T.C.
ESKİŞEHİR OSMANGAZİ ÜNİVERSİTESİ
SAĞLIK BİLİMLERİ ENSTİTÜSÜ
BİYOİSTATİSTİK VE TIBBİ BİLİŞİM ANABİLİM DALI

TÜRETİLMİŞ İKİLİ HETEROJEN VERİ YAPILARINDA GENEL, SAĞLAM VE KESİN LOJİSTİK REGRESYON YÖNTEMLERİNİN KARŞILAŞTIRILMASI

YÜKSEK LİSANS TEZİ

MUZAFFER BİLGİN

TEZ DANIŞMANI
YRD. DOÇ. DR. ERTUĞRUL ÇOLAK

OCAK-2012

T.C.
ESKİŞEHİR OSMANGAZİ ÜNİVERSİTESİ
SAĞLIK BİLİMLERİ ENSTİTÜSÜ
BİYOİSTATİSTİK VE TIBBİ BİLİŞİM ANABİLİM DALI

TÜRETİLMİŞ İKİLİ HETEROJEN VERİ YAPILARINDA GENEL, SAĞLAM VE KESİN LOJİSTİK REGRESYON YÖNTEMLERİNİN KARŞILAŞTIRILMASI

YÜKSEK LİSANS TEZİ

MUZAFFER BİLGİN

TEZ DANIŞMANI
YRD. DOÇ. DR. ERTUĞRUL ÇOLAK

OCAK-2012

KABUL VE ONAY SAYFASI

Muzaffer BİLGİN'in Yüksek Lisans Tezi olarak hazırladığı **"Türetilmiş İkili Heterojen Veri Yapılarında Genel, Sağlam ve Kesin Lojistik Regresyon Yöntemlerinin Karşılaştırılması"** başlıklı bu çalışma Eskişehir Osmangazi Üniversitesi Lisansüstü Eğitim ve Öğretim Yönetmeliği'nin ilgili maddesi uyarınca değerlendirilerek **"KABUL"** edilmiştir.

12/01/2012

Üye: Prof. Dr. Kazım ÖZDAMAR

Üye: Doç. Dr. Didem ARSLANTAŞ

Üye: Doç. Dr. K. Setenay ÖNER

Üye: Doç. Dr. Fezan MUTLU

Üye: Yrd. Doç. Dr. Ertuğrul ÇOLAK

Eskişehir Osmangazi Üniversitesi Sağlık Bilimleri Enstitüsü Yönetim Kurulu'nun 13/01/2012 tarih ve 901/4201 sayılı kararı ile onaylanmıştır.

Prof.Dr. Kazım ÖZDAMAR
Enstitü Müdürü

ÖZET

Sağlık alanında yapılan araştırmalarda ikili şekilde gözlenen bağımlı değişken içeren veri setleri ile sıklıkla karşılaşılmaktadır. Örneğin bazı fenomenler var-yok, ölü-sağ, başarılı-başarısız gibi ikili biçimde sonuçlanabilmektedir. Bu sonuçların ortaya çıkmasında birçok faktör söz konusudur. Bu ilişkinin incelenmesinde bağımlı değişken kategorik yapıda olduğu için lojistik regresyon yöntemi en çok kullanılan yöntemlerden biridir.

Lojistik regresyon yönteminde kullanılan model oluşturma tekniği, istatistik alanında kullanılan diğer model yapılandırma teknikleri ile benzerdir ve lojistik regresyon analizinin amacı en az sayıda bağımsız değişken kullanarak en iyi uyuma sahip olacak şekilde bağımlı ile bağımsız değişkenler arasındaki ilişkiyi tanımlayabilen bir model kurmaktır.

