıt değişken iki değerlidir. Yanıt iki değerli olduğunda lineer regresyon model uygun olmadığından lojistik regresyon model kullanılır. Lojistik regresyon model son yıllarda biyoloji, tıp, ekonomi, tarım ve taşıma sahalarında yaygın olarak kullanılmaktadır. Lojistik regresyon modelinde veriler iç ilişkili olduğu durumunda en çok olabilirlik kestirimi yetersiz kalır. Bu problemi yok etmek için lineer regresyonda kullanılan bazı kestiriciler lojistik regresyona da uygulanmıştır.

Bu çalışmadaki esas amacımız, lineer regresyonda kullanılan Liu kestiricisini lojistik regresyona uyarlamak ve daha önceden verilmiş olan bazı lojistik regresyon kestiricilerle karşılaştırmaktır.

ÖZET

Bu çalışma dört bölümden oluşmaktadır.

Birinci bölümde, lojistik regresyon modeli ve özellikleri verilmektedir. Ayrıca, lojistik regresyonun kullanıldığı çalışma alanları ve tarihsel gelişmeler ile lineer regresyona göre avantajları verilmektedir.

İkinci bölümde, lojistik regresyonda kullanılan yansız kestiriciler verilmiştir. Ayrıca, lojistik regresyonda verilerin iç ilişkili olması ve iç ilişkinin yansız kestiricilere etkisi verilmektedir. Bundan başka, veriler iç ilişkili olduğu durumlarda kullanılan bazı yanlı kestiriciler tanıtılmaktadır.

Üçüncü bölümde, lojistik regresyonda veriler iç ilişkili olduğunda kullanılan bazı yanlı kestiriciler alanında yapılmış olan bazı çalışmalar incelenerek bu kestiriciler ile ilgili temel bilgiler verilmiştir.

Son bölümde, lineer regresyonda kullanılan özel bir kestirim lojistik regresyona uyarlanmıştır. Son olarak, bu kestirici ile yanlı ve yansız kestiriciler karşılaştırılmıştır.

ABSTRACT

This study consists of four chapters.

In the first chapter, logistic regression model and its properties are given. Moreover, previous progresses and the study fields in logistic regression are given with its advantages in comparison linear regression.

In the second chapter, unbiased estimators in logistic regression are given. Moreover, collinearity and its effects to the unbiased estimation in logistic regression are given. Furthermore, some biased estimators are introduced when the data are collinear.

In the third chapter, some fundamental information are given about some biased estimators in logistic regression when the data are collinear by analyzing the previous works in this field.

In the final chapter, a special estimation in linear regression is adapted to the logistic regression. Finally, this estimator is compared with both biased and unbiased estimators.

BÖLÜM 1.

LOJİSTİK REGRESYON MODEL

1.1. GİRİŞ

Regresyon yöntemleri, bir tepki değişken ile bir veya daha fazla açıklayıcı değişken arasındaki ilişkileri inceleyen, her biri veri analizinin integral bileşenidir. Genel bir durum ise sonuç değişkeninin kesikli olup, iki veya daha fazla olası değer almasıdır. Lojistik regresyon model bu durumun analizini yapan standart yöntem haline gelmiştir. 1950' lerde genelleştirilmiş lineer model keşfedilmeden önce, lojistik regresyon biyoistatistik uygulamalarında kullanılmıştır.

Lojistik regresyon yöntemi kullanılarak yapılan analizin amacı, istatistikte kullanılan diğer model oluşturma yöntemleri ile aynıdır, yani, yanıt değişken ile açıklayıcı değişkenlerin kümesi arasındaki ilişkiyi inceleyen en uygun ve biyolojik olarak mümkün olan en ucuz modeli saptamaktır. Bu açıklayıcı değişkenler genellikle *"ortak değişken (covariate)"* olarak adlandırılır.

Lojistik regresyon modelini diğer modellerden ayıran özellik, lojistik regresyondaki tepki değişkeninin ikili veya iki değerli (dichotomous) olmasıdır. Lojistik ve lineer regresyon arasındaki bu fark hem parametrik model seçiminde hem de yaklaşımlarda gözlemlenir. Herhangi bir regresyon probleminde en önemli nicelik, verilen bir bağımsız değişken değeri için sonucun ortalama değeridir. Bu "koşullu ortalama" olarak adlandırılır. Y yanıt değişkeni ve x açıklayıcı değişken olmak üzere,

$E\left(Y|x\right)$ şeklinde gösterilir. Lineer regresyonda, bu ortalama x ' in bir denklemi olarak

$$E\left(Y|x\right) = \beta_0 + \beta_1 x$$

şeklinde yazılabilir. Bu ifadeden görüldüğü gibi, x' in $\left(-\infty, +\infty\right)$ aralığındaki değerlerine göre $E\left(Y|x\right)$ değer alır. Tablo 1.1 de 100 kişide yapılan incelemeler sonucu, AGE yaş, CHD koroner kalp hastalığının olması ("1") ve olmaması ("0") olmak üzere tablo oluşturulmuş [14].

AGE	n	CHD Olmayan(0)	CHD Olan(1)	Ortalama
20-29	10	9	1	0,10
30-34	15	13	2	0,13
35-39	12	9	3	0,25
40-44	15	10	5	0,33
45-49	13	7	6	0,46
50-54	8	3	5	0,63
55-59	17	4	13	0,76
60-69	10	2	8	0,80
Toplam	100	57	43	0.43

Tablo 1.1. CHD'ye göre yaş grubu frekans tablosu

İki değerli veriler ile ortalama, yukarıda da görüldüğü gibi 0 ile 1 arasındadır, yani $0 \le E\left(Y|x\right) \le 1$ dir. İki değerli sonuç değişkenlerinin analizinde kullanılabilen çok sayıda dağılım vardır, bu dağılımlardan lojistik dağılımı seçmek için başlıca iki sebep:

1) matematiksel açıdan çok esnek ve kolay kullanılabilir bir fonksiyon olması,

2) biyolojik olarak anlamlı yorumlara elverişli olmasıdır.

Lojistik dağılımında $E\left(Y|x\right)=\pi(x)$ olmak üzere, lojistik regresyon modeli

$$\pi(x)=\frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}} \tag{1.1}$$

şeklindedir. Lineer regresyon modeli esnek ve yorumu kolay olduğundan, lojistik regresyon modeli aşağıda verilen lojit dönüşümü kullanılarak lineer modele dönüştürülür.

$$g(x)=\ln\left[\frac{\pi(x)}{1-\pi(x)}\right]=\beta_0+\beta_1 x$$

Kolayca görüleceği gibi $g(x)$ lineerdir, x'e bağlı olarak ($-\infty,+\infty$) aralığında değer alır. Lineer ve lojistik model arasındaki bir başka önemli fark sonuç değişkeninin koşullu dağılım içermesidir. Lineer regresyon modelinde sonuç değişkeninin gözlemi $y=E(Y\mid x)+\varepsilon$ olarak yazılabilir. Buradaki ε hatadır ve koşullu ortalamadan sapan gözlemleri ifade eder. Genel varsayım, ε' nun 0 ortalamalı ve bağımsız değişkenlerin seviyeleri arasında sabit varyanslı bir normal dağılıma sahip olduğudur. Bu da x bilinirken sonuç değişkeninin koşullu dağılımının $E\left(Y|x\right)$ ortalamalı ve sabit varyanslı normal dağılım olması demektir. Fakat bu durum iki değerli sonuç değişkeni için geçerli değildir. Bu durumda x bilinirken sonuç değişkeni $y=\pi(x)+\varepsilon$ olarak ifade edilebilir. Burada ε iki olası değerden biridir; ya $y=1$ ise $\pi(x)$ olasılığıyla $1-\pi(x)$ dir, yada $y=0$ ise $1-\pi(x)$ olasılığıyla $\pi(x)$ dir. Böylece ε '0' ortalamalı ve $\pi(x)\left[1-\pi(x)\right]$ varyanslı dağılıma sahip olur.

Özet olarak, sonuç değişken iki değerli olduğunda regresyon analizinde:

- Regresyon denkleminin koşullu ortalaması 0 ile 1 arasında formüle edilmelidir. (1.1) ile verilen $\pi(x)$ lojistik regresyon modeli bu koşulu sağlar.
- Binom dağılım, ε nun dağılımını tanımlar ve analizin temeline dayanan istatistiksel dağılımdır.
- Lineer regresyon kullanarak yapılan analiz varsayımları lojistik regresyonda da geçerlidir.

1.2. LOJİSTİK REGRESYONUN LİNEER REGRESYON İLE İLİŞKİSİ

Model oluşturmanın en sık kullanılan yöntemi, sonuç değişkeni sürekli olan lineer regresyon modelidir.

Lojistik regresyon ile lineer regresyon arasındaki en belirgin fark; lojistik regresyonda sonuç değişkeninin ikili ya da çoklu olmasıdır. Aralarındaki bu fark hem parametrik model seçimine hem de varsayımlara yansımaktadır.

Lineer regresyon analizinde olduğu gibi, lojistik regresyon analizinde de bazı değişken değerleri göz önüne alınarak kestirim yapılmaya çalışılır. Fakat bu iki analiz arasında üç önemli fark vardır.

a) Lineer regresyonda kestirilecek bağımlı değişken sürekli ancak lojistik regresyonda bağımlı değişken kesikli değerler alır.

b) Lineer regresyon analizinde bağımsız değişkenin çok değişkenli normal dağılıma sahip olma koşulu aranırken lojistik regresyonda böyle bir koşul aranmaz.

c) Lineer regresyon analizinde bağımlı değişkenin değeri kestirilirken, lojistik regresyon analizinde bağımlı değişkenin alabileceği değerlerden birinin gerçekleşme olasılığı kestirilir.

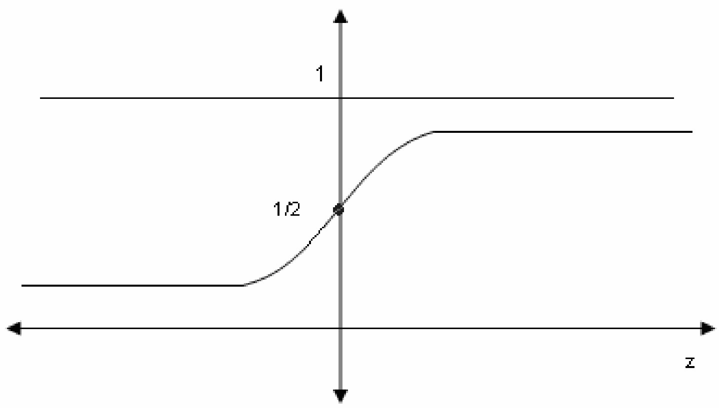
Dolayısıyla lojistik regresyona başlarken, lineer regresyonda kullanılan yöntemlerden yararlanacağız.

1.3. NEDEN LOJİSTİK REGRESYON DAHA İYİDİR?

Lojistik regresyonun alternatif olmasının en önemli özelliklerinden biri, lojistik regresyonda risk kestiriminin her zaman 0 ile 1 arasında olmasıdır. Diğer bir özellik ise yanıt iki değerlidir. Birinci özelliği göstermek için, lojistik modelin dayandığı matematiksel formu tanımlayan lojistik fonksiyon

$$f(z) = \frac{1}{1+e^{-z}} \quad , \quad -\infty < z < \infty$$

olarak alınsın. z, $(-\infty,\infty)$ arasında değer aldıkça bu fonksiyonun değerleri şekil 1' deki gibi olur.



Şekil 1.1: S-biçimi

Şeklin sağ tarafında z, $+\infty$ da değer aldıkça $f(z)$=1 , sol tarafında ise, z, -∞ da değer aldıkça $f(z)$=0 olduğu, yani 0< $f(z)$<1 olduğu görülüyor. $f(z)$' nin 0 ile 1 arasında değer alıyor olması lojistik modeli daha kullanışlı hale getirir.

Lojistik model, elde edilecek risk kestiriminin her zaman 0 ile 1 arasında olmasını garanti eder. Bu nedenle lojistik modelde hiçbir zaman risk kestirimi 1'in üstünde ve 0'ın altında elde edilmez. Bu diğer modellerde her zaman sağlanmadığından model seçimi yapılması durumunda ilk tercih lojistik model olmalıdır.

Lojistik modelin tercih edilmesinin bir başka sebebi; lojistik fonksiyonun şeklinden kaynaklanır. z, -∞' dan değer almaya başlandığında, $f(z)$ önce 0' a yakınsar sonra 1'e doğru artar. Bu sonuçların oluşturduğu şekle S-biçimi denir.

Lojistik fonksiyonun S-biçimi, z değişkeni birkaç risk faktörünün etkilerini kapsayan bir göstergeyi ve $f(z)$ de verilen bir z için riski temsil ettiğinden, birçok sağlık birimi için özellikle epidemiolojistler (salgın hastalık bilimcileri) için lojistik fonksiyonun S-biçiminin kullanımı çok caziptir.

$f(z)$' nin S-biçimi, bireysel risk üzerindeki z'nin etkisinin eşiğe (threshold) gelinene kadar küçük z'ler için en az olduğunu gösteriyor. Epidemiolojistler bu eşiği, hastalık durumlarının bir çeşidine uygulanacağını düşünmüştür. Başka deyişle, bir epidemiolojik araştırmanın çok değişkenli doğası ele alındığında S-biçim modeli geniş uygulama alanlarına sahiptir [17].

Özetle, lojistik model aşağıdaki nedenlerden dolayı daha kullanışlıdır, çünkü lojistik fonksiyonda:

- Kestirimler 0 ile 1 arasındadır.
- S-biçimi, bir hastalık için birkaç risk faktörünün etkilerini kapsar.

Lojistik modelin tercih edilmesinin bir başka önemli sebebi ise, çalışma alanlarının çoğunda yanıt değişken olarak iki değerli (dichotomous) değişkenlerin kullanılmasıdır. Örneğin; bir hastalığın olması durumunda Y=1, olmaması durumu için de Y=0 alındığında, bu değişken ile kan basıncı, yaş ve sigara içme alışkanlığı gibi açıklayıcı değişkenler arasında ilişki olup olmadığının bilinmesi gerekir. Normallik varsayımına dayanan lineer regresyon bu durum için uygun değildir.

$$y_i = \mu_i + \varepsilon_i \quad , \quad i = 1,2,...,n$$
$$\varepsilon_i \sim N(0,\sigma^2) \quad ve \quad \mu_i = \beta_0 + \beta_1 x_i$$

Lineer regresyon modeli göz önüne alındığında y_i ve ε_i iki değerli dağılıma sahip değişkenler olduğundan normallik varsayımı beklenmez. Varyansın homojenliği sadece $\mu_i = \mu_j$, $i, j = 1,...,n$ durumunda mevcuttur ki, bu ilgilenilen durum değildir.

1.4. LOJİSTİK REGRESYONUN KULLANIM ALANLARI VE TARİHSEL GELİŞİMİ

Lojistik regresyon modelleri, son yıllarda biyoloji, tıp, ekonomi, tarım, veterinerlik ve taşıma sahalarında yaygın olarak kullanılmaktadır. Lojistik modelin biyolojik deneylerin analizi için kullanımı ilk olarak Berkson (1944) tarafından önerilmiş, Cox (1970) bu modeli gözden geçirerek çeşitli uygulamalarını yapmış, özet gelişmeler ise ilk Anderson (1979, 1983) tarafından verilmiştir. Ayrıca verilerin lojistik modele uyumu ile ilgili birçok çalışmalar da yapılmıştır. Bunlar arasında Aranda-Ordaz (1981) ve Johnson (1985) tarafından yapılan çalışmalar en önemlileridir. Pregibon (1981) iki grup lojistik modelde etkin (influential), aykırı (outlier) gözlemleri ve belirleme ölçütlerini (diagnostic), Lesaffre (1986), Lesaffre ve Albert (1989) ise çoklu grup lojistik modellerde etkin ve aykırı gözlemlerle belirleme ölçütlerini incelemişlerdir. Lojistik regresyon modellerinin yaygın bir şekilde kullanılır hale gelmesi, hatayı kestirim yöntemlerinin geliştirilmesi ve lojistik regresyon modellerinin daha ayrıntılı incelenmesine sebep olmuştur. Cornfield (1962), lojistik regresyondaki katsayı kestirim işlemlerinde ayırıcı fonksiyon yaklaşımını ilk kez kullanarak popüler hale getirmiştir.

Halpern ve Blackwelden (1971)' de çok değişkenli lojistik fonksiyonda discriminant fonksiyon ve en çok olabilirlik kestiricisini karşılaştırmıştır. Aynı karşılaştırmayı bu kez Hosmer, D.W., Hosmer, T. ve Fisher(1983), sürekli ve kesikli değişkenler, üzerine yapmıştır. Hosmer, Wang, Lin ve Lemeshow (1978) lojistik regresyonda en çok olabilirlik kestirimi kullanarak bir program geliştirmiştir. Çoklu log-regresyon analizinde Lemeshow ve Hosmer odds oranların kestirimi üzerine çalışma yapmıştır. Epidemilolojik veriler kullanılarak Kleinbaum, Kupper ve Mongenstern (1982)' de lojistik regresyon analizi yapmış, Gren ve Symons (1983) ve Hauck (1985) lojistik fonksiyon ile orantılı tehlikeli (hazard) modeli karşılaştırmıştır. Ardından Abbott (1985) yaşam analizde log-regresyon yapmış, Bren ve Arnesen (1985) biraz daha geliştirerek risk faktörlerini seçim için lojistik regresyon ve cox regresyon modellerinin karşılaştırmasını yapmış, Albert ve Anderson (1984) ve Chambless ve Boyle (1985) lojistik modelde MLE üzerine çalışmalar yapmıştır.

Shaeffer (1986)' da veriler iç ilişkili olduğu durum için alternatif olarak Ridge, Stein ve Temel Bileşenler kestiricilerini inceleyerek, bunların En çok Olabilirlik kestiricisi ile karşılaştırılmasına ilişkin bir simülasyon çalışması yapmıştır. Aguilera, Escabias ve Valderamann (2006)' da çoklu doğrusal iç ilişkili durumunu için Temel Bileşenler kestiricisini geliştirerek bunun için bir simulasyon çalışması yapmıştır. Kestiriciler üzerine, Liang (1987) Mantel-Haenszel kestirim yöntemini uygulamış ve Hjont (1988) modelin yanlış olduğu durumlar için çalışmalar yapmıştır.

Duffy (1986) ve Duffy ve Santner (1989) lojistik regresyonda En Çok Olabilirlik kestiricisine alternatif bir kestirici önermişlerdir. Bu kestirici ceza terimi $\frac{\left\|\underline{\beta}-\underline{\mu}\right\|^2}{2\sigma^2}$ olan en çok cezalı olabilirlik kestiricisidir, burada μ , önsel dağılımın ortalaması ve σ^2 varyansıdır.

Nyquist (1991), Schaeffer ve Marx'ın Ridge kestiricisine benzer sınırlandırılmış bir kestirici önermiştir. Önerdiği bu kestiricinin Ridge kestiricisinden farkı, bu kestiricinin yinelenerek hesaplanıyor olmasıdır. Nyquist' in kestiricisi

$$\hat{\underline{\beta}}_r(k)=\left(X'\hat{V}X+kI\right)^{-1}X'\hat{Z} \quad , \quad \hat{\underline{Z}}=(\hat{\underline{z}},\ldots,\hat{\underline{z}}) \text{ ve } \hat{\underline{Z}}_i=\sum_{j=1}^{p+1}x_{ij}\beta_j+(y_i-\mu_i)\frac{\partial\eta_i}{\partial\mu_i}$$

şeklindedir. Schaffer ve Nyquist'in kestiricileri arasındaki fark, $\hat{V}$ matrisinin ve $\hat{z}_i$ vektörünün $\hat{\underline{\beta}}$'da değilde $\hat{\underline{\beta}}_r(k)$' da değerlendirilmesidir.

Le Cessie ve Von HouWeligen(1992), Duffy ve Santner (1989)'in bu yaklaşımından Ridge lojistik regresyon kestiricisinin Nygquist' inkine benzeyen farklı bir kullanımını önermiştir. Çalışmalarında özel dağılımı 0 ortalamalı ve k^{-1} varyanslı normal dağılım olarak ele almıştır.

Lee (1984) basit dönüşümlü (cross-over) deneme planları için lineer lojistik modeller üzerinde durmuştur. Bonney (1987) lojistik regresyon modelinin kullanımı ve geliştirilmesi üzerinde çalışmıştır. Robert (1987) lojistik regresyonda standart kikare, olabilirlik oran (G^2), "yalancı (pseudo)" en çok olabilirlik kestirimleri, uyum mükemmelliği ve hipotez testleri üzerine çalışmalar yapmışlardır. Duffy (1990) lojistik regresyonda hata terimlerinin dağılışı ve parametre değerlerinin gerçek değerlere yaklaşımını incelemiştir. Başarır (1990) klinik verilerde çok değişkenli lojistik regresyon analizi ve ayrımsama sorunu üzerinde çalışmıştır. Hsu ve Leonard (1995) lojistik regresyon fonksiyonlarında Bayes kestirimlerinin elde edilmesi işlemleri üzerine çalışmışlar ve lojistik regresyonda Monte Carlo dönüşümünün kullanılabileceğini göstermişlerdir. Akkaya ve Pazarlıoğlu (1998) lojistik regresyon modellerinin ekonomi alanında kullanımını örneklerle incelemişlerdir. Cox (1998) kardiovasküler hastalıklar ve hipertansiyon arasındaki ilişkiyi incelemişlerdir. Gardside ve Glueck (1995) insanlarda beslenme şekli, sigara ve alkol kullanımı, fiziksel aktivite gibi risk faktörlerinin kalp hastalığı üzerindeki etkilerini incelemiştir [35].

Kloiber, Winn, Shaffer ve Hassanein (1996), Buescher, Larson, Nelson ve Lenihan (1993) kadınlarda düşük doğum ağırlığını etkileyen risk faktörlerini; Santos (1998) kafein tüketimi ve düşük doğum ağırlığı arasındaki ilişkiyi, Sable ve Herman (1997) erken doğum ve düşük doğum ağırlığı arasındaki ilişkiyi incelemişlerdir.

BÖLÜM 2.

LOJİSTİK REGRESYONDA PARAMETRELERİN KESTİRİLMESİ

2.1. ALIŞILMIŞ PARAMETRE KESTİRİCİLERİ

2.1.1. En Çok Olabilirlik Kestirimi (MLE)

Lojistik regresyon, kanonik bağ fonksiyonu kullanılan ikili veya binom dağılımlı değişken modelidir. Veriler gruplanmış olsun, i-inci veri noktasında n_i deneme birimi ve $x_i\underline{\beta} = \beta_0 + \beta_1 x_{i1} + ... + \beta_k x_{ik}$ olmak üzere (1.1) modeli,

$$\begin{aligned} E(y_i) &= n_i\, \pi(x_i) \\ &= n_i \frac{1}{1+e^{-x_i'\underline{\beta}}} \qquad , \quad i = 1, 2, ..., m \end{aligned}$$

olarak yazılır. $y_1, y_2, ..., y_m$, bağımsız binom rastgele değişkenlerinin gözlem değerleridir.

$$Var(y_i) = n_i\, \pi(x_i)\left[1 - \pi(x_i)\right]$$

$$\sum_{i=1}^{m} n_i = \text{n}$$

n ve π için olasılık fonksiyon; $\binom{n}{y}\pi^{y}(1-\pi)^{n-y}$ dir. Log-olabilirlik belirlendiğinde $\binom{n}{y}$, $\underline{\beta}$ parametresi içermediğinden atılır ve böylece lojistik regresyon modeli için Log-olabilirlik fonksiyonu

$$\ln\left[\boldsymbol{\mathcal{L}}(\underline{\pi};\underline{y})\right]=\sum_{i=1}^{m}\left\{y_i\ln\left[\frac{\pi(x_i)}{1-\pi(x_i)}\right]+n_i\ln\left[1-\pi(x_i)\right]\right\} \tag{2.1}$$

şeklinde elde edilir. Burada $\ln\left[\frac{\pi(x_i)}{1-\pi(x_i)}\right]$, lojit olarak adlandırılır ve

$$\ln\left[\frac{\pi(x_i)}{1-\pi(x_i)}\right]=x_i'\underline{\beta}$$
$$=\beta_0+\sum_{j=1}^{k}x_{ij}\beta_j \quad , \quad i=1,2,\ldots,m \quad , \quad m\geq k+1$$

şeklinde elde edilir. Sonuç olarak, (2.1) denklemi

$$\ln\left[\boldsymbol{\mathcal{L}}(\underline{\beta};\underline{y})\right]=\sum_{i=1}^{m}\sum_{j=1}^{k}y_i x_{ij}\beta_j-\sum_{i=1}^{m}n_i\ln\left(1+\exp\sum_{j=1}^{k}x_{ij}\beta_j\right) \tag{2.2}$$

olarak yazılabilir. (2.2) denkleminin β_j'ye göre en büyük değerini bulmak için, (2.2) denkleminin matris formu,

$$\ln\left[\boldsymbol{\mathcal{L}}(\underline{\beta};\underline{y})\right]=\underline{\beta}'X\underline{y}-\sum_{i=1}^{m}n_i\ln\left(1+\exp(x_i'\underline{\beta})\right) \tag{2.3}$$

şeklinde oluşturulur. (2.3) denkleminin $\underline{\beta}$' ya göre türevinin alınması ile

$$\frac{\partial\ln\left[\boldsymbol{\mathcal{L}}(\underline{\beta};\underline{y})\right]}{\partial\underline{\beta}}=X'\underline{y}-\sum_{i=1}^{m}\left[\frac{n_i}{1+e^{x_i'\beta}}\right]e^{x_i'\beta}x_i$$

elde edilir.

$$\frac{e^{x_i'\beta}}{(1+e^{x_i'\beta})}=\frac{1}{(1+e^{-x_i'\beta})}=\pi(x_i)$$

olduğundan,

$$\frac{\partial \ln\left[\mathcal{L}(\underline{\beta};\underline{y})\right]}{\partial \underline{\beta}}=X'\underline{y}-\sum_{i=1}^{m} n_i\pi(x_i)x_i$$

şeklinde yazılır. Binom rastgele değişkeninin ortalaması $n_i\pi(x_i)$ olduğundan, sağ taraf matris gösterimiyle $X'(\underline{y}-\underline{\mu})$ olarak yazılır, burada $\underline{\mu}$,

$$\underline{\mu}=\begin{bmatrix}\mu_1\\ \mu_2\\ \vdots\\ \mu_m\end{bmatrix} \qquad ve \qquad \mu_i=n_i\pi(x_i)$$

şeklindedir. Sonuç olarak en çok olabilirlik kestirimi

$$X'(\underline{y}-\underline{\mu})=0 \tag{2.4}$$

"skor denkleminin" çözümü ile bulunur. n bağımsız gözlem, p ortak değişken ve η_i bağ fonksiyonuna sahip lojistik model göz önüne alındığında, bu model için skor fonksiyonunun j-inci elemanı

$$U(\beta_j)=\frac{\partial l(\underline{\beta})}{\partial \beta_j}=\sum_{i=1}^{n}\frac{y_i-\pi_i}{\pi_i\left(1-\pi_i\right)}\frac{\partial \pi_i}{\partial \eta_i}x_{ij}$$

olarak yazılır. Ayrıca bu model için skor vektörü ise

$$S(\underline{\hat{\beta}})=\left[U(\hat{\beta}_0)\quad U(\hat{\beta}_1)\quad \dots\quad U(\hat{\beta}_p)\right]'=X'(\underline{y}-\underline{\hat{\pi}})$$

olarak elde edilir. Burada $X=\left(\underline{x}_1',\underline{x}_2',...,\underline{x}_n'\right)$ tasarım matrisi, $\underline{y}=\left(y_1,y_2,...,y_n\right)$ gözlem vektörü ve $\hat{\underline{\pi}}=\left(\hat{\pi}_1,\hat{\pi}_2,...,\hat{\pi}_n\right)$, $\underline{\pi}'=\left(\pi_1,\pi_2,...,\pi_n\right)$ olasılık vektörünün en çok olabilirlik kestirimidir. $v_i=\pi_i(1-\pi_i)=Var(y_i)$ ve $V=diag(v_i)$ olmak üzere bilgi matrisi

$$i(\underline{\beta})=X'VX$$

şeklindedir.

Schaefer (1979) ve Bradly ve Gart (1962) çalışmalarında asimptotik sonuçların aşağıdaki koşulları sağladığı üzerinde durmuştur. Bu koşullar:

i) $\left|x_{ij}\right|$, tüm i ve j ler için sınırlıdır.

ii) $\lim\limits_{n\to\infty}\dfrac{i(\underline{\beta})}{n} \to Q$, (Q sonlu determinanta sahip pozitif tanımlı matris.)

Eğer bazı değişkenler birinci koşulu sağlamazsa, bu değişkenlerin sınırları keyfi olarak genişletilebilir. Oluşturulan kesilmiş (*truncated)* rasgele değişkenler orijinal değişkenlerle aynı asimptotik özelliklere sahiptir [9].

İkinci koşul ise, *x*' lerin dağılımının sonlu ikinci momente sahip olmasına denktir. Bu varsayım, asimptotik dağılımın kovaryans matrisinin iyi tanımlı ve uygun yapıya sahip olmasını garanti eder. Büyük örneklem için $\underline{\beta}$' nın en çok olabilirlik kestirimi:

i) $E\left(\underline{\beta}-\hat{\underline{\beta}}\right)=\underline{0}$

ii) $Var\left(\hat{\underline{\beta}}\right)=\left(X'VX\right)^{-1}$

iii) $MSE(\hat{\underline{\beta}})=\dot{I}z\left(Var\left(\hat{\underline{\beta}}\right)\right)+\left\|\left(\hat{\underline{\beta}}'\,nın\;yanı\right)\right\|^2\approx\sum\lambda_i$, λ_i, $i(\underline{\beta})$'nın özdeğerleri özelliklerini sağlar [33]. En çok olabilirlik kestirimi yansız olmasına rağmen, küçük örneklemler için genellikle yanlı çıkar. En çok olabilirlik kestiriminin yanının bir yaklaşımı için Taylor açılımı kullanılarak, $S(\underline{\beta})$ skor vektörünün kestirimi,

$$0 = S(\hat{\underline{\beta}}) = \begin{vmatrix} S(\beta_0) \\ \vdots \\ S(\beta_p) \end{vmatrix} + \begin{vmatrix} \left(\hat{\underline{\beta}} - \underline{\beta}\right)' S'(\beta_0) \\ \vdots \\ \left(\hat{\underline{\beta}} - \underline{\beta}\right)' S'(\beta_p) \end{vmatrix} - 0,5 \begin{vmatrix} \left(\hat{\underline{\beta}} - \underline{\beta}\right)' S''(\beta_0)\left(\hat{\underline{\beta}} - \underline{\beta}\right) \\ \vdots \\ \left(\hat{\underline{\beta}} - \underline{\beta}\right)' S''(\beta_p)\left(\hat{\underline{\beta}} - \underline{\beta}\right) \end{vmatrix} \tag{2.5}$$

olarak bulunur. $S'(\beta_i) = -X'VX$ (Y'ye bağlı olmayan bir sabit) ve $E\left(S(\beta_i)\right) = 0$ eşitlikleri kullanılarak, (2.5) denleminin her iki tarafının beklenen değerinin alınmasıyla

$$0 = \begin{vmatrix} S'(\beta_0) E\left[\left(\hat{\underline{\beta}} - \underline{\beta}\right)\right] \\ \vdots \\ S'(\beta_p) E\left[\left(\hat{\underline{\beta}} - \underline{\beta}\right)\right] \end{vmatrix} - 0,5 \begin{vmatrix} E\left[\left(\hat{\underline{\beta}} - \underline{\beta}\right)' S''(\beta_0)\left(\hat{\underline{\beta}} - \underline{\beta}\right)\right] \\ \vdots \\ E\left[\left(\hat{\underline{\beta}} - \underline{\beta}\right)' S''(\beta_p)\left(\hat{\underline{\beta}} - \underline{\beta}\right)\right] \end{vmatrix} \tag{2.6}$$

elde edilir. Ayrıca

$$\begin{vmatrix} S'(\beta_0) E\left[\left(\hat{\underline{\beta}} - \underline{\beta}\right)\right] \\ \vdots \\ S'(\beta_p) E\left[\left(\hat{\underline{\beta}} - \underline{\beta}\right)\right] \end{vmatrix} = -\left(X'VX\right) E\left[\left(\hat{\underline{\beta}} - \underline{\beta}\right)\right]$$

ve

$$E\left[\left(\hat{\underline{\beta}} - \underline{\beta}\right)' S''(\beta_i)\left(\hat{\underline{\beta}} - \underline{\beta}\right)\right] = S''(\beta_i)\, \dot{I}z \left\{E\left[\left(\hat{\underline{\beta}} - \underline{\beta}\right)' \left(\hat{\underline{\beta}} - \underline{\beta}\right)\right]\right\}$$
$$= \dot{I}z \left\{S''(\beta_i) Var(\hat{\underline{\beta}})\right\} \approx \dot{I}z \left\{S''(\beta_i)\left(X'VX\right)^{-1}\right\}$$

olduğundan (2.6) denklemi

$$0=-\left(X'VX\right)E\left[\left(\underline{\hat{\beta}}-\underline{\beta}\right)\right]-0,5\begin{vmatrix} \dot{I}z\ \left\{S''(\beta_0)\left(X'VX\right)^{-1}\right\} \\ \vdots \\ \dot{I}z\ \left\{S''(\beta_p)\left(X'VX\right)^{-1}\right\}\end{vmatrix}$$

olarak elde edilir. Bu durumda en çok olabilirlik kestiricisinin yanlılığı,

$$E\left[(\underline{\hat{\beta}}-\underline{\beta})\right]=-0,5\left(X'VX\right)^{-1}\begin{vmatrix} \dot{I}z\ \left\{S''(\beta_0)\left(X'VX\right)^{-1}\right\} \\ \vdots \\ \dot{I}z\ \left\{S''(\beta_p)\left(X'VX\right)^{-1}\right\}\end{vmatrix} \tag{2.7}$$

elde edilir. b, Taylor serisinin ilk teriminin kullanılmasıyla hesaplanan yaklaşık yan olmak üzere, $\underline{\hat{\beta}}_T=\underline{\hat{\beta}}-b$ eşitliğine karşılık gelen yukarıdaki yaklaşıma dayalı kestirici Anderson ve Richardson (1979), Schaefer (1983), Copas (1988) ve Cordeiro ve McCoullagh (1991) tarafından kullanılmıştır. $\underline{\beta}=\underline{\hat{\beta}}$ da elde edilen, $i(\underline{\hat{\beta}})$ bilgi matrisi ve $S''(\beta_i)$, skor vektörünün i-inci elemanının ikinci türevinin oluşturduğu matris olmak üzere,

$$b=-0,5\ i(\underline{\hat{\beta}})^{-1}\begin{vmatrix} \dot{I}z\ \left\{S''(\beta_0)\,i(\underline{\hat{\beta}})^{-1}\right\} \\ \vdots \\ \dot{I}z\ \left\{S''(\beta_p)\,i(\underline{\hat{\beta}})^{-1}\right\}\end{vmatrix}$$

eşitliği elde edilir. Buradan b yanlılığı,

$$b=-0,5\ i(\underline{\hat{\beta}})^{-1}\sum_{i=1}^{n}\underline{x}_i(0,5-\hat{\pi})\hat{\pi}(1-\hat{\pi})\underline{x}_i{}'\,i(\underline{\hat{\beta}})^{-1}$$

denkleminin hesaplanmasıyla bulunur.

2.1.2. Ağırlıklandırılmış En Küçük Kareler Kestirimi (WLS)

Lojistik regresyon için (2.4) denklemindeki skor fonksiyonunu ele alalım. (2.4) denklemi ve WLS arasında bir bağlantı vardır, bu bağlantının bir taslağı aşağıda verilmektedir. Örneğin ağırlıklandırılmış kalan karelerin toplamı,

$$S=\sum_{i=1}^{m}\left[\frac{(y_i-\mu_i)^2}{\sigma_i^2}\right] \tag{2.8}$$

olarak verildiğinde, $\mu_i = n_i\pi(x_i)$ dir ve $\sigma_i^2 = n_i\pi(x_i)\left[1-\pi(x_i)\right] = n_i\dfrac{e^{-x_i'\beta}}{\left(1+e^{-x_i'\beta}\right)^2}$ olacak şekilde i-inci veri noktasının binom varyansıdır. Aynı şekilde sabit σ_i^2 varyanslı

$$\min S = \min_{\beta}\sum_{i=1}^{m}\left[\frac{\left(y_i-\mu_i\right)^2}{\sigma_i^2}\right]$$

elde edilir. S'nin türevinin alınması ile

$$2\left[\frac{\sum_{i=1}^{m}\left(y_i-\mu_i\right)}{\sigma_i^2}\right]\left(\frac{\partial\mu_i}{\partial\beta}\right)$$

ifadesi elde edilir. Burada

$$\mu_i = n_i\pi(x_i) = n_i\frac{1}{1+e^{-x_i'\beta}} \quad\Rightarrow\quad \frac{\partial\mu_i}{\partial\beta} = n_i\pi(x_i)\left[1-\pi(x_{i\partial})\right]x_i = \sigma_i^2 x_i$$

olduğundan, σ_i^2 ile ağırlıklandırılmış kalan kareleri toplamının en küçük yapılmasıyla,

$$\sum_{i=1}^{m}(y_i-\mu_i)x_i = 0$$

denklemi elde edilir ve bu (2.4) denklemindeki $X'(y-\mu)$ =0 skor denklemine denktir. Sonuç olarak, tekrarlı yeniden ağırlıklandırılmış en küçük kareler yöntemi "skor denkleminin" çözümünü üretmek için kullanılır ve MLE'nin $b_0, b_1, ..., b_k$ sayısal değerleri elde edilir [26].

2.1.3. Tekrarlı Ağırlıklandırılmış En Küçük Kareler Kestirimi (IWLS)

X bağımsız değişkenlerin matrisi ve $\tilde{x}_i'$ bu matrisin i-inci sütunu olmak üzere (1.2) modeli

$$\begin{aligned}\pi_i &= 1\Big/\left(1+\exp\left[-\tilde{x}_i'\beta\right]\right)\\ &= f\left(\tilde{x}_i',\beta\right)\end{aligned}$$

dır. Burada β, β_j ' lerin $\left((p+1)\times 1\right)$ vektörü, y, $(n\times 1)$ vektör ve $\hat{\pi}$, $\hat{\pi}_i = f\left(\tilde{x}_i',\hat{\beta}\right)$ nin $(n\times 1)$ vektörüdür. β ' nın kestirimi tekrarlı ağırlıklandırılmış en küçük kareler (IWLS) tekniği ile de bulunabilir. ML veya IWLS kestiriminde, β 'nın $(l+1)$ -inci tekrarı

$$\hat{\beta}_{l+1} = \hat{\beta}_l + (X'\hat{V}_l X)^{-1} X'(y-\hat{\pi}_l) \tag{2.9}$$

şeklindedir. $\hat{\pi}_l$, $\hat{\beta}_l$ ve $\hat{V}_l = diag\left\{\hat{\pi}_{li}(1-\hat{\pi}_{li})\right\}$ 'nin kullanılmasıyla elde edilen π 'nin kestirim vektörüdür ve $\hat{\pi}_{li}$, $\hat{\pi}_l$ 'nin i-inci elemanıdır. Bu tekrarlar yakınsama elde edinceye kadar devam eder. Sonuç kestirim $\hat{\beta}$ ile gösterilir ve çok değişkenli normal dağılıma yaklaşması beklenir.