Sağlık alanında yapılan çalışmalarda ikili yapıda gözlenen veri setlerinin analizlerinde en yaygın kullanılan genel lojistik regresyon yöntemlerinin uygulanabilmesi, büyük örnek hacmine ve koşulsuz olabilirlik fonksiyonunun kullanılmasına bağlıdır. Ancak genel lojistik regresyon yöntemleri, örnek hacmi küçük, çarpık, seyrek ya da bağımlı değişkenin beklenenin dışında sapan değerler alması durumunda (heterojen veri seti) geçerli ve güvenilir sonuçlar vermeyebilirler. Bu durumda alternatif yöntemlerin kullanılması güvenilir sonuçların elde edilmesi için gereklidir. Alternatif yöntemler arasında en yaygın olarak kullanılan lojistik regresyon yöntemi kesin lojistik regresyon analizidir. Ancak son yıllarda sağlam lojistik regresyon yöntemleri de alternatif yöntemler arasında yerini almaktadır. Yapılan literatür taramaları sonucunda çok sayıda sağlam lojistik regresyon yöntemine rastlanılmıştır. Croux ve Haesbroeck, Bianco ve Yohai tarafından ortaya atılan sağlam lojistik regresyon yöntemini modifiye ederek diğer sağlam lojistik regresyon yöntemlerine göre hızlı ve stabil sonuç veren bir algoritma geliştirmişlerdir. Bu nedenle sağlam lojistik regresyon yöntemi olarak Croux ve Haesbroeck tarafından geliştirilen yöntem bu tez çalışmasına dahil edilmiştir.

Bu çalışmanın amacı, ikili yapıda bağımlı değişken içeren heterojen veri setlerinin analizlerinde Genel lojistik regresyon, Sağlam lojistik regresyon ve Kesin lojistik regresyon yöntemlerinin performanslarını karşılaştırmaktır.

Yöntemler; parametre tahminlerinin yanlılıkları ve standart hataları kullanılarak ve farklı örnek büyüklüğünde, farklı bozulma oranında simülasyon çalışmaları yapılarak karşılaştırıldı. Yöntemlerin karşılaştırılmasında Monte Carlo simülasyon yöntemi kullanıldı ve analizler R v2.13.2 ve SAS 9.0 paket programlarında yapıldı. Grafikler Minitab 15.0 programında oluşturuldu.

Simülasyon analizleri sonucunda; bozulma oranının %0 olduğu homojen veri setlerinde üç yöntemin de benzer sonuçlar verdiği gözlendi. Bozulmanın var olduğu veri setlerinde sağlam lojistik regresyon yönteminin, genel lojistik regresyon yöntemi ve kesin lojistik regresyon yöntemine göre daha yansız parametre tahminleri verdiği ve sağlam lojistik regresyon yönteminin parametre tahminlerine ilişkin standart hataları düzelterek daha güvenilir sonuçlar verdiği belirlendi.

Anahtar Kelimeler: İkili Gözlemeler, Genel Lojistik Regresyon Yöntemi, Sağlam Lojistik Regresyon Yöntemi, Kesin Lojistik Regresyon Yöntemi, Bozulma Oranı, Yanlılık, Standart Hata.

SUMMARY

The data sets that contain binary dependent variable often encountered in research in the field of health. For example, there are some phenomena such as yes-no, alive - dead and successful - unsuccessful. There are many factors that affect the observation of these results. For certain categories of the dependent variable is the study of this relationship, the logistic regression method is one of the most widely used methods.

Model building technique used in logistic regression analysis is similar to other model building techniques used in statistical field. The purpose of logistic regression analysis is to establish model that can define the relationship between dependent and independent variables by using a minimum number of independent variables having the best fit.

Asymptotic logistic regression is the most common methods used in binary data sets in the field of health studies. The application of this method depends on the use of large sample volume and the unconditional likelihood function. However, the asymptotic logistic regression methods may not release reliable results when the sample size is small, skewed, sparse or contaminated. In this case, the use of alternative methods is required to achieve reliable results. Exact logistic regression analysis is the most widely used method among alternative methods. On the other hand, robust logistic regression methods have become one of the alternative methods in recent years. Croux and Haesbroeck developed an algorithm that works fast and stable than other robust regression methods for the robust logistic regression method proposed by Bianco and Yohai. For this reason, the method improved by Croux and Haesbroeck included in this study.

The purpose of this study, compare the performance of asymptotic logistic regression, robust logistic regression and exact logistic regression on homogeneous contaminated data sets that contains binary dependent variable.

The methods were compared using biases of the parameter estimation and standard errors in different sample size and contamination rate and the comparisons

were performed using Monte Carlo simulation method. The simulations were achieved using R v2.13.2 and SAS 9.0 package programs. The graphs were drawn on Minitab 15.0 program.

As a result of simulation analyses, it was observed that there were no significant differences among the three methods in the homogeneous data sets having 0% contamination rate. In contaminated data sets, it was observed that robust logistic regression methods yielded less biased parameter estimates than asymptotic and exact logistic regression methods, also robust logistic regression methods released more reliable results by adjusting the standard errors for the parameter estimates.

Key Words: Binary Data, Asymptotic Logistic Regression Method, Robust Logistic Regression Method, Exact Logistic Regression Method, Contamination Rate, Bias, Standard Error.

İÇİNDEKİLER

TABLO DİZİNİ

ŞEKİL DİZİNİ

SİMGE VE KISALTMALAR DİZİNİ

GLR	: Genel Lojistik Regresyon
SLR	: Sağlam Lojistik Regresyon
KLR	: Kesin Lojistik Regresyon
GLM	: Generalized Linear Models (Genelleştirilmiş Doğrusal Modeller)
S	: Bozulma Oranı
SAS	: Statistical Analysis System

1. GİRİŞ VE AMAÇ

Sağlık alanındaki incelemeler, teknolojideki gelişmeler sayesinde karmaşık haldeki problemlerin çözümünde ayrıntılı sonuçlar elde etmemizde yardımcı olmayı amaçlamaktadır. Son zamanlarda basit, tek değişkenli ve ikili ilişkiler ile açıklanmaya çalışılan sorunlar, günümüzde de detaylı bir biçimde çoklu ya da çok değişkenli yöntemlerle açıklanmaktadır. Bilimsel çalışmalarda incelenemeyen değişkenler ve bu değişkenler arasındaki ilişkiler, bilimin ve bilimsel yöntemlerin gelişmesi ile günümüzde incelebilir duruma gelmiştir [8].

Sağlık alanında değişkenlerin arasındaki ilişkiyi, değişkenler arasındaki neden sonuç ilişkilerinin belirlenmesi, risk faktörlerinin belirlenmesi ve bu risk faktörlerinden majör ve minör yapıda olanların saptanması ileri istatistiksel yöntemlerle yapılmaktadır.

Günümüzde sağlık alanında yapılan araştırmalarda güvenilir sonuçlar elde etmek için, biyoistatistiksel yöntemlere ihtiyaç duyulmaktadır [8]. Sağlık alanında yapılan çalışmalarda incelenen olayı etkileyen birden fazla bağımsız değişken bulunmaktadır. İncelenen olayın karmaşık olması, olayı açıklamamıza yardımcı olacak olan bağımsız değişken sayısını arttırmaktadır. Bu bağımsız değişkenler sürekli veya kesikli yapıda olabilmektedir. Analizlerin gerçekleşmesinde istatistiksel modellerin kullanılması gerekmektedir. Bağımlı değişkeni etkileyen bağımsız değişkenleri içeren model kurulurken çok iyi seçimler yapılmalıdır.

Elde edilen verilerin analizini gerçekleştirebilmek için kuramsal istatistik modelin matematiksel fonksiyonlar şeklinde ifade edilmesi gerekmektedir. Bu fonksiyonlar, incelenen veriler yardımı ile ileride gerçekleşebilecek olaylar hakkında tahmin yapılması ve bağımlı değişkene etki eden majör faktörlerin belirlenmesini sağlamaktadır [8]. İncelediğimiz olayın çözümünde kullanacağımız modelin seçiminde dikkat etmemiz gereken noktalar vardır. Model kurmamızdaki temel amaç, ileriye yönelik çalışmalar içindir. Elimizdeki bağımsız değişkenler ile olabilecek olan risk faktörlerinin tanımlanması ve bağımlı değişken üzerindeki etkilerinin belirlenmesi temel amaçlardan biridir. Aynı amaç için çok farklı modeller kurulabilmektedir. Her

model kurma işleminin bazı riskleri vardır. Bu riski en aza indirmemiz için en uygun modeli seçmemiz gerekmektedir.

Doğada gözlenen fenomenlerin bazıları var-yok, başarılı-başarısız, ölü-yaşıyor gibi ikili biçimde, bazı fenomenler hiç-az-fazla, yok-orta-çok gibi ikiden fazla seçenekler biçiminde sonuçlanabilmektedir [29, 38]. Bu sonuçların ortaya çıkmasında bir çok faktör söz konusudur. Bu ilişkinin incelenmesinde bağımlı değişken belirli kategorilerde olduğu için lojistik regresyon yöntemi en çok kullanılan yöntemlerden biridir [8, 29].