2.2. DOĞRUSAL İÇ İLİŞKİ

Regresyon modelleri, çeşitli uygulamalar için kullanılır. Bir regresyon modelinin kullanılışlığını çarpıcı bir biçimde etkileyen ciddi bir problem doğrusal iç ilişki veya açıklayıcı değişkenler arasındaki doğrusal bağlılıktır. Tasarım matrisi tekil ise açıklayıcı değişkenler arasında iç ilişki vardır.

$\hat{\mu} = \hat{E}\left[Y_i \mid \underline{x}_i\right]$ maksimum veya minimum değer alması ve en çok olabilirlik kestiricisinin parametre uzayının sınırlarına yaklaşması durumunda da başka çeşit iç ilişki vardır. Bu iç ilişkiye de ML-iç ilişki denir [18]. Her iki iç ilişki aynı anda bulunabilir.

Bu iki farklı iç ilişki formu göstermek için,

$$diag(v_1\, v_2) = \begin{bmatrix} 0,25 & 0 \\ 0 & \alpha \end{bmatrix}, \;\; X = \left[\underline{x}_1\, \underline{x}_2\right] = \begin{bmatrix} 1 & 1 \\ 2 & \lambda \end{bmatrix}$$

olarak ele alınsın. Buradan bilgi matrisinin kestirimi,

$$X'\hat{V}X = \begin{bmatrix} 0,25+4\alpha & 0,25+2\alpha\lambda \\ 0,25+2\alpha\lambda & 0,25+\alpha\lambda^2 \end{bmatrix}$$

olarak elde edilir. $\hat{V}$ matrisinin tekil olması ancak ve ancak $\alpha = 0$ olması ile mümkündür, bu durumda da bilgi matrisinin kestirimi

$$X'\hat{V}X = \begin{bmatrix} 0,25 & 0,25 \\ 0,25 & 0,25 \end{bmatrix}$$

dır. $X'\hat{V}X$ matrisi tüm λ' lar için tekildir. Fakat $v_2 = \alpha = 0$ olması ancak ve ancak $\hat{\mu}_2 = \hat{E}\left[Y_i = 1 \mid \underline{x}_i\right] = \mu_{\max}\ veya\ \mu_{\min}$ olması ile mümkündür. Lojistik regresyonda ise $v_2 = \alpha = 0$ olması ancak ve ancak $\hat{\mu}_2 = \hat{E}\left[Y_i = 1 \mid \underline{x}_i\right] = 1\ veya\ 0$ olması ile mümkündür. Buda $\hat{\beta}$ nın artı sonsuz veya eksi sonsuzlukta değer almasıyla gerçekleşir. Bu durumda

ML-iç ilişki kesinlikle vardır ve tasarım matrisi tekil olmadığı halde bilgi matrisi tekildir.

X tasarım matrisinin tekilliği λ' nın değerine bağlıdır. Eğer λ=2 ise X matrisinin sütunları eşit ve tekil olur. Bu durumda bilgi matrisinin kestirimi,

$$X'\hat{V}X = \begin{bmatrix} 0,25+4\alpha & 0,25+2\alpha\lambda \\ 0,25+2\alpha\lambda & 0,25+\alpha\lambda^2 \end{bmatrix} = \begin{bmatrix} 0,25+4\alpha & 0,25+4\alpha \\ 0,25+4\alpha & 0,25+4\alpha \end{bmatrix}$$

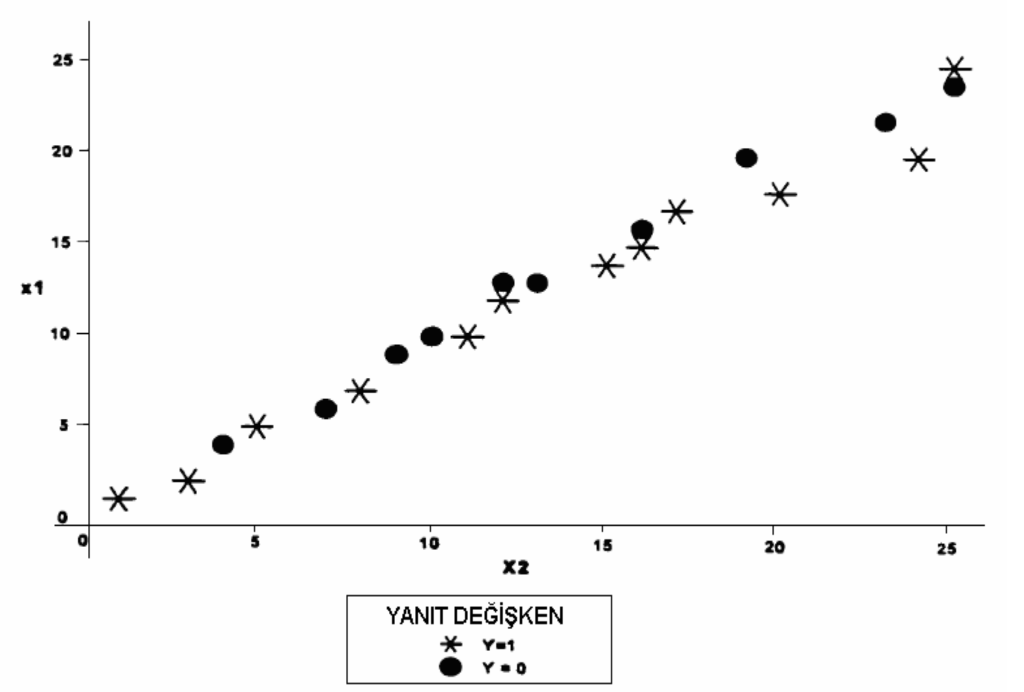
olarak bulunur. $X'\hat{V}X$ matrisi tüm α' lar için tekildir. Bu durumda da açıklayıcı değişkenler arasında iç ilişki vardır.

$X'\hat{V}X$ matrisi kötü-koşullu olduğunda; X matrisi tekil ise $(\lambda \to 2)$ açıklayıcı değişkenler arasında iç ilişki vardır, $\hat{V}$ matrisi tekil ise $(\alpha \to 0)$ ML-iç ilişki vardır.

Lesaffre ve Marx (1993) çalışmalarında biri açıklayıcı değişkenler arasında iç ilişki olan, diğeri ise açıklayıcı değişkenler arasında iç ilişki olmayan ama ML-iç ilişkili iki yapay örnek vermişler, bunlar sırasıyla Şekil 2.1 ve Şekil 2.2 de gösterilmiştir.

MODEL	$\hat{\beta}$	s($\hat{\beta}$)	p-değeri
X1	0.46	0.20	0.02
X2	0.45	0.19	0.02
X1 X2	2.75 -2.75	2.90 2.80	0.34 0.42

Tablo 2.1: İki açıklayıcı değişkenli lojistik regresyonda iç ilişki tablosu
(Lesaffre & Marx, 1993)

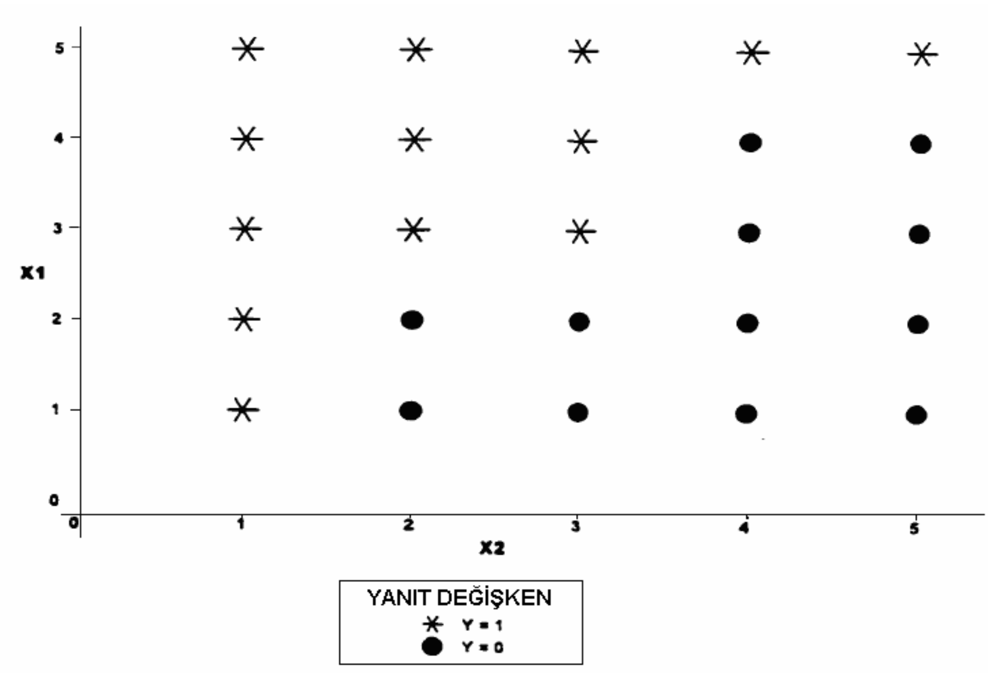


Şekil 2.1: İki açıklayıcı değişkenli lojistik regresyonda iç ilişki grafiği

MODEL	$\hat{\beta}$	$s(\hat{\beta})$	p-değeri
X1	-1.02	0.41	0.01
X2	0.99	0.4	0.01
X1 X2	-5.02 5.12	4.17 4.30	.23 .23

Tablo 2.2: İki açıklayıcı değişkenli lojistik regresyonda ML-iç ilişki tablosu

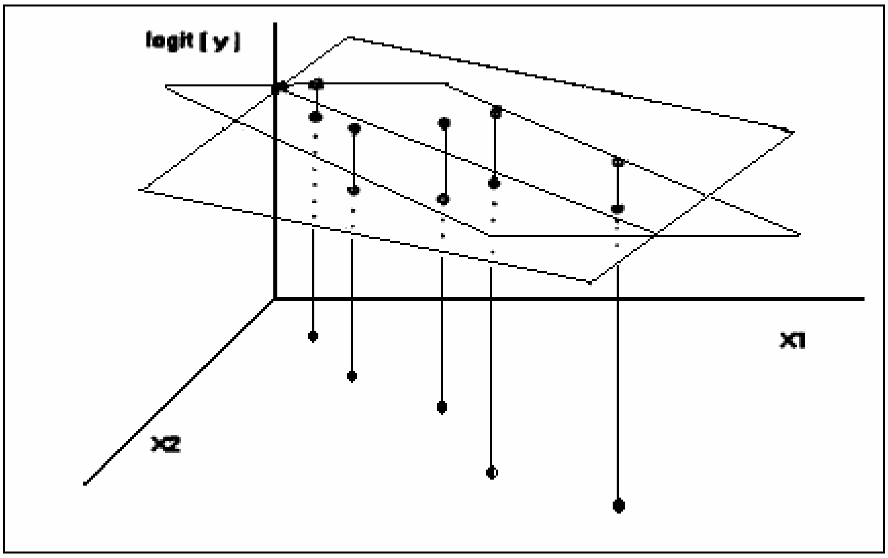
(Lesaffre & Marx, 1993)



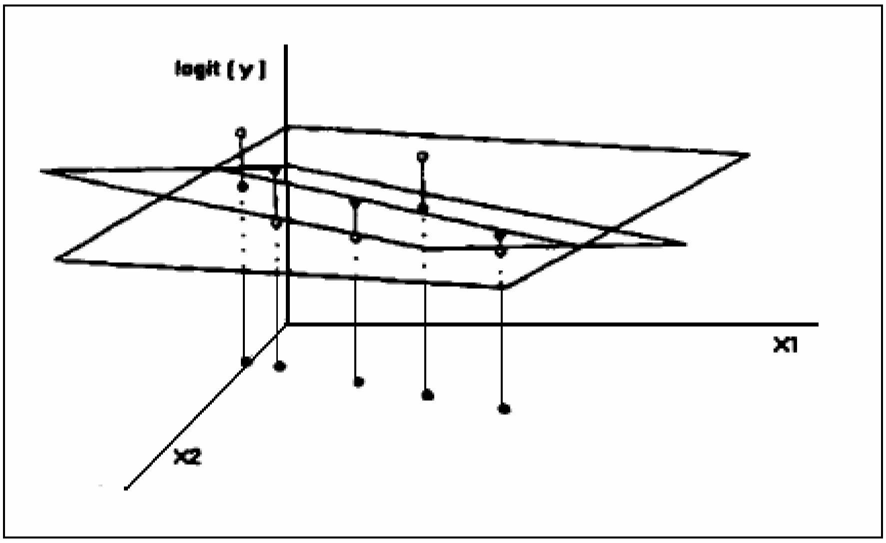
Şekil 2.2: İki açıklayıcı değişkenli lojistik regresyonda ML-iç ilişki grafiği.

Yukarıda verilen iki değişkenli örnekler için Tablo 2.1 ve Tablo 2.2 de bulunan her açıklayıcı değişkenin tek tek ele alınıp diğeri ile karşılaştırılmasıyla iç ilişkinin ve ML-iç ilişkinin etkisi anlaşılmaktadır. Bu değişkenler beraber ele alındığında varyansların ve parametrelerin kestirim değerlerinin arttığı görülebilir.

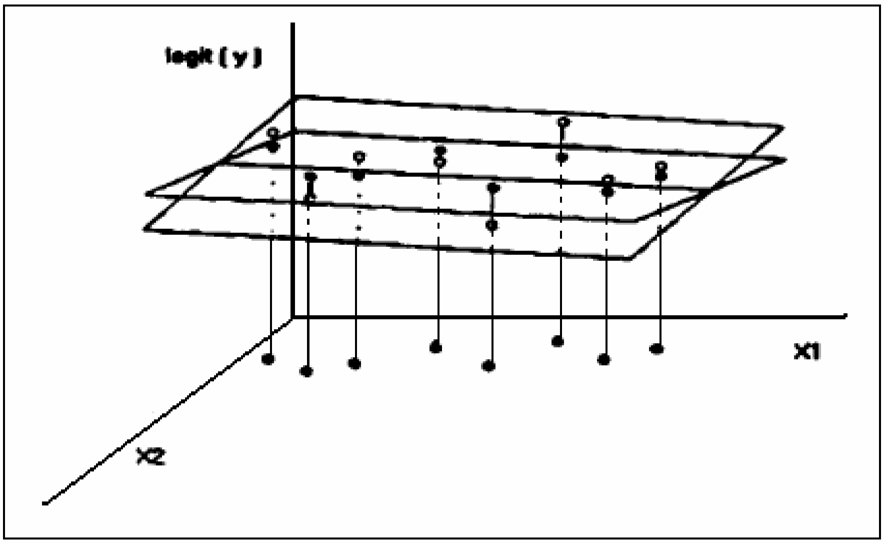
Bu iki tip iç ilişki arasında büyük fark vardır. Açıklayıcı değişkenler arasındaki iç ilişki lineer modellerdeki gibidir. Bireysel parametrelerin değerlerine ya da işaretlerine ilişkin yeterli bilgi olmamakla birlikte bazı parametrelerin lineer kombinasyonları ile ilgili bilgi vardır. Bu doğrusal kombinasyon tasarım matrisinin birinci bileşenidir, bu da Şekil 2.3a, 3b ve 3c de gösterilmiştir. Şekil 2.3a da kesen (intercept) parametresi ile ilgili iyi bilgi olduğu fakat diğer parametrelerle ilgili bilgi olmadığı görülür. Kesen parametresindeki değişim olabilirlik fonksiyonunda büyük değişikliklere yol açar. Benzer olarak Şekil 2.3b’de de parametrelerden sadece biri ile ilgili iyi bilgi olduğu görülür. Şekil 2.3c de ise parametrelerin hiçbiri ile ilgili bilgi elde edilememektedir.



a)



b)



c)

Şekil 2.3: Açıklayıcı değişkenler arasındaki güçlü iç ilişki

Lineer regresyonda iç ilişkiyi belirlemek için birçok yol vardır. Bunlardan bazıları aşagıdaki gibi verilebilir:

i) *Açıklayıcı değişkenler arasındaki basit korelasyon*: X ' in i-inci sütunu $\underline{x}_i$ ve j-inci sütunu $\underline{x}_j$ $(i \neq j;\ i, j = 1, 2, ..., p)$ arasındaki korelasyon olan r_{ij} 0,7 den büyükse iç ilişki problemi vardır.

ii) R_i^2 *çoklu korelasyon katsayısı*: Herhangi bir R_i^2 $(i = 1, 2, ..., p)$ 1'e yakınsarsa iç ilişki problemi vardır.

iii) VIF_i *Varyans şişirme faktörü*: Varyans şişirme faktörü: $VIF_i = \left(1 - R_i^2\right)^{-1}$, $i = 1, 2, ..., p$ dir. VIF_i>4 olması R_i^2>0,75 olduğu ve VIF_i>10 olması R_i^2>0,9 olduğu anlamına gelir. Eğer VIF_i>10 ise iç ilişki vardır ve alternatif yanlı kestiriciler kullanılır (Snee,1983).

iv) $X'X$ *matrisinin özdeğeri*: $X'X$ matrisinin en küçük özdeğeri 0' a yakınsarsa iç ilişki problemi vardır.

v) $X'X$ *matrisinin koşul sayısı*: λ_1, $X'X$ matrisinin en büyük (maksimal) özdeğeri ve λ_p, $X'X$ matrisinin en küçük (minimal) özdeğeri olmak üzere k_x koşul sayısı: $k_x = \dfrac{\lambda_1}{\lambda_p}$ şeklinde tanımlanır. Buna göre eğer:

k_x<100 ise iç ilişki yoktur,

100<k_x<900 ise iç ilişki problemi olabilir,

k_x>900 ise güçlü iç ilişki problemi vardır.

vi) C_i *koşul indeksi*: λ_i, $X'X$ matrisinin i-inci özdeğeri olmak üzere i-inci koşul indeksi: $C_i = \dfrac{\lambda_1}{\lambda_i}$ şeklinde tanımlanır. C_i koşul sayısı ölçütleri ile aynı özelliklere sahiptir.

vii) $X'X$ *matrisinin determinantı*: $X'X$ matrisinin determinantı 0' a yakınsarsa iç ilişki problemi vardır.

viii) T_i *tolerans faktörü*: Tolerans faktörü iç ilişkiyi belirleyerek, değişkenleri modele yerleştirmekte kullanılan bir ölçüttür. Tolerans faktörü: $T_i = 1/VIF_i$, $i = 1, 2, ..., p$ şeklindedir. Bir değişken ile modelde bulunan diğer değişkenler arasındaki korelasyon, tolerans faktörünü geçerse o değişken modelden kaldırılır.

ix) *Varyansların oranı*: $\lambda^*_1 > \lambda^*_2 > ... > \lambda^*_{p+1}$ tasarım matrisinin birim uzunluğa standartlaştırılmasıyla oluşan $X^{*\prime}X^*$ bilgi matrisinin özdeğerleri ve $\underline{\gamma}^*_i = (\gamma^*_{i1}, ..., \gamma^*_{i(p+1)})$ $X^{*\prime}X^*$ bilgi matrisinin i-inci özvektörü olamak üzere, i-inci bileşenin varyansı,

$$Var(\hat{\beta}_i) = \sum_{j=1}^{p+1} \frac{\left(\gamma^*_{ij}\right)^2}{\lambda^*_i}$$

dır. Buradan i-inci bileşenin varyans oranı $prop_{ij}$ ile gösterilmek üzere,

$$prop_{ij} = \frac{\dfrac{\left(\gamma^*_{ij}\right)^2}{\lambda^*_i}}{\sum_{k=1}^{p+1} \dfrac{\left(\gamma^*_{kj}\right)^2}{\lambda^*_k}}$$

dir. Böylece varyans oranlarının bir matrisi Tablo 2.3 deki gibi oluşturulabilir. Küçük bir özdeğerin olması en az iki büyük varyans oranının varlığını gösterir, bu durumda da iç ilişki problemi vardır [24;34].

Sıralı özdeğerler	$Var(\hat{\beta}_1)$	$Var(\hat{\beta}_2)$	. . .	$Var(\hat{\beta}_{p+1})$
λ_1^*	$prop_{11}$	$prop_{12}$	. . .	$prop_{1(p+1)}$
λ_2^*	$prop_{21}$	$prop_{22}$	. . .	$prop_{2(p+1)}$
. . .	. . .	. . .	. . .	. . .
λ_{p+1}^*	$prop_{(p+1)1}$	$prop_{(p+1)2}$	. . .	$prop_{(p+1)(p+1)}$

Tablo 2.3: Standartlaştırılmış bilgi matrisinin i-inci bileşenlerinin varyans oranları. (Marx&Smith, 1990b)

Lineer regresyon için geçerli olan bu ölçütler lojistik regresyona uygulanabilir.

2.3. DOĞRUSAL İÇ İLİŞKİNİN EN ÇOK OLABİLİRLİK LOJİSTİK KESTİRİCİSİNE ETKİSİ

Doğrusal iç ilişki için birçok ölçüm önerilmiştir ve bunlardan en çok kullanılanları,

i) R_j^2, j-inci bağımsız değişkenin regresyondaki belirleyicilik katsayısı,

ii) $\left(\delta_j'\delta_j\right)$, (i) de söz edilen regresyondaki kalan kareler toplamı,

iii) μ_j, $\left(X'VX\right)$ matrisinin $\mu_1 \leq \mu_2 \leq ... \leq \mu_p$ şeklinde sıralı ayırtedici (latent) kökleridir.

Verilen bu karakterizasyonlar için,

i) R_j^2 *'nin bazı j ler için 1' e yakınsaması,*

ii) $\left(\delta_j'\delta_j\right)$ *'nin bazı j ler için 0' a yakınsaması,*

iii) μ_j *'nin bazı j ler için 0' a yakınsaması*

durumunda doğrusal iç ilişkinin olduğu söylenebilir. Doğrusal iç ilişkinin derecesi bu ölçümlerin limitlerine (sırasıyla 1,0 ve 0) olan yakınlığı ile saptanır. Bu ölçümlerin avantaj ve dezavantajları vardır. Örneğin; R_j^2 'nin avantajı, kolay yorumlanır olması ve en küçük kareler kestiriminin değişim artışının ölçümünü, VIF=$1/1-R_j^2$ olduğundan, VIF'e yakın olmasıdır, dezavantajı ise birkaç tane bir arada olan doğrusal iç ilişkiyi tanımada başarısız olmasıdır.

$\hat{\beta}$'nın kovaryans matrisi $\left(X'VX\right)^{-1}$ dir ve büyük n değeri için $Var\left(\hat{\beta}_j\right)$ yaklaşık olarak $\left(X'VX\right)^{-1}$ matrisinin (j,j)-inci elemanı olan $\left(X'VX\right)^{jj}$ dir. x_j, X matrisinin *j*-inci sütunu ve X_j, X matrisinin *j*-inci sütununun silinmesiyle kalan $\left(n\times p\right)$ matris, X matrisi $\left[x_j, X_j\right]$ şeklinde bölünmüş matris olarak yazıldığında [30],

$$\begin{aligned} Var(\hat{\beta}_j) &\approx \left(X'\hat{V}X\right)^{jj} \\ &= \left\{x_j'Vx_j - x_j'VX_j\left(X_j'VX_j\right)^{-1}X_j'Vx_j\right\}^{-1} \end{aligned} \tag{2.10}$$

şeklinde bulunur. (2.10) modelinde, a_j sabitlerin vektörü olmak üzere, bazı *j*' ler için x_j yerine $x_j = X_j a_j + \delta_j$ yazıldığında,

$$Var(\hat{\beta}_j) \approx \left(\delta_j' S \delta_j\right)^{-1}$$

olarak elde edilir. Buradaki *S,* $S = V - VX_j\left(X_j'VX_j\right)^{-1}X_j'V$ matrisidir. Uygulamada x_j nin elemanları sonludur, bu nedenle $n\to\infty$ için $\lim\left(1/n\right)\left(X'VX\right)$ matrisi sonlu pozitif tanımlıdır. *V* sınırlı olduğundan, *S* de sınırlı elde edilir. Böylece veriler daha fazla doğrusal iç ilişkili olduğunda, bazı *j* ler için $\delta_j'\delta_j \to 0$ ve $Var\left(\hat{\beta}_j\right)\to\infty$ olur. Eğer doğrusal iç ilişkinin derecesi büyükse, kestirimlerin biri veya daha fazlası belirsiz olacaktır ve kestirimler bağımsız değişkenlerin doğru etkilerini yansıtmayacaktır.

2.3.1. Lojistik Regresyonda Doğrusal İç İlişkiye örnek

Önce kurulmuş iki veri küme analizine sonra verinin gidişatına bakılacak. Her iki veri kümesi de 0-1 ikili yanıt değişkenler ve x,y açıklayıcı değişkenlerden oluşuyor

[18]. Tablo 2.4' de görüleceği gibi açıklayıcı değişkenler tek başına iken yanıtta anlamlı (significant) etkiye sahip oldukları halde, birlikte iken Wald testine göre bu anlamın kaybolduğu görülür. Bu da standart çoklu regresyonda bir iç ilişki probleminin olduğunu gösterir.

Type	Model	coeff	SE(coeff)	P(Wald)	%corr
UNIV	x	-1,02	0,41	0,01	%72
UNIV	y	0,99	0,40	0,01	%68
MULT	x y	-5,02 5,12	4,17 4,30	0,23 0,23	%84

Tablo 2.4 : Birinci veri kümesi için x ve y değişkenlerinin katsayı değerleri

İşaretlerdeki bazı değişikliklerle Tablo 2.5 de de aynı sonuçlar elde edilir. İki açıklayıcı değişken bir arada olduğunda, şişirilmiş varyans kestirimi ve şişirilmiş regresyon katsayıları elde ediliyor. Medikal uygulamalarda sıkça kullanılan bir yol açıklayıcı değişkenlerden birinin çıkarılmasıdır, böylece birinci veri kümesinde doğru sınıflandırma oranı %84 den %72 veya %68' e düşer, ikinci veri kümesinde ise bu oran %92' den %92 veya %84' e düşer.

Type	Model	coeff	SE(coeff)	P(Wald)	%corr
UNIV	x	0,46	0,20	0,02	%92
UNIV	y	0,45	0,19	0,02	%84
MULT	x y	2,75 -2,27	2,90 2,84	0,34 0,42	%92

Tablo 2.5 : İkinci veri kümesi için x ve y değişkenlerinin katsayı değerleri

Açıklayıcı değişkenlerden birinin silinmesiyle birinci veri kümesinde verimde bazı kayıplar olur. Bu kayıplar, veri kümesinin boyutu ve performans ölçümünde sınıflandırma oranının güçlüğüdür, daha büyük farklar da saçılım grafiğinde görülür.

Kestiricilerin asimptotik kovaryans matrisi

$$(X'VX)^{-1}$$

dir. Burada X tasarım matrisi, $p(x_i)$, x_i ortak değişkenli vektörlü lojistik modelde i-inci gözlemin olasılığı olmak üzere V, $V = diag(v_i)$, $v_i = p(x_i)[1-p(x_i)]$, dır.

Fisher bilgi matrisi

$$W = X'VX$$

şeklindedir. Eğer $S = V^{1/2}X$ 'nin rankı tam sütun ranktan az ise Fisher matrisi tekildir. ($V^{1/2}$,yarı pozitif tanımlı olan $n \times n$ tipindeki V matrisinin kareköküdür.) Fakat bütün doğrusal sonlu regresyon katsayılarının kümesi için V matrisi pozitif tanımlı olduğundan W sadece X tam sütun ranklı olmadığında singülerdir [18].

2.4. ALTERNATİF KESTİRİCİLER

2.4.1. Ridge Kestiricisi

Lineer regresyonda iç ilişki problemini yok etmek için Hoerl ve Kennard'ın (1970a,1970b) öne sürdüğü Ridge kestirimi kullanılabilir. Schaefer (1979) ve Schaffer, Roi ve Wolfe (1984) çalışmalarında Ridge kestirimini verilerin iç ilişki olduğu durumda lojistik regresyon için geliştirmiştir. Ağırlıklı tasarım matrisinin sütunları arasında iç ilişki olduğunda en çok olabilirlik kestiricisinin beklenen normu çok büyük olduğundan Ridge kestiriminin kullanımı daha uygundur. Ridge kestiricisinin normu en çok olabilirlik kestiricisinin normundan daha küçüktür ve Ridge kestiricisinin $\left[(y-\hat{\pi})'(y-\hat{\pi})\right]$ ağırlıklı hata kareler toplamı da en çok olabilirlik kestiricisininkinden daha küçüktür. Ridge kestiricisi $\hat{\beta}_R$ ile gösterilir ve aşağıdaki şekildedir,

$$\hat{\beta}_R = \left(X'VX + kI \right)^{-1} X'VX\hat{\beta} \ . \tag{2.11}$$

Buradaki k Ridge parametresi olarak adlandırılır. k ' nın 0 değeri için Ridge kestiricisi en çok olabilirlik kestiricisine eşittir. k arttıkça Ridge kestiricisinin normu küçülür ve 0' a yaklaşır. $\hat{\beta}_R$, $\hat{\beta}$ nın ve $\left(X'VX \right)^{-1}$ in bir fonksiyonu olduğundan hesaplanması kolaydır. $\hat{\beta}$ ve $\left(X'VX \right)^{-1}$ in ikisi de lojistik regresyon için mevcut paket programlarından kolayca elde edilir.

k ' nın sabit ve stokastik değeri için Ridge regresyonun asimptotik özellikleri aşağıdaki gibidir [29],

i) $\hat{\beta}_R$, $\hat{\beta}$ ' nın tutarlı bir kestirimidir,

ii) $\sqrt{n}\left(\hat{\beta}_R - \hat{\beta} \right)$ 0 ortalamalı ve $\left(X'VX \right)^{-1}$ varyanslı normal dağılıma yakınsar

iii) $\left| x_{ij} \right|$, i ve j ' nin tüm değerleri için sınırlıdır.

iv) Q sonlu determinanta sahip pozitif tanımlı bir matris olmak üzere

$$\lim_{n \to \infty} \frac{i(\beta)}{n} \to Q \ \text{dır.}$$

v) k sabit olmak üzere, $\hat{k}$ Uyarlayıcı (adaptive) Ridge parametresinin limiti

$$\lim_{n \to \infty} \frac{\hat{k}}{n} \to k \ \text{dır.}$$

Buradan Ridge kestiricisinin (2.11) deki denklemi,

$$\hat{\beta}_R = \left(\frac{X'VX}{n} + \frac{kI}{n} \right)^{-1} \frac{X'VX}{n} \hat{\beta}$$

olarak yazılabilir. En çok olabilirlik kestiricisinin asimptotik özelliklerinin ve (v) özelliğinin kullanılmasıyla Ridge regresyon kestiricisinin dağılımı $\left(Q + kI \right)^{-1} Q\hat{\beta}$

ortalamalı ve $(Q+kI)^{-1}Q(Q+kI)^{-1}$ varyanslı normal dağılıma yaklaşır [16]. $\hat{k}$ Uyarlayıcı Ridge parametresinin limiti k =0' a yaklaştığında Ridge regresyon kestiricisi

$$\hat{\beta}_R \sim N[\beta,(X'VX)^{-1}]$$

olarak elde edilir. Bu kestirici tutarlı ve asimptotik olarak en çok olabilirlik kestiricisine denktir.

Ridge regresyon kullanmaktaki asıl amaç, hata kareler ortalaması (MSE) ve varyansı daha küçük olan en çok olabilirlik kestirimini küçülten bir kestirim yöntemi olmasıdır. Çoklu regresyonda k ' nın seçimi hala çözümsüzdür, bu yüzden k 'nın seçimi için kesin bir kural henüz olmamakla birlikte verilen bazı seçim yolları,

a) $k = \dfrac{1}{\max\limits_j \left(\gamma_j'\hat{\beta}\right)^2}$

b) $k = \dfrac{p+1}{\hat{\beta}'\hat{\beta}}$

c) $k = \dfrac{1}{\hat{\beta}'\hat{\beta}}$

[11;13;30] şeklindedir. Buradaki $\hat{\beta}$, β ' nın en çok olabilirlik kestiricisi, γ_j', $\left(X'\hat{V}X\right)$ bilgi matrisinin j-inci özdeğerlerine karşılık gelen ayırtedici vektörleri ve p açıklayıcı değişken sayısıdır. Çoklu regresyonda iyi sonuç veren ve lojistik regresyona genişletilmişi olan

$$k = \frac{p+1}{\hat{\beta}'\hat{\beta}}$$

olarak seçilen k parametresi lojistik regresyon için umut vericidir. Bunlara ek olarak Lee ve Silvapulle (1988) Uyarlayıcı Ridge parametre kestiricisi için

d) $\hat{k} = \dfrac{[\dot{I}z\ (X'\hat{V}X)]}{\hat{\beta}'(X'\hat{V}X)\hat{\beta}}$

sabitini belirlemişlerdir.

2.4.2. Temel Bileşenler Kestiricisi

Genelleştirilmiş lineer modellerde temel bileşenler yönteminin kullanımı, bilgi matrisinin spektral ayrışımı kullanılarak yeniden parametreleştirilmesiyle ele alınır. Yeniden parametreleştirme modelin kanonik formu olarak bilinir ve

$$\underline{\eta} = X\underline{\beta} + \underline{\varepsilon} = Z\underline{\alpha} + \underline{\varepsilon}$$

olarak yazılır. Buradaki $\underline{\alpha} = M'\underline{\beta}$ dir. Burada, M ' nin i-inci sütunu bilgi matrisinin i-inci özdeğerine karşılık gelen özvektörlerden oluşan matristir. $Z = XM$, $(p+1)$ tane temel bileşen olarak bilinen $Z = \left(\underline{z}_1, \ldots, \underline{z}_{(p+1)}\right)$ olacak şekildeki ilişkili olmayan z_i değişkenleri üretir.

Bu model için $\Lambda = diag(\lambda_i)$ ve $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_{p+1}$ bilgi matrisinin özdeğerleri olmak üzere, $\dot{I}z\ \{Var(\hat{\alpha})\}$ ' i asimptotik olarak,

$$\begin{aligned} \dot{I}z\ \{Var(\hat{\alpha})\} &= \dot{I}z\ \left\{\left(Z'VZ\right)^{-1}\right\} \\ &= \dot{I}z\ \left\{\left(M'X'\Lambda XM\right)^{-1}\right\} = \dot{I}z\ \left\{\Lambda^{-1}\right\} \\ &= \sum_{i=1}^{p+1} \frac{1}{\lambda_i} \end{aligned} \tag{2.12}$$

şeklinde elde edilir. $\underline{z}_i$ bileşeni ile tanımlı ağırlıklı tasarım matrisinin varyans oranı,

$$prop(z_i) = \frac{\lambda_i^{-1}}{\sum_{i=1}^{p+1} \lambda_i^{-1}}$$

şeklindedir. Eğer bilgi matrisi iyi koşullu ise en küçük özdeğere ilişkin bileşenin varyansa katkısı çok az olur. Bu bileşen modelden çıkarıldığında çok küçük bir bilgi kaybedileceğinden ve bileşenler dik olduğundan kalan parametrelerin kestirimi bu elemeden etkilenmeyecektir. Fakat (2.12) eşitliğinden de görüldüğü gibi en küçük özdeğerin $\dot{I}z\left\{Var(\hat{\alpha})\right\}$' e katkısı çok büyüktür. En küçük özdeğere ilişkin bileşen modelden çıkarılırsa $\dot{I}z\left\{Var(\hat{\alpha})\right\}$' in değeri çok azalacaktır. Modele katkıları az olan birden fazla bileşen modelden çıkarılabilir. Bu bileşenlerin bazılarının silinmesiyle parametreler hakkında az miktarda bilgi kaybedilir ancak kestiricinin varyansında büyük oranda artış meydana gelir.

Bu sonuçlardan faydalanılarak iki farklı temel bileşen kestiricisi ele alınmıştır [31]. Birinci tip temel bileşen kestiricisi tekrarlı yöntem ile elde edilir ve

$$\hat{\beta}_{tb} = \left(X'X\right)^{+} X'y + \sum_{l=1}^{t}\left(X'\hat{V}_l X\right)^{+} X'\left(y-\hat{\pi}_l\right)$$

şeklindedir. Buradaki t, toplam tekrar sayısı, $\lambda_1^* \geq \lambda_2^* \geq \ldots \geq \lambda_{p+1}^*$ ler $\left(X'X\right)$ matrisinin özdeğerleri, m_i^*, $\left(X'X\right)$ matrisinin i-inci özdeğerine karşılık gelen özvektör ve r, silinen bileşen sayısı olmak üzere

$$\left(X'X\right)^{+} = \sum_{i=1}^{p+1-r} \frac{m_i^*\left(m_i^*\right)'}{\lambda_i^*}$$

ve

$$\left(X'\hat{V}_l X\right)^{+} = \sum_{i=1}^{p+1-r} \frac{m_{il}\left(m_{il}\right)'}{\lambda_{il}}$$

şeklinde elde edilir. Buradaki $\lambda_{1l} \geq \lambda_{2l} \geq \ldots \geq \lambda_{(p+1)l}$, $\left(X'\hat{V}X\right)$ matrisinin l-inci tekrarındaki özdeğerleri ve m_{il}, $\left(X'\hat{V}X\right)$ matrisinin i-inci özdeğerine karşılık gelen özvektördür. Bu tekrarlı kestirici her tekrarda $\left(X'\hat{V}X\right)$'in değerini gerektirir. Bundan dolayı Schaefer (1986) tek adımda hesaplanacak başka bir temel bileşen lojistik kestiricisi öne sürmüştür. $\hat{\beta}$, β' nın en çok olabilirlik kestiricisi ve $\left(X'\hat{V}X\right)^{+} = \sum_{i=1}^{p+1-r} \frac{m_i\left(m_i\right)'}{\lambda_i}$ olmak üzere ikinci temel bileşen lojistik kestiricisi,

$$\hat{\beta}_{tb} = \left(X'\hat{V}X\right)^{+}\left(X'\hat{V}X\right)\hat{\beta}$$

olarak verilir. Marx (1988) ve Marx ve Smith (1990a), uygun bilgi matrisi kullanarak temel bileşenler kestirimini genelleştirilmiş lineer modellere genişletmişler ve tekrarlı yöntemde en çok olabilirlik kestiricisinin yaklaşımına gerek duyulmadığı için tekrarlı yöntemin tek adım yönteminden daha iyi olduğunu göstermişlerdir. Genelleştirilmiş lineer modelde ve lojistik regresyonda iç ilişki varlığında tek adım ve tekrarlı temel bileşenler kestirimleri en çok olabilirlik kestiriminden daha iyi sonuçlar verir [23;31].