Lojistik regresyon modelleri, günümüzde biyoloji, tıp, epidemiyoloji, ekonomi, tarım ve veterinerlik alanlarında yaygın olarak kullanılan bir yöntemdir.

Lojistik regresyon analizi, istatistik alanında kullanılan diğer model yapılandırma teknikleri ile benzerdir ve lojistik regresyon analizinin amacı en az sayıda bağımsız değişkeni kullanarak en iyi uyuma sahip olacak şekilde bağımlı ile bağımsız değişkenleri arasındaki ilişkiyi tanımlayabilen bir model kurmaktır [17].

Doğrusal regresyon analizinde bağımlı değişkenin alabileceği değerler tahmin edilirken, lojistik regresyon analizinde bağımlı değişkeninin alacağı değerlerden birinin gerçekleşme olasılığı tahmin edilerek risk faktörlerinin belirlenmesi sağlanır [28, 37].

Lojistik regresyon modelleri ilk olarak Berkson (1944) tarafından epidemiyolojik denemelerde kullanılmıştır. Berkson tarafından kullanılan lojistik regresyon modeli, Cox (1970) tarafından yeniden modifiye edilerek farklı uygulamalar üzerinde denenmiştir. 1979-1983 yılları arasında Anderson tarafından da kullanılarak yaygınlaşmıştır [17].

Sağlık alanında yapılan çalışmalarda genel (asimptotik) lojistik regresyon yöntemlerinin uygulanabilmesi, büyük örnek hacmine ve koşulsuz olabilirlik fonksiyonunun kullanılmasına bağlıdır. Ancak genel lojistik regresyon yöntemleri, örnek hacmi küçük, çarpık, seyrek ya da bağımlı değişkenin beklenenin dışında sapan (aykırı) değerler alması (heterojen veya kontamine olmuş veri seti) durumunda geçerli ve güvenilir sonuçlar vermeyebilirler. Bu şartlar altında *Sağlam Lojistik regresyon*

(SLR) ve *Kesin Lojistik Regresyon* (KLR) yöntemleri*, Genel Lojistik Regresyon* (GLR) yöntemlerine alternatif yöntemler olarak karşımıza çıkmaktadır.

KLR yönteminin temeli 1970 yılında Cox tarafından atılmıştır. Bu yöntemlerin teorisi ve algoritmaları da Hirji, Mehta ve Patel (1987) tarafından oluşturulmuştur [25]. Literatür taramalarında SLR yöntemine ilişkin değişik yaklaşımlar gözlenmiştir. Croux ve Haesbroeck (2003), Bianco ve Yohai (1996) tarafından ortaya atılan SLR yöntemini modifiye ederek diğer SLR yöntemlerine göre hızlı ve stabil sonuç veren bir algoritma geliştirmişlerdir. Bundan dolayı bu tez çalışmasında, Croux ve Haesbroeck (2003) tarafından geliştirilen bu algoritma SLR yöntemi olarak kullanılmıştır.

Bu çalışma;

1. GLR, SLR ve KLR yöntemlerini açıklamak,
2. Bu üç lojistik regresyon modeli arasındaki farkları belirlemek,
3. Heterojen veri yapısı olduğu durumda GLR, SLR ve KLR yöntemlerinin performanslarını, simülasyon çalışması ile karşılaştırmak,

amaçlarını gerçekleştirmek üzere yapılmıştır.

2. GENEL BİLGİLER

Regresyon yöntemleri bağımlı değişken ile bir veya birden fazla bağımsız değişkenler arasındaki ilişkiyi açıklayan veri analizi yöntemlerinin en temel bileşenidir. Bağımlı değişkenin kategorik yapıda olması, iki veya daha fazla sayıdaki kategorilerin gözlenme olasılık değerlerinin belirlenmesi sık karşılaşılan bir durumdur [15]. Bu durumda en çok kullanılan yöntemler lojistik regresyon yöntemleridir ve son yıllarda lojistik regresyon yöntemleri bu durumu açıklamada kullanılan standart yöntemler arasında yer almaktadır [3, 15].