2.4.3. Stein Kestiricisi

Lojistik regresyon modelde, β ' nın en çok olabilirlik kestiricisinin varyansı

$$Var(\hat{\beta}) = E\left\{\left[\beta - E(\hat{\beta})\right]\left[\beta - E(\hat{\beta})\right]'\right\} \approx \left(X'VX\right)^{-1}$$

olarak yazılır. Her iki tarafın izi alındığında:

$$\begin{aligned} \text{İz}\ \left\{E\left[\beta - E(\hat{\beta})\right]\left[\beta - E(\hat{\beta})\right]'\right\} &= E\left(\text{İz}\ \left\{\left[\beta - E(\hat{\beta})\right]\left[\beta - E(\hat{\beta})\right]'\right\}\right) \\ &= E\left(\left\{\left[\beta - E(\hat{\beta})\right]\left[\beta - E(\hat{\beta})\right]'\right\}\right) \\ &= E\left(\hat{\beta}'\hat{\beta}\right) - E\left(\hat{\beta}'\right)E\left(\hat{\beta}\right) \approx \text{İz}\ \left(X'VX\right)^{-1} \end{aligned}$$

elde edilir. $\left[E\left(\hat{\beta}\right)\right]' E\left(\hat{\beta}\right) \geq 0$ olduğundan

$$\begin{aligned} E\left(\hat{\beta}'\hat{\beta}\right) &= \text{İz}\ (\Lambda) \\ &= \sum_{i=1}^{p+1} \frac{1}{\lambda_i} \end{aligned}$$

olur. $\lambda_1 > \lambda_2 > ... > \lambda_{p+1}$ olmak üzere $\Lambda = diag\left(\lambda_1, \lambda_2, ..., \lambda_{p+1}\right)$, $X'VX$ bilgi matrisinin özdeğerlerinin oluşturduğu köşegen matrisidir. λ_i özdeğerleri 0' a yaklaştıkça iç ilişki ortaya çıkar ve en çok olabilirlik kestiricisi olan $\hat{\beta}$ ' nın normunun karesinin beklenen değeri büyür. Stein (1960) lineer regresyon üzerine çalışmasında en çok olabilirlik kestiricisini küçültecek Stein kestirim tekniğini öne sürmüştür. Schaefer (1976) bu kestiriciyi lojistik regresyona genişletmiştir. $\hat{\beta}$, en çok olabilirlik kestiricisi olmak üzere Stein kestiricisi

$$\hat{\beta}_S = c\,\hat{\beta} \quad , \quad 0 < c < 1$$

ile verilmiştir. Stein kestiricisinin hata kareler ortalaması (MSE)

$$\begin{aligned} MSE(\hat{\beta}_S) &= E\left(\left[\hat{\beta}_S - \hat{\beta}\right]' \left[\hat{\beta}_S - \hat{\beta}\right]\right) = E\left(\left[c\hat{\beta} - \hat{\beta}\right]' \left[c\hat{\beta} - \hat{\beta}\right]\right) \\ &= \dot{I}z\left[Var(\hat{\beta})\right] + \left[E\left(\left[c\hat{\beta} - \hat{\beta}\right]\right)\right]' E\left(\left[c\hat{\beta} - \hat{\beta}\right]\right) \\ &= c^2\, \dot{I}z\left(X'\hat{V}X\right) + (c-1)^2\, \beta'\beta \end{aligned}$$

olarak elde edilir. Stein kestiricisi $\hat{\beta}_S$ ' yi minimum yapan c değeri,

$$\frac{\partial\, MSE(\hat{\beta}_S)}{\partial c} = 2c\, \dot{I}z\left(X'\hat{V}X\right) + 2(c-1)\beta'\beta = 0$$

$$c = \frac{\beta'\beta}{\beta'\beta + \dot{I}z\left(X'\hat{V}X\right)}$$

olarak bulunur. c ' nin bu değeri için lojistik regresyonda elde edilen kestirimin performansı Ridge regresyon kestiricisinin performansından daha kötüdür [29].

c ' nin seçimi için önerilen bir başka ölçüt ise hata kareler ortalamasını (MSE) değil de

$$L = E\left(\left[\hat{\beta}_S - \beta\right]' X'VX\left[\hat{\beta}_S - \beta\right]\right) = E\left(\left[c\hat{\beta} - \beta\right]' X'VX\left[c\hat{\beta} - \beta\right]\right)$$

beklenen değerini minimum yapan c değeridir [23].

M, i-inci sütunu $\underline{m}_i = (m_{i1},...,m_{i(p+1)})$ olacak şekildeki $X'VX$ matrisinin i-inci özvektörlerinin oluşturduğu matris, $\Lambda = diag(\lambda_i)$ olmak üzere, bilgi matrisinin spektral ayrışımından elde edilen

$$L = E\left(\left[c\hat{\beta} - \beta\right]' M\Lambda M'\left[c\hat{\beta} - \beta\right]\right)$$

ve λ_i, $X'VX$ bilgi matrisinin özdeğeri ve $\alpha = M'\beta$ kullanılmasıyla

$$L = E\left(\left[c\hat{\alpha} - \alpha\right]' \Lambda\left[c\hat{\alpha} - \alpha\right]\right)$$

$$= E\left(\sum_{i=1}^{p+1}\left[c\hat{\alpha}_i - \alpha_i\right]^2 \lambda_i\right)$$

olarak elde edilir. Buradan

$$L = \sum_{i=1}^{p+1} E\left(\left[c\hat{\alpha}_i - \alpha_i\right]^2 \lambda_i\right) = \sum_{i=1}^{p+1} Var\left(c\hat{\alpha}_i\right) + \left\{E\left(\left[c\hat{\alpha}_i - \alpha_i\right]\right)\right\}^2 \lambda_i$$

$$L = c^2\left(p+1\right) + \sum_{i=1}^{p+1}\left(c-1\right)^2 \alpha_i^2 \lambda_i$$

olur. L' yi minimum yapan c değeri ise:

$$\frac{\partial L}{\partial c} = 2c(p+1) + \sum_{i=1}^{p+1} 2\left(c-1\right)\alpha_i^2\lambda_i = 0$$

$$c = \frac{\sum_{i=1}^{p+1}\alpha_i^2\lambda_i}{(p+1) + \sum_{i=1}^{p+1}\alpha_i^2\lambda_i}$$

olarak bulunur. c' nin bu değerinin performansı ile ilgili çalışmalar Valverde (1997) tarafından yapılmıştır.

BÖLÜM 3.

BAZI YANLI KESTİRİCİLER İLE İLGİLİ GELİŞMELER

3.1. RIDGE LOJİSTİK KESTİRİCİSİ (Schaeffer,Roi,Wolfe, 1984)

Genel lojistik model;

$$y_i = \pi_i + \varepsilon_i \ , \ \text{i=1,2,...,n} \tag{3.1}$$

şeklinde verilir. Burada,

$\pi_i : \pi_i = P(y_i = 1) = 1 \big/ 1 + \exp\left\{-\beta_0 - \beta_1 x_{i1} \ldots - \beta_p x_{ip}\right\}$

y_i :0 veya 1 değeri alan i-inci gözlem değeri,

$\varepsilon_i : E(\varepsilon_i) = 0$, $v_i = \left\{\pi_i\left(1-\pi_i\right)\right\}$, $V = diag\left\{v_i\right\}$

β_j :bilinmeyen parametreler,

x_{ij} : gözlemlenebilir bağımsız değişkenlerdir.

X bağımsız değişkenlerin matrisi ve $\tilde{x}_i'$, X matrisin i-inci satırı olmak üzere π_i,

$$\pi_i = 1 \Big/ \left(1 + \exp\left[-\tilde{x}_i'\beta\right]\right)$$

$$= f\left(\tilde{x}_i', \beta\right)$$

şeklinde yazılabilir. Burada β, β_j nin $\left((p+1)\times 1\right)$ vektörüdür. β'nın ML kestirimi $\hat{\beta}$,

$$X'(y - \hat{\pi}) = 0$$

ML denkleminin çözümüyle bulunur. Buradaki y, $(n\times 1)$ vektör ve $\hat{\pi}$, $\hat{\pi}_i = f\left(\tilde{x}_i', \hat{\beta}\right)$ nin $(n\times 1)$- vektörüdür.

Bu çalışmada çoklu doğrusal iç ilişki için; i) R_j^2'nin bazı j ler için 1' e yakınsaması, ii) $\left(\delta_j'\delta_j\right)$'nin bazı j ler için 0' a yakınsaması, iii) μ_j'nin bazı j ler için 0' a yakınsaması ölçütleri ele alınmaktadır. Doğrusal iç ilişkinin derecesi bu ölçütlerin limitlerine olan yakınlığı ile saptanır.

Schaeffer, Roi ve Wolfe (1984) lojistik regresyonda kullanmak için Ridge kestiricisini ele almış ve

$$E(\hat{\beta}'\hat{\beta}) = \hat{\beta}'\hat{\beta} + \dot{I}z\ \left\{Var\left(\hat{\beta}'\right)\right\} + \hat{\beta}'\text{nın yanının karesi} \geq \hat{\beta}'\hat{\beta} + \Sigma\ Var\left(\hat{\beta}'\right)$$

olduğunu göstermiş. Buradan hareketle verilerin doğrusal iç ilişkili olduğu durumda en çok olabilirlik kestiriminin ortalamadan çok uzun olacağını belirtmişlerdir. $\hat{\beta}$, IWLS kestiricisi olduğunda, $\hat{\beta}$ yaklaşık olarak Ağırlıklı Hata kareler toplamını (WSSE) en küçük yapar. Buradan Ridge lojistik kestirici, WSSE yi arttıracak minimum uzunluğa sahip kestirici olarak belirlenir [30]. Sabit kestiriciler için $\hat{\beta}$ yerine $\hat{B}$ yazılarak WSSE deki artış,

$$\phi = \left[\hat{\pi}(\hat{\beta}) - \hat{\pi}(\hat{B})\right]' V^{-1}\left[\hat{\pi}(\hat{\beta}) - \hat{\pi}(\hat{B})\right] + 2\left[\hat{\pi}(\hat{\beta}) - \hat{\pi}(\hat{B})\right]' V^{-1}\left[y - \hat{\pi}(\hat{\beta})\right] \quad (3.1)$$

ile verilir. Buradaki $\hat{\pi}(\hat{B})$, β’nın kestirimi olarak $\hat{B}$’ nın kullanılması ile elde edilen π’nin kestirimidir. $\hat{\pi}$ ların doğrusallaştırılması ile

$$\phi = \left(\hat{B} - \hat{\beta}\right)' X'VX\left(\hat{B} - \hat{\beta}\right) \quad (3.2)$$

eşitliği elde edilir. $\hat{B}$’nın uzunluğunun en küçük yapılmasıyla Ridge lojistik kestirici

$$\hat{\beta}_R = \left(X'VX + kI\right)^{-1} X'VX\hat{\beta} \quad (3.3)$$

olarak elde edilir. Buradaki k, küçük ϕ değerli (3.2) deki kısıtlamayı sağlayacak şekilde seçilir. Uygulamalarda k’ yı seçtikten sonra ϕ’ yi hesaplamak kolaylaşırken, V yerine kestirimi olan $\hat{V}$ yazılarak

$$\hat{\beta}_R = \left(X'\hat{V}X + kI\right)^{-1} X'\hat{V}X\hat{\beta}$$

ifadesi kullanılır. Ancak lojistik regresyon programlarında (3.3) denklemindeki kestirim yöntemi bulunduğu için (3.3) denkleminin kullanımı daha caziptir.

$MSE(\hat{\beta})$, k ve $\hat{\beta}(0)=\hat{\beta}$ ya göre sabit olduğundan, k ' nın sadece pozitif yönde 0' a yaklaşması ile $MSE\left(\hat{\beta}(k)\right)$ ' nın negatif eğilimde olduğunu göstermek gerekir. $\hat{\beta}$ ' nın MSE değeri $O(1/n^3)$ sıralamasından bulunabilir [4]. (1.2) ML denklemindeki $\hat{\pi}$ için Taylor açılımı kullanılarak yeniden düzenlenmesi ile,

$$(\hat{\beta} - \beta) = \left(XVX\right)^{-1} X'\varepsilon - (1/2)\left(XVX\right)^{-1} X'm^* \qquad (3.4)$$

eşitliği elde edilir. Burada ε , ε_i nin bir vektörü ve m^* , $\left(\hat{\beta} - \beta\right)$ daki karesel biçimlerin bir vektörüdür. $\left(\hat{\beta} - \beta\right)$ için m^* yerine yaklaşık bir çözümün konmasıyla (3.4) denklemi

$$(\hat{\beta} - \beta) = \left(XVX\right)^{-1} X'\varepsilon - (1/2)\left(XVX\right)^{-1} X'm \qquad (3.5)$$

elde edilir. Burada m , ε ' nun karesel biçimlerinin bir vektörüdür. $E(\varepsilon) = 0$ ve $Var(\varepsilon) = V$ kabul edildiğinden, $\hat{\beta}$ ' nın yanlılığı ve varyansı (3.5) denkleminin kullanılması ile

$$\begin{aligned} MSE(\hat{\beta}) = \dot{I}z\ &\left\{\left(X'VX\right)^{-1}\left[X'VX + \left(1/4\right)X'Var(m)X - X'Cov(\varepsilon, m)X\right]\left(X'VX\right)^{-1}\right\} \\ &+ 1/4\left(X'h\right)'\left(X'VX\right)^{-2}\left(X'h\right) \end{aligned} \qquad (3.6)$$

olarak elde edilir. Buradaki h , m ' nin beklenen değeridir. n yeterince büyük alındığında $O(1/n^3)$ terimleri ihmal edilebilir.

(3.3) de $\hat{\beta}(k)$, $\hat{\beta}$ 'nın lineer dönüşümü olduğundan, $\hat{\beta}$ 'nın varyansı ve yanlılığı kullanılarak $\hat{\beta}(k)$ 'nın varyansı ve yanlılığı elde edilebilir, ayrıca $\hat{\beta}(k)$ 'nın MSE si,

$$\begin{aligned} MSE\left(\hat{\beta}(k)\right) &= \dot{I}z\ \left\{\left(X'VX\right)^{-1}\left[X'VX+\left(1/4\right)X'Var(m)X-X'Cov(\varepsilon,m)X\right]\left(X'VX\right)^{-1}\right\} \\ &+\left\{\left(-k\beta-1/2\,X'h\right)'\left(X'VX+kI\right)^{-2}\left(-k\beta-1/2\,X'h\right)\right\} \end{aligned} \tag{3.7}$$

olarak elde edilir. λ_j, $\left(X'VX\right)$ matrisinin $\lambda_0 \le \lambda_1 \le ... \le \lambda_p$ şeklinde sıralı ayırtedici kökleri ve γ_j ayırtedici vektörleri olsun. Bu değerler (3.7) denkleminde kullanılması ile

$$\begin{aligned} MSE\left(\beta(k)\right) &= \Sigma\left(\lambda_j+k\right)^{-2}\left\{\lambda_j+\gamma_j X'\left[1/4\,Var(m)-Cov(\varepsilon,m)\right]X\gamma_j\right\} \\ &+\Sigma\left(\lambda_j+k\right)^{-2}\left\{\gamma_j'\left(-k\beta-1/2\,X'h\right)\right\}^2 \end{aligned} \tag{3.8}$$

olarak elde edilir. Başka bir durum belirtilmediği sürece j için tüm toplamlar 0 dan p ye kadar alınır.

Genelliği bozmadan, veriler arasında sadece bir çoklu doğrusal iç ilişki olduğu ve sadece λ_0 'ın 0' a yaklaştığı kabul edilsin. n ' nin büyük olması ve $O(1/n^3)$ terimlerinin ihmal edilmesi varsayımından, (3.8) in k' ya göre türevi ($k \to 0$ için)

$$MSE'\left(\hat{\beta}(0^+)\right) = -2\Sigma\lambda_j^{-2}\left\{1-\left(1/2\right)\left(\gamma_j'X'h\right)\left(\gamma_j'\beta\right)\right\}$$

olarak bulunur [29]. Burada c, x_{ij} ' nin sınırlarına bağlı bir sabit olmak üzere

$$\left[\left(1/2\right)\left(\gamma_j'X'h\right)\left(\gamma_j'\beta\right)\right] \le c\sqrt{\lambda_j}$$

eşitsizliğini sağlar. λ_j ler pozitif olduğundan,

$$MSE'\left(\hat{\beta}(0^+)\right) \le -2\lambda_j^{-2}\left(1-c\sqrt{\lambda_j}\right) \tag{3.9}$$

olarak elde edilir. Burada $\lambda_j^{-2}\left(1-c\sqrt{\lambda_j}\right)$, $\sqrt{\lambda_j}=4\!\!\big/\!_{3c}$ değeri için sonlu minimuma sahiptir. Böylece (3.9),

$$\begin{aligned}MSE'\left(\hat{\beta}(0^+)\right) &\le -2\lambda_0^{-2}\left(1-c\sqrt{\lambda_0}\right)-2p\left\{1-c\left(4\!\!\big/\!_{3c}\right)\right\}\Big/\left(4\!\!\big/\!_{3c}\right)^4\\ &\le -2\lambda_0^{-2}\left(1-c\sqrt{\lambda_0}\right)+\left(2p/3\right)\left(3c/4\right)^4\end{aligned}$$

olarak yeniden düzenlenebilir. Çoklu doğrusal iç ilişkinin derecesi çok şiddetli oldukça, $\lambda_0 \to 0$ ve $MSE'\left(\hat{\beta}(0^+)\right)\to -\infty$ olur. Buradan büyük n ve iç ilişkili veriler için, $k \to 0^+$ oldukça $MSE\left(\hat{\beta}(k)\right)$ negatif olur ve böylece $MSE\left(\hat{\beta}(k)\right)\le MSE\left(\hat{\beta}\right)$ olacak şekilde bir k bulunabilir.

Böylece veriler çoklu doğrusal iç ilişkili ve n yeterince büyük olduğunda bazı k lar için Ridge Lojistik kestirici $\hat{\beta}$' nınkinden daha küçük MSE' ye sahiptir. Aşağıdaki deneysel çalışmada, örneklem büyüklüğü 250 ve çoklu doğrusal iç ilişkinin derecesi şiddetli ($R^2=0.90$) ise Ridge kestiricisinin daha küçük MSE ye sahip olduğu gösterilmiştir.

Bu deneysel çalışmanın amacı: 1) n' yi ve ML kestiriciden daha iyi performansa sahip Ridge kestirici için doğrusal iç ilişkinin derecesini belirlemek, 2) k' nın verilerden kestirilebildiğini belirlemektir. k değeri verilerden kestirilebildiğinde Ridge kestirici, $\hat{\beta}$' dan daha küçük MSE ye sahip olacaktır.

k 'nın seçimi için kesin bir kural olmamakla birlikte verilen bazı seçim yolları aşağıdaki gibidir.

I) $k=\dfrac{1}{\max\limits_j\left(\gamma_j'\hat{\beta}\right)^2}$, γ_j', $\left(X'\hat{V}X\right)$' in ayırtedici vektörleri [11;12]

II) $k=\dfrac{p+1}{\hat{\beta}'\hat{\beta}}$ [13]

Uygulamalarda I) ve II)’ yi hesaplamak kolaydır. I) ve II)’ ye ek olarak diğer bir seçim,

III) $k = \frac{1}{\hat{\beta}'\hat{\beta}}$ [30]

dir. Bu deneysel çalışmadaki deneysel veriler 1968-1975 yılları arasında on iki yanık ihtisas enstitüsünden alınmıştır (Feller ve Crane,1971). Bağımlı değişken ölümcül yanıklar ve bağımsız değişkenler ölümcül yanıkların anlamlı açıklayıcı değişkenlerinden seçilmiştir (Cornell ve Feller,1979). Bağımsız değişkenler cinsiyet, yaş, vücudun toplam yanığı (TB), üçüncü derece yanık (FTB) olarak alınmıştır. Buradaki TB ve FTB iç ilişkilidir. Bu iç ilişkinin etkisini incelemek için yaş, TB ve FTB değişkenler merkezileştirilmeden iç ilişkisinin derecesi yapay olarak arttırılmıştır. Tablo 3.1’ de tüm değişken kümelerinin R_j^2 değerleri verilmiştir.

Değişkenler	Cinsiyet	Yaş	$(\text{Yaş})^2$	TB	$(\text{TB})^2$	FTB	$(\text{FTB})^2$
Merkezileştirilmiş Karesel Terimler	%4.44	%36.86	%37.89	%76.99	%77.61	%86.56	%84.88
Merkezileştirilmemiş Karesel Terimler	%4.40	%90.63	%90.64	%93.59	%94.46	%92.24	%92.03

Tablo 3.1: Deneysel çalışmadaki merkezileştirilmiş ve merkezileştirilmemiş verilerin R_j^2 değerleri.

Örneklem boyutunun Ridge kestiricisi üzerindeki üzerindeki etkisini belirlemek için 14313 birimlik veri kümesi 250 birimlik 57 alt kümeye rasgele bölünmüştür ve bölünme 500 birimlik 28 küme elde etmek için tekrar yapılmıştır. Bu veri kümesiyle k’ nın üç değeri için üç Ridge kestirimi ve En Çok olabilirlik kestirimi lojistik regresyon kullanılarak hesaplanmış ve bu hesaplamalar Tablo 3.2 ile karşılaştırılmıştır. Çoklu regresyonda Ridge kestiricilerinin performansının, p bağımsız değişken sayısı ve $\underline{\beta}$ ile iç ilişkili vektör arasındaki θ açısından etkilendiği görülmektedir. p arttıkça veya $\theta = 90^o$ olunca Ridge kestiricisinin performansı iyileşir.

Kestiriciler	Boyut	İç İlişkinin derecesi	
		Az	Çok
MLE	250	495.06 ± 507.56 (57)	2131.50 ± 2479.17 (57)
	500	143.43 ± 122.45 (28)	881.72 ± 836.87 (28)
Ridge Kestirici [k'nın(I) değeri için]	250	369.67 ± 361.42 (57)	1496.05 ± 1852.10 (57)
	500	124.76 ± 100.95 (28)	688.55 ± 715.56 (28)
Ridge Kestirici [k'nın(II) değeri için]	250	397.27 ± 385.52 (57)	1597.11 ± 1903.68 (57)
	500	130.21 ± 106.25 (28)	881.72 ± 836.87 (28)
Ridge Kestirici [k'nın(III) değeri için]	250	222.17 ± 123.04 (57)	561.41 ± 544.26 (57)
	500	115.16 ± 78.76 (28)	353.69 ± 316.06 (28)

Tablo 3.2: Örneklem boyutuna göre ML ve Ridge kestiricilerin MSE ± MSE'nin standart sapması

3.2. VERİLERİN İÇ İLİŞKİLİ OLDUĞU LOJİSTİK REGRESYONDA ALTERNATİF KESTRİCİLER (Schaeffer, 1986)

Lojistik model

$$\pi = P(y = 1 \mid x_1, x_2, ..., x_{p-1})$$
$$= \left[1 + \exp(-\beta_0 - \beta_1 x_1 - ... - \beta_{p-1} x_{p-1}\right]^{-1}$$
$$\pi = \left[1 + \exp(-x'\beta)\right]^{-1} \qquad (3.10)$$

eşitliği ile verilir. Burada y, ikili (0/1) değerler alan yanıt değişken, β, parametre vektörü, $x' = (1, x_1, ..., x_{p-1})$ bağımsız değişken vektörü ya da risk faktörleridir.

Tekrarlı ağırlıklandırılmış en küçük kareler kestirimi (IWLS) ile elde edilen β' nın kestiriminin $(l+1)$-inci tekrarı

$$\hat{\beta}_{l+1} = \hat{\beta}_l + (X'\hat{V}_l X)^{-1} X'(y - \hat{\pi}_l)$$

şeklindeki (2.9) denklemidir.

$(X'\hat{V}_l X)^{-1}$, $\hat{\beta}$' nın kovaryans matrisinin bir yaklaşımı olarak kullanılır. $\hat{V}_l = diag\{\hat{\pi}_{li}(1-\hat{\pi}_{li})\}$ ve $\hat{\pi}_i$, (1.1) denkleminde β için $\hat{\beta}$'yı sağlayan π_i'nin en çok olabilirlik kestirimidir.

İç ilişkinin genel tanımı, kesin veya yaklaşık olarak tekil olan X' de sonuçlanan bağımsız değişkenler arasında kesin veya yakın lineer bağımlılık olmasıdır. Bu çalışmada iç ilişki için

i) R_j^2 : kalan bağımsız değişkenlerde j-inci bağımsız değişkenin regresyonunda determinasyon katsayısı.

ii) μ_j : $(X'X)$'in $\mu_1 \leq \mu_2 \leq ... \leq \mu_p$ sıralı ayırtedici kök

ölçütleri kullanılmıştır. Eğer R_j^2 1'e veya μ_j 0'a yakınsarsa bu durumda iç ilişki vardır. Her iki ölçümünde avantaj ve dezavantajları vardır. Örneğin, R_j^2'ye başvurulma nedeni kolay yorumlanır olması ve dik olması durumda en küçük kareler kestiriminin değişim artışını ölçen VIF' e (varyans şişirme faktörü) yakın olmasıdır. R_j^2'nin dezavantajı ise, birkaç tane bir arada olan veriler arasındaki iç ilişkiyi tanımada başarısız olmasıdır.

j-inci en küçük kareler kestiriminin varyansı $\sigma^2 (X'X)^{jj}$ olduğunda $(X'X)^{jj}$, $(X'X)^{-1}$'in (j, j)-inci elemanı olup

$$
\begin{aligned}
(X'X)^{jj} &= (1-R_j^2)^{-1}(USSQ_j)^{-1} \\
&= \sum_{k=1}^{p} \mu_k^{-1} \gamma_{kj}^2
\end{aligned}
$$

olduğu gösterilebilir. $USSQ_j$, j-inci bağımsız değişkenin düzeltilmemiş (uncorrected) kareler toplamı, γ_{kj} ayırtedici vektörün j-inci elemanıdır. Bazı j ler için, iç ilişkinin derecesi çoklu regresyondaki en küçük karelerin varyanslarını arttırdığı gibi bu artış kestirimi de belirsizleştirir. Sezgisel olarak, lojistik regresyondaki iç ilişki, $\left(X'\hat{V}X\right)^{-1}$ kullanılarak bulunan $Var\left(\hat{\beta}\right)$' nın yaklaşımında da buna benzer problemlere neden olabilir.

Bu çalışmaya göre, sadece iki değişken arasında iç ilişki varsa μ_1' in 0 a yakın olduğu varsayılmaktadır. μ_k^*' lar $\left(X'\hat{V}X\right)$ nin ayırtedici kökleri olmak üzere, $\mu_k^* \leq \left(\frac{1}{4}\right)\mu_1$ olduğunu göstermek kolaydır ve ayırtedici kök iç ilişkinin ölçütü olduğundan iç ilişki lojistik regresyonda da sorun olur. Ayrıca, γ_k^*, μ_k^* ile ilgili $\left(X'\hat{V}X\right)$ in ayırtedici vektörü ve γ_{kj}^*, γ_k^*'nın j-inci elemanı olmak üzere,

$$Var(\beta_j) = \left(X'\hat{V}X\right)^{jj}$$
$$= \sum_{k=1}^{p}\left(\mu_k^*\right)^{-1}\left(\gamma_{kj}^*\right)^2 \geq \left(\mu_1^*\right)^{-1}\left(\gamma_{1j}^*\right)^2 \geq 4\mu_1^{-1}\left(\gamma_{1j}^*\right)^2$$

şeklindedir. Bazı j ler için, γ_{1j}^* sıfır olmayacaktır ve böylece karşılık gelen varyans kestirimi büyük ve belirsiz olacaktır.

Sonuç olarak iç ilişki, ML lojistik regresyon kestiriminde kesinlik problemlerine sebep olur. Bağımsız değişkenler arasındaki iç ilişkinin bulunuşunda lojistik modeldeki β'nın kestirimi çok belirsizdir ve analizin amacı tehlikeye girecektir. İç ilişki problemi bir kesinlik problemi olduğundan, bir sonraki bölümde kesinlik problemine değinen alternatif tahmin ediciler gösterilecektir. Bu çalışmada verilen alternatif üç kestirici; (1) Ridge Kestiricisi, (2) Temel Bileşenler Kestiricisi ve (3) Stein Kestiricisidir.

İç ilişki problemi Hoerl ve Kennard (1970a.b) ın Ridge kestirim tekniğinin kullanılmasıyla ele alınabilir. Schaffer, Roi ve Wolfe (1984) lojistik regresyonda Ridge kestiricisi kullanmıştır. Ridge kestiricisi $\hat{\beta}_R$ ile gösterilir ve

$$\hat{\beta}_R = \left(X'VX + kI\right)^{-1} X'VX\hat{\beta}$$

olarak elde edilir, buradaki k Ridge parametresi olarak adlandırılır.

k nın seçimi çoklu regresyonda hala çözümsüzdür, bu yüzden k 'nın seçimi için kesin bir kural henüz yoktur. Bununla beraber Hoerl, Kennard ve Baldwin (1975) çoklu regresyonda iyi çalışan bir alternatif bulmuştur ve lojistik regresyona genişletilmişi olan

$$k = \frac{p}{\hat{\beta}'\hat{\beta}}$$

değeri umut vericidir [30]. k 'nın bu seçim kuralı Ridge regresyon kestiricisi ele alınırken kullanılmıştır.

Webster, Gunst ve Mason (1974)' ının gösterimine benzer bir gösterim kullanılmasıyla, Temel Bileşenler Kestiriciyi göstermek için bazı ek teknik terimlere ihtiyaç vardır. $\left(\mu_{jl}^{*}, \gamma_{jl}^{*}\right)$ sıralı ayırtedici kökler olsun ve $\left(X'VX\right)$'in vektörlerine karşılık gelsin, buna göre

$$\left(X'VX\right) = \sum_{j=1}^{p} \mu_{jl}^{*} \gamma_{jl}^{*} \gamma_{jl}^{*\prime}$$

şeklinde yazılır. ML ve IWLS tekrar yöntemi için tipik bir başlangıç noktası en küçük kareler kestirimidir. Bu tekrarlama

$$\begin{aligned} \hat{\beta} &= \hat{\beta}_0 + \sum_{l=0}^{L} \left(X'\hat{V}_l X\right)^{-1} X'\left(y - \hat{\pi}_l\right) \\ &= \left(X'X\right)^{-1} X'y + \sum_{l=0}^{L} \left(X'\hat{V}_l X\right)^{-1} X'\left(y - \hat{\pi}_l\right) \end{aligned} \tag{3.11}$$

şeklindedir. Buradaki L, kestirimin yakınsak olduğu tekrar sayısı ve $\hat{\beta}_0$, β'nın en küçük kareler kestirimidir. Eğer tek bir iç ilişki olduğu varsayılırsa, μ_1, 0'a yakınsar bu durumda μ_{1l}^{*}' de 0'a yakınsar. Burada $\left(X'X\right)$ ve $\left(X'\hat{V}X\right)$ deki çok küçük değişimin birinci temel bileşenin silinmesiyle kaybolduğu kanıtlanacaktır. Temel bileşenler kestiricisini elde etmek için (3.11) denkleminde $\left(X'X\right)^{-1}$ için

$$\sum_{j=2}^{p} \mu_j^{-1} \gamma_j \gamma_j'$$

ve $\left(X'\hat{V}X\right)^{-1}$ için

$$\sum_{j=2}^{p} \mu_{jl}^{*-1} \gamma_{jl}^{*} \gamma_{jl}^{*-1}$$

kullanılır. Böylece Temel bileşenler kestiricisi

$$\hat{\beta}_{tb} = \sum_{j=2}^{p} \mu_j^{-1} a_j \gamma_j + \sum_{l=0}^{L} \sum_{j=2}^{p} \mu_{jl}^{*-1} a_{jl}^{*} \gamma_{jl}^{*}$$

olur. Burada

$$a_j = \gamma_j' X'y$$

ve

$$a_{jl}^{*} = \gamma_{jl}^{*} X'\left(y - \hat{\pi}_l\right)$$

dir. $\left(X'X\right)^{+} = \sum_{j=2}^{p} \mu_j^{-1} \gamma_j \gamma_j'$ ve $\left(X'\hat{V}X\right)^{+} = \sum_{j=2}^{p} \mu_{jl}^{*-1} \gamma_{jl}^{*} \gamma_{jl}^{*-1}$ olmak üzere, daha basit bir gösterimin kullanılmasıyla Temel bileşenler kestiricisi

$$\hat{\beta} = \left(X'X\right)^{+} X'y + \sum_{l=0}^{L} \left(X'\hat{V}_l X\right)^{+} X'\left(y - \hat{\pi}_l\right) \tag{3.12}$$

şeklinde elde edilir. Fakat $\left(X'\hat{V}X\right)^{-1}$ matrisinin her tekrarlamada hesaplanması gerektiğinden Temel bileşenler kestiricisini elde etmek zordur, bu yüzden aşağıdaki alternatif önerilmektedir.

Lojistik fonksiyon için, x sabit olmak üzere $\hat{\beta}$ daki küçük değişiklikler $\hat{\pi}_i$ de ve dolayısıyla $\hat{V}$ de çok küçük değişiklikler meydana getirir. Böylece $\left(X'\hat{V}X\right) \approx \left(X'\hat{V}_l X\right)$ ve $\left(X'\hat{V}X\right)^{+} \approx \left(X'\hat{V}_l X\right)^{+}$ olur. $\hat{\beta}$, Temel bileşenler çoklu regresyondaki en küçük kareler tahminin dönüşümüne benzer biçimde,

$$\begin{aligned}
\hat{\beta}_{tb} &= \left(X'\hat{V}X\right)^{+}\left(X'\hat{V}X\right)\hat{\beta} \\
&= \left(X'\hat{V}X\right)^{+}\left(X'\hat{V}X\right)\left(X'X\right)^{-1}X'y + \sum_{l=0}^{L}\left(X'\hat{V}X\right)^{+}\left(X'\hat{V}X\right)\left(X'\hat{V}_lX\right)^{-1}X'\left(y-\hat{\pi}_l\right) \\
&\approx \left(X'\hat{V}X\right)^{+}C^{*}X'y + \sum_{l=0}^{L}\left(X'\hat{V}X\right)^{+}X'\left(y-\hat{\pi}_l\right) \\
&\approx \left(X'X\right)^{+}X'y + \sum_{l=0}^{L}\left(X'\hat{V}X\right)^{+}X'\left(y-\hat{\pi}_l\right)
\end{aligned}$$

elde edilir. Burada $\left(X'\hat{V}X\right)$, $C^{*}\left(X'X\right)$ ve $\left(X'\hat{V}X\right)^{+}$, $C^{*\,-1}\left(X'X\right)^{+}$ yaklaşımı yapılmıştır. Dolayısıyla $\hat{\beta}$' nın basit bir dönüşümü, yaklaşık olarak (3.12) denklemindeki Temel bileşenler kestiricisine eşittir.

Üçüncü alternatif kestirici için,

$$E(\hat{\beta}'\hat{\beta}) \geq \dot{I}z\left(Var(\hat{\beta})\right) = \dot{I}z\left\{(X'\hat{V}X)^{-1}\right\} \geq \mu_1^{*\,-1} \geq 4\mu_1^{-1}$$

olduğu gösterilebilir. İç ilişki varlığında $\hat{\beta}$ ortalama olarak fazla uzun olacaktır. ML kestiricisini küçültmek için,

$$\hat{\beta}_s = c\hat{\beta} \quad , \quad 0 < c < 1$$

olan Stein kestiricisi tanımlanmıştır. c' nin seçimi için Stein (1960)'ın öne sürdüğü

$$c = \frac{\left(\hat{\beta}'\hat{\beta}\right)}{\left[\hat{\beta}'\hat{\beta} + \dot{I}z\left\{\left(X'\hat{V}X\right)^{-1}\right\}\right]}$$

değeri kullanılır.

Bu kestirici çoğu lojistik regresyon programlarında kolaylıkla elde edilebilen niceliklere bağlı olduğundan bu kestiricinin kullanımı kolaydır.

Özet olarak, Ridge, Temel Bileşenler ve Stein Kestiricilerin varyansları En Çok Olabilirlik kestiricisinin varyansından daha küçük varyansa sahip olduğundan, iç ilişkiden daha az etkilenir. Böylece bu kestiriciler iç ilişkinin varyans üzerindeki etkilerini azaltarak, iç ilişki problemiyle ilgilenir ve daha kesin kestirimle sonuçlanır.

Bu çalışmada, veriler iç ilişkili olduğu durumlar için ML kestiricisine alternatif olarak çeşitli lojistik kestiriciler gösterilmiştir. Simülasyon sonuçları ile lojistik regresyondaki iç ilişkinin ML kestiricisini ciddi olarak etkilediğini ispatlanmıştır. Bu etki, iç ilişkinin yönü ve β paralel olduğunda büyük ve dengesiz kestirimler üretmesidir. Bu durumda hem Ridge kestiricisi hem de Temel Bileşenler kestiricisi MSE' yi azaltmada ve düşük değerli hata kareler üretmede daha iyi performans göstermiştir. Stein kestiricisinin performansının da iyi olduğu gösterilmiştir. Ancak iç ilişki vektörü ile β' nın dik olması durumunda, ML kestiricisinin performansı büyük ölçüde daha iyidir.

Bu alternatif kestiricilerin avantajı, iç ilişkinin etkisini azaltması ve kestirilen bağımsız değişkenlerin etkilerinin belirsizliğini ve doğruluğunu değerlendirmesidir. Üç kestiricinin başka bir avantajı ise, ML kestiricisinin basit bir dönüşümü olduklarından var olan lojistik regresyon programlarından kolayca elde edilmesidir.

3.3. YÜKSEK DERECEDE ÇOKLU İÇ İLİŞKİLİ LOJİSTİK REGRESYONUN KESTİRİMİ İÇİN TEMEL BİLEŞENLERİN KULLANIMI (Aguilera, Escabias, Valderrama, 2006)

Çoklu iç ilişki durumunda lojistik model parametrelerinin kestirimini bulmak ve sürekli ortak değişkenli problemin boyutunu azaltmak için, orijinal açıklayıcı değişkenlerin en iyi temel bileşenlerinin indirgenmiş kümesini açıklayıcı değişkenler olarak kullanılması önerilmiştir. Önerilen temel bileşenler için lojistik regresyon modelinin performansı bir simülasyon çalışması ile analiz edilmiştir.

Bu çalışmada Massy (1965)' nin ortaya koyduğu lineer durumda temel bileşenler regresyon modelinin (PCR) bir genişletilmişi olan temel bileşen lojistik regresyon model (PCLR) önerilmiştir.