Lojistik regresyon modelleri, bağımlı değişkenin her hangi bir kategorisinin gözlenme olasılığının logitini, model parametrelerinin doğrusal bir fonksiyonu olduğunu kabul etmektedir. Bağımlı değişkenin Bernoulli dağılıma uyduğu durumlarda kullanılan lojistik regresyon yöntemi özellikle sağlık alanında yapılan araştırmalarda çok sık başvurulan veri analizi yöntemidir [36].

Veri analizine başlamadan önce önemli olan nokta; analiz için kullanılan yöntemin amacı ile model kurma teknikleri arasındaki bağıntının iyi anlaşılmasıdır. Özellikle veri analizinde kullanılacak olan bağımlı değişkenin yapısı, dağılımı ve hangi tür ölçekle elde edilmiş olduğu önemli bir noktadır. Örneğin doğrusal regresyon yöntemleri bağımlı değişkenin sürekli olduğu ve normal dağılım gösterdiği durumlarda kullanılan yaygın regresyon yöntemleridir. Lojistik regresyon ile doğrusal regresyon arasındaki en büyük ayırım bağımlı değişkeninin lojistik regresyon analizinde kesikli olmasıdır [15].

Veri analizinde kullanılacak olan bağımlı ve bağımsız değişken yapısına göre değişik yöntemler karşımıza çıkmaktadır. Birer örnek verilecek olunursa; bağımlı değişkenin sürekli ve normal dağılım gösterdiği, bağımsız değişkenlerinin kesikli yapıda olması durumunda *varyans analizi* yöntemleri kullanılabilmektedir. Hem bağımlı hem de bağımsız değişkenlerin kesikli yapıda olması durumunda *log-linear analiz* yöntemleri kullanılabilmektedir. Bütün değişkenlerin sürekli olması durumunda *Doğrusal regresyon analizleri* kullanılabilmektedir [28].

Lojistik regresyon yöntemleri doğrusal regresyon yöntemlerinin dayandığı temelleri kullansa da lojistik regresyon yöntemleri ile doğrusal regresyon yöntemlerinin farkı parametrik modelin seçimi ve varsayımlarıdır [15]. Bu iki yöntem arasındaki 3 temel fark;

1) Doğrusal regresyon analizinde bağımlı değişken sürekli yapıdayken lojistik regresyon analizinde kesikli yapıdadır,
2) Doğrusal regresyon analizinde bağımlı değişkenin değeri, lojistik regresyon analizinde ise bağımlı değişkenin alabileceği kategorilerin gerçekleşme olasılığı tahmin edilir,
3) Doğrusal regresyon analizinde bağımlı değişkenin normal dağılım varsayımı söz konusu iken, lojistik regresyonun uygulanabilmesi için bağımlı değişkenin sürekli olması ve normal dağılım varsayımı ön koşulu içermesi gerekmez [15, 29, 40].

Lojistik regresyon analizinin açıklanmasında kullanılan bazı kavramlar, gösterimler ve modeller aşağıda kısaca açıklanmıştır. Bölüm 2.1'de ikili lojistik regresyon modelleri için ortak kullanılan gösterimler ile GLR yöntemi birlikte açıklanmıştır. Bu bölümde yer alan logit fonksiyon, katsayıların önemliliğinin testinde kullanılan analizler, parametre tahminlerinin güven aralıklarının hesaplanması her üç yöntem içinde geçerlidir. Bölüm 2.2'de SLR yöntemi ve Bölüm 2.3'deki KLR yöntemi için parametre tahminleri ile parametre tahminlerinin anlamlılıklarında kullanılan yaklaşımlar açıklanmıştır. Çünkü bu üç yöntem arasındaki en temel fark parametre tahminlerinin elde edilmesinde gözlenmektedir.