$\mathcal{X}_1, \mathcal{X}_2, ..., \mathcal{X}_p$ hatasız gözlemlenmiş sürekli değişkenler kümesi, bu değişkenlerin n gözlemlerin matrisi de $\mathcal{X} = \left(x_{ij}\right)_{n\times p}$ dir. $Y = \left(y_1, ..., y_n\right)'$, $\mathcal{X}$' deki gözlemlere göre $\mathcal{Y}$ ikili yanıt değişkenin rasgele örneklemi ve $y_i \in \{0,1\}, i = 1,...,n$ dir. Logistik regresyon model

$$y_i = \pi_i + \varepsilon_i \quad , \qquad i = 1,2,...,n \tag{3.13}$$

π_i, $\left(\mathcal{X}_1 = x_{i1}, \mathcal{X}_2 = x_{i2}, ..., \mathcal{X}_p = x_{ip}\right)$ bilinirken $\mathcal{Y}$' nin beklenen değeridir ve

$$\pi_i = P\left\{\mathcal{Y} = 1 | \left(\mathcal{X}_1 = x_{i1}, \mathcal{X}_2 = x_{i2}, ..., \mathcal{X}_p = x_{ip}\right)\right\} = \frac{\exp\left\{\beta_0 + \sum_{i=1}^{p} x_{ij}\beta_j\right\}}{1 + \exp\left\{\beta_0 + \sum_{i=1}^{p} x_{ij}\beta_j\right\}} \tag{3.14}$$

dir. Buradaki $\beta_0, \beta_1, ..., \beta_p$' ler model parametreleri, ε_i, 0 ortalamalı bağımsız hatalar ve $Var(\varepsilon_i) = \pi_i\left(1 - \pi_i\right)$, $i = 1,...,n$.

Lojit dönüşüm $l_i = \ln\left(\dfrac{\pi_i}{1 - \pi_i}\right)$, $i = 1,...,n$ dir. Buradaki $\dfrac{\pi_i}{1 - \pi_i}$ gözlemlenmiş $x_i = \left(x_{i1}, ..., x_{ip}\right)$ değerleri için $\mathcal{Y} = 1$ yanıtı için odds değerini gösterir.

Lojistik regresyon model bağ fonksiyonu olarak lojit dönüşümün kullanılması ile genelleştirilmiş lineer model olarak elde edilebilir, böylece $L = X\beta$ matris biçiminde ifade edilebilir[25]. Buradaki $L = \left(l_1, l_2, ..., l_n\right)'$ lojit dönüşüm vektörü, $\beta = \left(\beta_0, \beta_1, ..., \beta_p\right)'$ parametre vektörü, $\underline{1} = (1,1,...,1)'$ birlerin n boyutlu vektörü olmak üzere $X = \left(\underline{1} \mid \mathcal{X} \right)$ tasarım matrisidir. (3.14) denkleminde j-inci prediktör değişken bir birimle artarken, diğer prediktörler kontrol edildiğinde, j-inci parametrenin üssü başarının ($\mathcal{Y} = 1$) odds oranı olarak elde edilir ve bu da

$$\theta\left(\Delta\mathcal{X}_j = 1 \mid \mathcal{X}_k = x_k, \forall k \neq j \right) = \frac{\dfrac{\pi\left(x_i,\ldots,x_{j+1},\ldots,x_p\right)}{1-\pi\left(x_i,\ldots,x_{j+1},\ldots,x_p\right)}}{\dfrac{\pi\left(x_i,\ldots,x_j,\ldots,x_p\right)}{1-\pi\left(x_i,\ldots,x_j,\ldots,x_p\right)}} = \exp\left\{\beta_j\right\}$$

şeklindedir. Buradaki $x_1,\ldots,x_p$ açıklanabilir değişkenlerin tekil gözlemleridir.

Lojistik modeli kestirmek için en çok kullanılan yöntem Hosmer ve Lemeshow (1989) ve Ryan (1997) çalışmalarında olduğu gibi en çok olabilirliktir. $L\left(Y;\beta\right)$ olabilirlik olmak üzere

$$L\left(Y;\beta\right) = \prod_{i=1}^{n} \pi_i^{y_i} \left(1-\pi_i\right)^{1-y_i}$$

dir.

Çok değişkenli temel bileşenler analizi, 20.yüzyıl başlarında Karl Pearson tarafından ortaya konmuş ve Harold Hötteling (1993) tarafından geliştirilmiştir. Temel bileşenler, en büyük varyansa sahip ilişkili olmayan verilerin silinmesiyle oluşan ilişkili değişkenlerin kümesini açıklamaya dayalı çoklu değişken tekniğidir.

p, sürekli değişkenlerin bir kümesini ve yeni $\mathcal{X} = (x_{ij})_{n\times p}$ şeklindeki matrisi oluşturan n gözlemleri ele alındığında yeni matrisin sütun vektörleri $\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_p$ şeklindedir.

$$s_{jk} = \tfrac{1}{n-1}\sum_{i=1}^{n}\left(x_{ij}-\overline{x}_j\right)\left(x_{ik}-\overline{x}_k\right) \text{ ve } \overline{x}_j = \tfrac{1}{n}\sum_{i=1}^{n} x_{ij} \quad (j = 1,2,\ldots,p)$$

olmak üzere örneklemin kovaryans matrisi $S=\left(s_{jk}\right)_{p\times p}$ ile gösterilsin. Gözlemlerin merkezileştirildiği yani $\bar{x}_1=...=\bar{x}_p=0$ ele alınırsa, örneklemin kovaryans matrisi $S=\frac{1}{n-1}\mathcal{X}'\mathcal{X}$ olur.

Örneklem, temel bileşenler $\mathcal{X}$ matrisinin sütunlarının maksimum varyanslı dik doğrusal yayılımı olan $\mathcal{Z}_j=\mathcal{X}\mathcal{V}_j$ $(j=1,...,p)$ ile gösterilir. Temel bileşenlerin katsayısı olan $\mathcal{V}_1,\mathcal{V}_2,...,\mathcal{V}_p$ vektörleri, temel bileşenlerin varyansları ve $\lambda_1\geq\lambda_2....\geq\lambda_p\geq 0$ özdeğerlerine karşılık gelen örneklem kovaryans matrisinin özvektörleridir.

Örneklem kovaryans matrisi $\mathcal{V}$ dik ve sütunları örneklem kovaryans matrisinin özdeğerlerinden oluşan $\mathcal{V}=\left(v_{jk}\right)_{p\times p}$ şeklindedir ve $\Delta=diag(\lambda_1,...,\lambda_p)$ olmak üzere $S=\mathcal{V}\Delta\mathcal{V}'$ şeklindedir, böylece gözlem matrisi $\mathcal{X}=\mathcal{Z}\mathcal{V}'$ olur.

Regresörlerin merkezileştirildiği varsayımı altında

$$\pi_i=\frac{\exp\left\{\beta_0+\sum_{i=1}^{p}x_{ij}\beta_j\right\}}{1+\exp\left\{\beta_0+\sum_{i=1}^{p}x_{ij}\beta_j\right\}}$$

modelinin başarı olasılığı, temel bileşen terimleriyle

$$\pi_i=\frac{\exp\left\{\beta_0+\sum_{j=1}^{p}\sum_{k=1}^{p}z_{ik}v_{jk}\beta_j\right\}}{1+\exp\left\{\beta_0+\sum_{j=1}^{p}\sum_{k=1}^{p}z_{ik}v_{jk}\beta_j\right\}}=\frac{\exp\left\{\beta_0+\sum_{k=1}^{p}z_{ik}\gamma_k\right\}}{1+\exp\left\{\beta_0+\sum_{k=1}^{p}z_{ik}\gamma_k\right\}}$$

şeklinde ifade edilir. Burada z_{ik} $(i=1,...,n;\ k=1,...,p)$, olmak üzere $\mathcal{Z}=\mathcal{X}\mathcal{V}$ temel bileşenler matrisinin elemanları ve $\gamma_k=\sum_{j=1}^{p}v_{jk}\beta_j$ dir. Lojit dönüşümlerle ve temel bileşenlerle, lojistik model matris biçiminde,

$$L=X\beta=ZV'\beta=Z\gamma \qquad (3.15)$$

şeklinde yazılabilir. Burada $Z=\left(\mathbf{1}\,|\,\mathcal{Z}\right)$, $V=\begin{pmatrix} 1 & \underline{\mathbf{0}'} \\ \overline{\mathbf{0}} | & v \end{pmatrix}$, $\mathbf{0}=(0,...,0)'$, $\mathbf{1}=(1,...,1)'$ dir. Lojit modelin parametreleri $\beta=V\gamma$ ile bulunabilir. En çok olabilirlik kullanılarak $\hat{\beta}=V\hat{\gamma}$ ve ön kestirim $\hat{Y}=\hat{\pi}$ elde edilir.

İç ilişki varlığında orijinal parametrelerin kestirimi için temel bileşenlerin silinmiş bir kümesi ele alınarak Temel Bileşenler Lojistik Regresyon modeli oluşturulur. Bunun için Z ve V ayrık matrisleri aşağıdaki gibi yazılır.

$$Z=\left(\begin{array}{cccc|ccc} 1 & z_{11} & \cdots & z_{1s} & z_{1s+1} & \cdots & z_{1p} \\ 1 & z_{21} & \cdots & z_{2s} & z_{2s+1} & \cdots & z_{2p} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 1 & z_{n1} & \cdots & z_{ns} & z_{ns+1} & \cdots & z_{np} \end{array}\right)=\left(Z_{(s)}\middle|Z_{(r)}\right) \quad , \quad (r=p-s)$$

$$V=\left(\begin{array}{cccc|ccc} 1 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & v_{11} & \cdots & v_{1s} & v_{1s+1} & \cdots & z_{1p} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & v_{p1} & \cdots & z_{ps} & z_{ps+1} & \cdots & z_{pp} \end{array}\right)=\left(V_{(s)}\middle|V_{(r)}\right)$$

$Z_{(s)}=XV_{(s)}$ ve $Z_{(r)}=XV_{(r)}$ yazılarak orijinal parametreler

$$\beta=V\gamma=V_{(s)}\gamma_{(s)}+V_{(r)}\gamma_{(r)} \quad , \quad \gamma=\left(\gamma_0\,\gamma_1\ldots\gamma_s\,|\,\gamma_{s+1}\ldots\gamma_p\right)'=\left(\gamma_s'\,|\,\gamma_r'\right)'$$

olarak açıklanır. (3.15) denklemi ile verilen temel bileşenlerin tüm terimli lojit modeli $L=Z\gamma=Z_{(s)}\gamma_{(s)}+Z_{(r)}\gamma_{(r)}$ şeklinde yazılır. s-temel bileşenli PCLR modeli $\left(\text{PCLR(s)}\right)$, denklemdeki son r-temel bileşenlerinin kaldırılmasıyla

$$y_i=\pi_{i(s)}+\varepsilon_{i(s)}$$

olarak elde edilir. Burada $\pi_{i(s)}$ (i=1,2,…,n) için

$$\pi_{i\,(s)} = \frac{\exp\left\{\gamma_0 + \sum_{j=1}^{s} z_{ij}\gamma_j\right\}}{1+\mathrm{xp}\left\{\gamma_0 + \sum_{j=1}^{s} z_{ij}\gamma_j\right\}}$$

şeklindedir. Bu model için $l_{i(s)} = \ln\left(\pi_{i(s)} / 1-\pi_{i(s)}\right)$ bileşenli lojit dönüşümünün matris yazılımı

$$L_{(s)} = Z_{(s)}\gamma_{(s)} = XV_{(s)}\gamma_{(s)} = X\beta_{(s)}$$

şeklindedir. Bu, PCLR modelindeki β parametrelerinin

$$\hat{\beta}_{(s)} = V_{(s)}\hat{\gamma}_{(s)}$$

şeklindeki en çok olabilirlik kestirimini verir.

BÖLÜM 4.

LOJİSTİK REGRESYON İÇİN LİU KESTİRİMİ

4.1. LİU KESTİRİCİSİ

Kejian (1993) lineer modelde çoklu iç ilişki olması durumunda iç ilişki problemini yok edecek bir kestirici önermiştir. Liu adı verilen bu kestirici

$$\hat{\beta}_d = \left(X'X + I\right)^{-1}\left(X'Y + d\hat{\beta}\right)$$

dır. Buradaki $\hat{\beta}$, en küçük kareler kestirimi ve d, $0 < d < 1$ dir. Liu kestiricisi Ridge ve Stein kestiricilerinin avantajlarını kapsamaktadır. Kejian (1993) teoride ve simülasyon çalışmasında Liu kestiricisinin Ridge kestiricisi ile benzer olduğunu göstermiştir. Liu kestiricisinin Ridge kestiricisinden avantajı ise, $\hat{\beta}_d$ ' nin d ' nin lineer bir fonksiyonu olmasıdır.

d ' nin seçimi için verilen bazı yöntemler:

i) C_L kriteri ile d değeri

$$d_{C_L} = 1 - \hat{\sigma}^2 \left[\sum_{i=1}^{p} \frac{1}{\lambda_i + 1} \Big/ \sum_{i=1}^{p} \frac{\lambda_i \hat{\alpha}_i^2}{\left(\lambda_i + 1\right)^2} \right]$$

ii) Mcdonald-Galarneau yöntemi ile d değeri

$$\hat{d}_{MG} = 1 - \frac{\sum_{i=1}^{p} \frac{\hat{\alpha}_i^2}{\lambda_1 + 1} - \left[\left(\sum_{i=1}^{p} \frac{\hat{\alpha}_i^2}{\lambda_1 + 1} \right)^2 - \hat{\sigma}^2 \sum_{i=1}^{p} \frac{\hat{\alpha}_i^2}{\left(\lambda_1 + 1\right)^2} \sum_{i=1}^{p} \frac{1}{\lambda_1} \right]^{\frac{1}{2}}}{\sum_{i=1}^{p} \frac{\hat{\alpha}_i^2}{\left(\lambda_1 + 1\right)^2}}$$

Bu çalışmada lineer modeldeki iç ilişkiyi yok etmek için verilen Liu kestiricisini Lojistik regresyona uyarlayarak en çok olabilirlik kestiricisi ve Ridge kestiricisi ile karşılaştırmaları yapılacaktır.

4.1.1. Liu Lojistik Kestirici için Yanlılık, Varyans Kovaryans ve Hata Kareler Ortalaması

Doğrusal iç ilişki problemi için bir başka alternatif kestirici Liu Kejian (1993)' nın öne sürdüğü Liu kestiricisidir. Bu kestiricinin lojistik regresyon için gösterimi aşağıdaki gibi bulunur.

$$\pi_i = 1/1 + \exp\{-x_i\beta\}$$

$$E(\hat{\beta}'\hat{\beta}) = \beta'\beta + \dot{I}z\ \left\{Var(\hat{\beta})\right\} + \left(Yan\,(\hat{\beta})\right)\left(Yan\,(\hat{\beta})\right)' \geq \beta'\beta + \sum Var(\hat{\beta}_j)$$

$$\phi = \left[\hat{\pi}(\hat{\beta}) - \hat{\pi}(\hat{B})\right]' V' \left[\hat{\pi}(\hat{\beta}) - \hat{\pi}(\hat{B})\right] + 2\left[\hat{\pi}(\hat{\beta}) - \hat{\pi}(\hat{B})\right]' V^{-1} \left[y - \hat{\pi}(\hat{\beta})\right]$$

$\hat{\pi}$ ' lar doğrusallaştırılarak

$$\phi = \left(\hat{B} - \hat{\beta}\right)' X'VX \left(\hat{B} - \hat{\beta}\right)$$

elde edilir. $\hat{B}$ ' nın uzunluğunu küçültmek için, Lagrangian probleminden

$$F = \hat{B}'\hat{B} + \frac{1}{k}\left[\left(\hat{B} - \hat{\beta}\right)' X'VX\left(\hat{B} - \hat{\beta}\right) - \phi\right]$$

şeklinde yazılır buradan da $0 = k^{1/2}\beta + \varepsilon'$, $d\hat{\beta} = \beta + \varepsilon'$ genişletilmesiyle

$$\hat{B} = \hat{\beta}_d = \left(X'VX + I\right)^{-1}\left(X'VX + dI\right)\hat{\beta} \tag{4.1}$$

Liu lojistik kestiricisi bulunur. Buradaki d, 0<d<1 aralığında değer alan bir parametredir.

$\hat{\beta}$' nın MSE' sini bulmak için, $O(1/n^2)$ dizilimi aşağıdaki gibi yapılır. m, ε da kuadratik form vektörü olmak üzere, Taylor açılımından

$$\begin{aligned}\hat{\beta}_d = \beta &+ \left(X'VX + I\right)^{-1}\left(I + d(X'VX)^{-1}\right)X'\varepsilon \\ &- \frac{1}{2}\left(X'VX + I\right)^{-1}\left(I + d(X'VX)^{-1}\right)X'm\end{aligned} \tag{4.2}$$

şeklinde yazılır. $E(\varepsilon) = 0$ ve $Var(\varepsilon) = V$ varsayıldığından, $\hat{\beta}$' nın yanı ve varyansı (4.2) denkleminden aşağıdaki gibi bulunur.

$$\begin{aligned}MSE(\hat{\beta}_d) = \dot{I}z &\left\{\left(X'VX + I\right)^{-1}\left(I + d(X'VX)^{-1}\right)\right. \\ &.\left[X'VX + \tfrac{1}{4}X'Var(m)X - X'Cov(\varepsilon, m)X\right] \\ &.\left.\left(X'VX + I\right)^{-1}\left(I + d(X'VX)^{-1}\right)\right\} + \left[(1-d)\beta - \tfrac{1}{2}X'h\right]' \\ &.\left(I + d(X'VX)^{-1}\right)'\left(X'VX + I\right)^{-2}\left(I + d(X'VX)^{-1}\right)\left[(1-d)\beta - \tfrac{1}{2}X'h\right]\end{aligned}$$

λ_j ler $\left(X'VX\right)$' in $\lambda_0 \leq \lambda_1 \leq \ldots \leq \lambda_p$ şeklinde sıralı ayırtedici kökleri ve γ_j ler de ayırtedici vektörleri olmak üzere

$$MSE(\hat{\beta}_d) = \sum\left(\lambda_j + 1\right)^{-2}\left(1 + d\lambda_j^{-1}\right)^2$$
$$.\left\{\lambda_j + \gamma_j X'\left[\tfrac{1}{4}Var(m) - Cov(\varepsilon, m)\right]X\gamma_j\right\} \tag{4.3}$$
$$+\sum\left(\lambda_j + 1\right)^{-2}\left(1 + d\lambda_j^{-1}\right)^2\left\{\gamma_j'\left[(1-d)\beta - \tfrac{1}{2}X'h\right]\right\}^2$$

yazılabilir.

4.2. LİU LOJİSTİK KESTİRİCİSİ İLE EN ÇOK OLABİLİRLİK KESTİRİCİSİNİN KARŞILAŞTIRILMASI

Genelliği kaybetmeden veriler arasında sadece bir iç ilişki olduğu ve sadece λ_0 ' ın sıfıra yaklaştığı varsayımı altında, (4.3) denkleminin *d* ye göre türevi alınıp, yeterince büyük n için $O(\frac{1}{n^2})$ sırası ihmal edildiğinde

$$\left[MSE(\hat{\beta}_d)\right]' = 2\sum\left(\lambda_j + 1\right)^{-2}\left(1 + d\lambda_j^{-1}\right)\left(\lambda_j\right)$$
$$-2\sum\left(\lambda_j + 1\right)^{-2}\left(1 + d\lambda_j^{-1}\right)^2\left(\gamma_j'\beta\right) \tag{4.4}$$
$$.\left\{\gamma_j'\left[(1-d)\beta - \tfrac{1}{2}X'h\right]\right\}$$

şeklinde bulunur. $d = 1$ için $\hat{\beta} = \hat{\beta}_d$ olduğundan $MSE'(\hat{\beta}) = MSE'(\hat{\beta}_{d=1})$ olur. d=1 için $MSE'(\hat{\beta}_{d=1})$,

$$\left[MSE(\hat{\beta}_{d=1})\right]' = 2\sum\left(\lambda_j + 1\right)^{-1} + 2\sum\lambda_j^{-2}\left\{\gamma_j'\left[\tfrac{1}{2}X'h\right]\right\}\left(\gamma_j'\beta\right)$$

dır. $\sum\left(\lambda_j + 1\right)^{-2}\left(1 + d\lambda_j^{-1}\right)^2 \geq 0$ olduğundan $0 < d < 1$ için $MSE'(\hat{\beta}) \geq MSE'(\hat{\beta}_d)$ olur. Yani $0 < d < 1$ için $MSE(\hat{\beta}) - MSE(\hat{\beta}_d)$ n.n.d. dir, bu da Liu lojistik kestiriminin en çok olabilirlik kestriminden daha iyi olduğu anlamına gelir.

4.3. LİU LOJİSTİK KESTİRİCİSİ İLE RİDGE KESTİRİCİSİNİN KARŞILAŞTIRILMASI

Liu kestiricisinin Ridge kestiricisinden daha iyi kestirici olması için $MSE(\hat{\beta}_d) - MSE(\hat{\beta}(k)) \leq 0$ olmalıdır.

Bunu göstermek için $\left(\left[MSE(\hat{\beta}_d) \right]' - \left[MSE(\hat{\beta}(k)) \right]' \right) \to -\infty$ olduğunu göstermek yeterlidir. Genelliği kaybetmeden veriler arasında sadece bir iç ilişki olduğu ve sadece λ_0 ' ın sıfıra yaklaştığı varsayımı altında, (4.3) denkleminin d ye göre türevi alınıp, yeterince büyük n için $O(\frac{1}{n^2})$ sırası ihmal edildiğinde

$$\left[MSE(\hat{\beta}_d) \right]' = 2\sum \left(\lambda_j + 1 \right)^{-2} \left(1 + d\lambda_j^{-1} \right) \left(\lambda_j \right) - 2\sum \left(\lambda_j + 1 \right)^{-2} \left(1 + d\lambda_j^{-1} \right)^2 \left(\gamma_j' \beta \right)$$
$$.\left\{ \gamma_j' \left[(1-d)\beta - \tfrac{1}{2} X'h \right] \right\}$$

şeklinde bulunmuştu.

Ridge kestiricisinin hata kareler ortalaması, λ_j, $\left(X'VX \right)$'in $\lambda_0 \leq \lambda_1 \leq ... \leq \lambda_p$ şeklinde sıralı ayırtedici kökleri ve γ_j ayırtedici vektörleri olmak üzere

$$\begin{aligned} MSE\left(\beta(k) \right) = {} & \Sigma \left(\lambda_j + k \right)^{-2} \left\{ \lambda_j + \gamma_j X' \left[\tfrac{1}{4} Var(m) - Cov(\varepsilon, m) \right] X \gamma_j \right\} \\ & + \Sigma \left(\lambda_j + k \right)^{-2} \left\{ \gamma_j' \left(-k\beta - \tfrac{1}{2} X'h \right) \right\}^2 \end{aligned} \tag{3.8}$$

şeklinde elde edilir. $O(\frac{1}{n^3})$ terimleri ihmal edildiğinde, $MSE\left(\beta(k) \right)$' nın k ' ya göre türevinden ve $k \to 0$ dan $\left[MSE\left(\hat{\beta}(0^+) \right) \right]' = -2\Sigma \lambda_j^{-2} \left\{ 1 - \left(\tfrac{1}{2} \right) \left(\gamma_j' X'h \right) \left(\gamma_j' \beta \right) \right\}$ elde edilmişti. $d = 1$ için $\left[MSE(\hat{\beta}_d) \right]'$,

$$\left[MSE(\hat{\beta}_{d=1})\right]' = 2\sum\left(\lambda_j + 1\right)^{-1}\lambda_j^{-1}\left[I - \tfrac{1}{2}\left(\gamma_j X'h\right)\left(\gamma_j \beta\right)\right]$$

dir. c, x_{ij} 'nin sınırlarına bağlı bir sabit olmak üzere [29]

$$\left[(½)\left(\gamma_j' X'h\right)\left(\gamma_j' \beta\right)\right] \leq c\sqrt{\lambda_j}$$

eşitsizliğinden Liu ve Ridge kestiricilerinin $d \to 1$ ve $k \to 0$ için MSE' lerinin türevleri

$$\left[MSE(\hat{\beta}_{1^+})\right]' \leq 2\sum\left(\lambda_j + 1\right)^{-1}\lambda_j^{-1}\left[1 - c\sqrt{\lambda_j}\right]$$

$$\left[MSE\left(\hat{\beta}(0^+)\right)\right]' \leq -2\sum\lambda_j^{-2}\left(1 - c\sqrt{\lambda_j}\right)$$

şeklinde elde edilir. Buradan

$$\left[MSE(\hat{\beta}_{d\to 1^+})\right]' \leq -\left(\sum\left(\lambda_j + 1\right)^{-1}\lambda_j\left[MSE\left(\hat{\beta}(k)\right)\right]'\right)$$

dir. $\lambda_j > 0$ ve $0 < \lambda_j / 1 + \lambda_j < 1$ olduğundan $\left(\left[MSE(\hat{\beta}_d)\right]' - \left[MSE(\hat{\beta}(k))\right]'\right)$ negatif sonsuzlukta değer alır. Löwner kısmi sıralamasına göre $MSE(\hat{\beta}_k) - MSE(\hat{\beta}_d)$ n.n.d. dir, $MSE(\hat{\beta}_d) \leq MSE(\hat{\beta}_k)$ dir. Bu durumda Liu lojistik kestiricisi Ridge Lojistik kestiricisinden daha iyi bir kestiricidir.

4.4. SONUÇLAR VE ÖNERİLER

Lojistik regresyon modelinde veriler iç ilişkili olabilir. Verilerin iç ilişkili olması durumunda bu modelde yansız kestiriciler yetersiz kaldığından yanlı kestiricilerin uygulandığı 3. bölümde verilmiş olan çalışmaların ışığında lojistik regresyon modeline, özel bir yanlı kestirici olan Liu kestiricisi uygulanmıştır.

Uygulanan Liu Lojistik kestiricisi ile yansız kestirici olan En Çok olabilirlik Lojistik kestiricisi MSE ölçütlerine göre karşılaştırılmıştır. Sonuç olarak Liu Lojistik kestiricisinin MSE değerinin daha küçük olduğu yani, bu kestiricinin daha iyi bir kestirici olduğu gösterilmiştir.

Ridge Lojistik kestiricisi [20] ve Liu Lojistik kestiricisi de MSE ölçütlerine göre karşılaştırılmış ve yine Liu Lojistik kestiricisinin daha iyi bir kestirici olduğu gösterilmiştir.

Bundan sonraki çalışmalarda, teorik olarak gösterilen bu karşılaştırmalar örneklem büyüklükleri ve veriler arasındaki açılar farklı alınarak bu karşılaştırmalar incelenebilir.

Verilen yanlı ve yansız kestiriciler ve Liu Lojistik kestirici için kalanlar bulunarak hata yayılım ortalaması (MDE) ölçütlerine göre karşılaştırmalar yapılabilir.

KAYNAKLAR

[1] Aguilers, A.M., Escabias, M. and Valderrama, M.J. (2006). Using Principal Components for Estimating logistics Regression with High-Dimensional multicollinear data. Computational Statistics & Data Analysis. 50, pp. 1905-1924.

[2] Akdeniz, F., Kaçıranlar, S. (1995). On the almost unbiased generalized Liu Estimator and unbiased estimation of the Bias and MSE. Commun. Statist.- Theory and Methods, 24(7), pp. 1789-1797.

[3] Anderson, J.A., Richardson, S.C. (1979). Logistic discrimination and bias correction in maximum likelihood estimation. Technometrics, 21, pp. 71-78.

[4] Bishop,Y.M.M., Feinberg, S.E. and Holland, P.W. (1975). Discrete Multivariate analysis: Theory and Practice. MIT Press, Cambridge, Massachusetts.

[5] Bradly, R.A., Gart, J.J. (1962). The Asymtotic Properties of ML Estimators When Sampling From Associated Population. Biometrika, 49, pp. 205-214.

[6] Cessie, S.L., Houwelingen, J.C.V. (1992). Ridge Estimators in Logistic Regression. *Applied Statistics*, Vol. 41, No. 1, pp. 191-201.

[7] Copas, J.B. (1988). Binary Regression Models for Contaminated Data. Journal of the Royal Statistical Society, Series B, 50, pp. 225-265.

[8] Cordeiro, G.M., McCullagh, P. (1991). Bias correction in generalized linear models. Journal of the Royal Statistical Society, Series B, 53, pp. 629-643.

[9] Cox, D.R., Hinkley, D.V. (1974). Theoretical Statistics. London: Chapman and Hall.

[10] Gunst, R.F., Mason, R.L. (1977b). Biased estimation in Regression: An evaluation Using Mean Square Error. JASA, 72, pp. 616-627.

[11] Hoerl, A.E., Kennard, R.W. (1970a). Ridge Regression: Biased estimation for Nonorthogonal problems. Technometrics, 12, pp. 55-67**.**

[12] Hoerl, A.E., Kennard, R.W. (1970b). Ridge Regression: Applications Nonorthogonal problems. Technometrics, 12, pp. 69-82**.**

[13] Hoerl, A.E., Kennard, R.W. and Baldwin, K.F. (1975). Ridge Regression: Some Simulations. Communications in Statistics, 4(2), pp. 1105-1123.

[14] Hosmer, D. and Lemeshow, S. (1989). Applied Logistic Regression. NY: Wiley & Sons.

[15] Kaçıranlar, S. (1995). Yanlı regresyon tahmin edicileri ve hata kareleri ortalaması kriterine göre karşılaştırmalar. Doktora tezi. Çukurova Üniversitesi, Fen Bilim. Ens.

[16] Kaçıranlar, S., Sakallıoğlu, S. (2001). Combining the Liu Estimator and the Principal component Regression Estimator. Commun. Statist.- Theory and Methods, 30(12), pp. 2699-2705.

[17] Kleinbaum D.G. and Klein, M. (2002). Logistic Regression: A Self-learning Text, Second Edition. Springer-Verlag New York.

[18] Lesaffre, E. And Marx, B.D. (1993). Collinearity in Generalized Linear Regression. Commun. Statist.-Theory Meth., 22(7), pp. 1933-1952.

[19] Liu, K. (1993). A new class of biased estimate in linear regression. Commun. Statistics: Theory and Methods, 22(2), pp. 393-402.

[20] Liu, K. (2003). Using Liu-Type Estimator to Combat Collinearity. Communications in Statistics: Theory and Methods, Vol.32, No.5, pp. 1009-1020.

[21] Mackinnon, M.J. & Putterman, M.L. (1989). Collinearity in Generalized Linear Models. Commun. Statist.-Theory Meth., 18(9), pp. 3463-3472.

[22] Manoukian, E.B. (1989). Modern Concepts and Theorems of Mathematical Statistics. New york:Springer Verlog.

[23] Marx, B.D. (1988). Ill-Conditioned Information Matrices and the Generalized Linear Models: An Asymptotically Biased Estimation Approach. Doctoral Dissertation. Virginia: Virginia Polytechnic Institute and State University.

[24] Marx, B.D. and Smith, E.P. (1990). Principal Component Estimators for Generalized Linear Regression. Biometrika, 77(1), pp. 23-31.

[25] McCullagh, P., Nelder, J.A. (1989). Generalized Linear Models. New York: Chapman and Hall.

[26] Myers, R.H., Montgomery, D.C., Vining, G.G. (2001). Generalized Linear Models: with applications in engineering and sciences. Wiley & Sons, Inc., New York.

[27] Ryan, T.P. (1997). Modern Regression Methods. Wiley, New York.

[28] Sarkar, N. (1992). A new Estimator combining the Ridge Regression and the Restricted Least Squares Methods of Estimation. Commun. Statist.-Theory Meth., 21(7), pp. 1987-2000.

[29] Schaefer, R.L. (1979). Multicollinearity and Logistic Regression. Doctoral dissertation. University of Michigan.

[30] Schaeffer, R.L., Roi, L.D. and Wolfe, R.A. (1984). A Ridge logistic estimator. Communications in Statistics: Theory and Methods, Vol.13, No.1, pp. 99-113.

[31] Schaefer, R.L. (1986). Alternative estimators in logistic regression when the data are collinear. J. Statist. Comput. Simul., Vol. 25, pp. 75-91.

[32] Urgan, N.N. ve Tez, M. (2007). Verilerin Lineer İç İlişkili Olduğu Lojistik Regresyonda bazı Yanlı Parametre Kestiricileri ve Hata Kareler Ortalamasına göre Karşılaştırılmaları. 5.İstatistik Kongresi, 20-24 Mayıs 2007. Antalya/Türkiye.

[33] Valverde, G.R. (1997). Study of Collinearity on Logistic Regression in Presence of Misclassification. Ph.D. Dissertation submitted on 23 July 1997, Univ. of Tulane.

[34] Weissfeld, L.A. and Sereika, S.M. (1991). A Multicollinearity Diagnostic for Generalized Linear Models. Commun. In Statistics-Theory and Methods, Vol. 20, pp. 1183-1198.

[35] http://kosbed.kou.edu.tr/sayi8/bircan.pdf

SEMBOL LİSTESİ

AGE	:	yaş
CHD	:	koroner kalp hastalığı
TB	:	vücudun tamamının yanık olması
FTB	:	üçüncü derecede yanık
$E(Y\|x)$	:	koşullu ortalama
ML	:	en çok olabilirlik
MLE	:	en çok olabilirlik kestirimi
λ_i	:	$X'X$ matrisinin i-inci özdeğeri
ε	:	rasgele hata vektörü
n_i	:	gruplandırılmış verilerde i-inci veri noktasındaki deneme birimi
X	:	tasarım matrisi
$\underline{y}$	:	$(n\times1)$ tipinde gözlemlerin vektörü
$i(\beta)$	:	bilgi matrisi
Q	:	sonlu determinanta sahip pozitif tanımlı matris
$S(\underline{\beta})$	:	score vektörü
S	:	ağırlıklandırılmış kalan kareler toplamı
VIF	:	varyans şişirme faktörü
k_x	:	$X'X$ matrisinin koşul sayısı
C_i	:	$X'X$ matrisinin i-inci koşul indeksi
T_i	:	tolerans faktörü
$prop_{ij}$	:	i-inci bileşenin varyans oranı
R_j^2	:	j-inci bağımsız değişkenin belirleyicilik katsayısı
μ_j	:	$(X'VX)$' in j-inci ayırtedici kökü

WSSE	:	ağırlıklandırılmış hata kareler toplamı
WLS	:	ağırlıklandırılmış en küçük kareler
IWLS	:	tekrarlı ağırlıklandırılmış en küçük kareler
PCR	:	temel bileşenler regresyon modeli
PCLR	:	temel bileşenler lojistik regresyon modeli
$\left(X'VX\right)^{jj}$	:	$\left(X'VX\right)^{-1}$'in (j,j)-inci elemanı
$\left(\delta_j'\delta_j\right)$	:	kalan kareler toplamı
W	:	Fisher bilgi matrisi
$\hat{\beta}_R$	:	Ridge kestiricisi
k	:	Ridge parametresi
$\hat{k}$	:	Uyarlayıcı Ridge parametresi
$\hat{\beta}_{tb}$	:	Temel bileşenler kestiricisi
$\hat{\beta}_S$	:	Stein kestiricisi
μ_k^*	:	$\left(X'\hat{V}X\right)$ nin ayırtedici kökleri
γ_j'	:	$\left(X'\hat{V}X\right)$' in ayrıtedici vektörleri
$\hat{\beta}_d$	:	Liu kestiricisi
n.n.d.	:	negatif olmayan tanımlı

TANIM LİSTESİ

Tanım 1. (Pozitif tanımlı) Simetrik A matrisi için $x'Ax$ karesel formu ele alınsın. $\forall x \neq 0$ için $x'Ax$ karesel formu 0' dan büyük ise A matrisine pozitif tanımlı denir.

Tanım 2. (Yarı Pozitif tanımlı) Simetrik A matrisi için $x'Ax$ karesel formu ele alınsın. En az bir $x \neq 0$ için $x'Ax$ karesel formu 0' dan büyük ya da eşit ise A matrisine pozitif yarı tanımlı denir.

Tanım 3. (Negatif olmayan tanımlı) Bir A matrisi pozitif tanımlı ya da pozitif yarı tanımlı ise A matrisine negatif olmayan tanımlı denir.

Tanım 4. (Spektral ayrışım) $V;$ pxp tipinde kolonları $X'X$ matrisinin özdeğerlerine karşılık gelen normalleştirilmiş özvektörler olan dik matris, $Z=XV$ ve $\alpha=V'\beta$ olmak üzere, $X'X$ matrisinin spektral ayrışımı; $X'X=V\Lambda V'$ ve $(Z'Z)^{-1}=\Lambda^{-1}$ şeklindedir.

Tanım 5. (C_L Kriteri) $SS_{Res}(L)$**,** L**'** ye bağlı kalan kareler toplamı olmak üzere

$$C_L = \frac{SS_{Res}(L)}{\hat{\sigma}^2} - n + 2 + 2\,\dot{I}z(H)$$

dir. (Mallows,1973). Liu Kestirici için H,

$$H = X\left(X'X + I\right)^{-1}\left(X'X + dI\right)X'$$

dır.

Tanım 6. (Löwner kısmi sıralaması) A ve B $m \times m$ tipinde iki matris olsun. B-A n.n.d. ise B matrisi A matrisinden daha büyüktür ve $A \leq B$ olarak yazılır.

ŞEKİL LİSTESİ

TABLO LİSTESİ

DİZİN

ÖZGEÇMİŞ

Nurkut Nuray URGAN, 1977 yılında Tarsus'ta doğdu. İlkokulu Tarsus Şehit İshak İlkokulunda, orta ve lise öğrenimini Abdulkerim Bengi Tarsus Anadolu Lisesinde tamamladı. 1995 yılında Dicle Üniversitesi Fen Edebiyat Fakültesi Matematik Bölümünü kazandı ve 1999 yılında mezun oldu, aynı yıl D.Ü. Fen Bilimleri Enstitüsü'nde yüksek lisansı kazandı. 1999-2000 döneminde Van'ın Edremit ilçesindeki Edremit İlköğretim Okulu ve Edremit Sarmansuyu İlköğretim okulunda görev yaptı. 2000 yılında D.Ü. Fen Edebiyat Fakültesi Matematik Bölümünde araştırma görevlisi olarak göreve başladı. 2002 yılında Dicle Üniversitesi Fen Bilimleri Enstitüsü Matematik Anabilim Dalı Uygulamalı Matematik programında Yüksek Lisansını tamamladı. Aynı yıl Dicle Üniversitesi Fen Bilimleri Enstitüsü Matematik Anabilim dalında Doktora öğrenimine başladı. Halen Dicle Üniversitesi Fen Edebiyat Fakültesi Matematik Bölümünde Araştırma Görevlisi olarak görev yapmaktadır.