2.1. İkili Lojistik Regresyon Modelleri İçin Ortak Kullanılan Gösterimler ve İkili GLR Yöntemi

Bağımlı değişkenin var-yok, hasta-sağlam gibi ikili kategoride gözlendiği lojistik regresyon modelidir [6, 9]. Bağımsız değişken/değişkenler ile ikili cevap içeren bağımlı değişken arasındaki bağıntıyı ortaya koyar [19]. Bu regresyon modelinde bağımlı değişken Bernoulli dağılımı gösterir [18]. n hacimli bir örnekte birbirinden

bağımsız n tane Bernoulli dağılımı gösteren rassal bağımlı değişkenler $Y_1, \dots, Y_n$ olduğu durumda, bu değişkenlere ilişkin gözlenen değerler vektörü aşağıdaki gibidir;

$$\boldsymbol{y}' = (y_1, \dots, y_n) \tag{2.1}$$

Her bir gözlem için, i=1,…,n

$$\boldsymbol{x_i}' = \left(1, x_{1i}, \dots, x_{pi}\right) \tag{2.2}$$

p elemanlı bağımsız değişkenler vektörü olsun. Tüm gözlemler için bağımsız değişkenler veri matrisi;

$$\boldsymbol{X} = (\boldsymbol{x_1}, \dots, \boldsymbol{x_n})' \tag{2.3}$$

şeklinde oluşur.

i. gözlem için ilgilenilen olayın gözlenme olasılığı $\pi(\boldsymbol{x_i}) = \mathrm{P}(Y_i = 1|\boldsymbol{x_i})$ şeklinde gösterilmektedir. Bu durumda her bir gözlem için ilgilenilen olayın gözlenme olasılığını gösteren olasılıklar vektörü aşağıdaki gibidir;

$$\boldsymbol{\pi} = [\pi(\boldsymbol{x_1}), \dots, \pi(\boldsymbol{x_n})]' \tag{2.4}$$

Verilen bağımsız değişkenlerin değerine bağlı olarak, bağımlı değişkenin beklenen değeri (ortalama değeri) $\mathrm{E}(\mathrm{Y}|\ \boldsymbol{x_i})$ şeklinde gösterilir. Bu değer koşullu ortalama olarak adlandırılır. Bağımlı değişken ikili yapıda olduğundan ve Bernoulli dağılımı gösterdiğinden dolayı bağımlı değişkenin beklenen değeri 0 ile 1 arasındadır. Bu durumda $\mathrm{E}(\mathrm{Y}|\ \boldsymbol{x_i}) = \pi(\boldsymbol{x_i})$ eşitliği karşımıza çıkmaktadır. Bağımsız değişkenler ile bağımlı değişkenin beklenen değeri arasında doğrusal bir model oluşturduğumuzda aşağıdaki eşitliği elde ederiz [15].

$$E(Y|\ \boldsymbol{x_i}) = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi} \tag{2.5}$$

$$E(Y|\ \boldsymbol{x_i}) = \boldsymbol{x_i'\beta} \tag{2.6}$$

olarak tanımlanır. Burada $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_p)'$ parametre vektörünü göstermektedir.

Bu eşitlikte hata terimini beklenen değeri 0 olduğundan eşitlikte hata terimi yer almamaktadır.

Eşitlik (2.5)'de eşitliğin sol tarafı olan koşullu ortalamanın alabileceği değerler 0 ve 1 arasındayken, eşitliğin sağ tarafı bağımsız değişkenlerin tipine ve gözlenme aralığına bağlı olarak -∞ ve +∞ arasında değerler alabilir. Bu sorunu ortadan kaldırmak için bağımlı değişkenin koşullu beklenen değerini çeşitli rotasyonlarla -∞ ve +∞ aralığında tanımlı hale getirmek gerekmektedir. Bu rotasyon için kullanılan en yaygın fonksiyon *Logit* fonksiyonudur.

2.1.1. Logit Fonksiyonu

Logit fonksiyonu, incelenen bir olayın gözlenme olasılığının (P), gözlenmeme olasılığına (1-P) oranının doğal logaritmasını (*ln*) verir ve aşağıdaki şekilde gösterilir (1).

$$Logit[P] = ln\left(\frac{P}{1-P}\right) \qquad (2.7)$$

R. A. Fisher ve Frank Yates de önerdiği gibi bu transformasyonu $\pi(\boldsymbol{x_i})$ cinsinden gösterilecek olursak [3].

Logit transformasyonu;

$$Logit[\,\pi(\boldsymbol{x_i})\,] = ln\left[\frac{\pi(\boldsymbol{x_i})}{1-\pi(\boldsymbol{x_i})}\right] = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi} \qquad (2.8)$$

şeklinde ifade edilir.

Modelin bağımlı değişkeninin sınırlarını genişletmek için uygulanan Logit transformasyonunun bazı özellikleri aşağıdaki gibi sıralanabilmektedir; [17].

- ✓ $\pi(\boldsymbol{x_i})$ arttıkça logit[$\pi(\boldsymbol{x_i})$] de artar.
- ✓ $\pi(\boldsymbol{x_i})$ < 0.5 olduğunda logit[$\pi(\boldsymbol{x_i})$] < 0 ve $\pi(\boldsymbol{x_i})$ > 0.5 olduğunda logit[$\pi(\boldsymbol{x_i})$] > 0 olur.
- ✓ $\pi(\boldsymbol{x_i})$, 0 ile 1 arasında iken logit[$\pi(\boldsymbol{x_i})$] -∞ ile +∞ arasında değerler alabilir.

Bu transformasyonun önemi, $Logit[\,\pi(\boldsymbol{x_i})\,]$'in doğrusal regresyon modelinin genel özelliklerini taşımasıdır. $Logit[\,\pi(\boldsymbol{x_i})\,]$ parametreleri bakımından doğrusal, sürekli ve $\boldsymbol{x_i}$'nin aldığı değerlere bağlı olarak -∞ ve +∞ arasında değişim gösterebilmektedir.

İkili kategoriye sahip olan bağımlı değişkende incelediğimiz kategori Y= 1 ile kodlanırken, diğer kategori ise Y= 0 olarak kodlanmaktadır. Kullanacağımız lojistik regresyon modelinin spesifik formu;

$$Logit[\,\pi(\boldsymbol{x_i})\,] = Logit[\mathrm{P}(Y_i = 1|\boldsymbol{x_i})] = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi} \quad (2.9)$$

$$Logit[\mathrm{P}(Y_i = 1|\boldsymbol{x_i})] = ln\left(\frac{\mathrm{P}(Y_i = 1|\boldsymbol{x_i})}{1 - \mathrm{P}(Y_i = 1|\boldsymbol{x_i})}\right) = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi} \quad (2.10)$$

$$P(Y_i = 1|\boldsymbol{x_i}) = \frac{e^{\beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi}}}{1 + e^{\beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi}}} \quad (2.11)$$

şeklinde ifade edilir.

Sürekli artan yığılımlı dağılım fonksiyonu $F(u) = 1/\big(1 + exp(-u)\big)$ kullanılarak $\mathrm{P}(Y_i = 1|\boldsymbol{x_i})$ ifadesi;

$$\mathrm{P}(Y_i = 1|\boldsymbol{x_i}) = F(\boldsymbol{x_i'\beta}) \quad (2.12)$$

şeklinde gösterilir.

Burada;

n: birim sayısını,

i: 1, 2, …, n

$P(Y_i = 1|\boldsymbol{x_i})$: i. birimin incelenen kategoriye eşit olma olasılığını ya da incelenen olay ile ilgili pozitif cevap verme olasılığını,

β_0: bağımsız değişkenler sıfır değerini aldığında bağımlı değişkenin beklenen değerini, yani sabiti,

$\beta_1,\ \beta_2, \ldots, \beta_p$: bağımsız değişkenlerin regresyon katsayılarını,

$X_{1i}, X_{2i}, \ldots, X_{pi}$: i. birime ait bağımsız değişkenleri,

p : Bağımsız değişken sayısını,

j = 0, 1,2, …, p

e = 2,718281828 sayısını göstermektedir [8, 23, 26, 29].

Burada $F(u)$ fonksiyonunu seçmemizin iki önemli sebebi vardır [40].

1. Matematik açısından bakıldığında kesinlikle esnek ve kolayca kullanılabilen bir fonksiyon olması,
2. Klinik olarak rahat yorumlanabilir olması.

2.1.2. Lojistik Regresyon Yönteminde Parametre Tahminleri ve Modelinin Uygunluğu

Lojistik regresyon analizinde bilinmeyen parametreler olan β_0, β_1, …, β_p değerlerinin tahminlenmesi doğrusal regresyonda analizinde kullanılan "En Küçük Kareler" yöntemi ile değil en çok olabilirlik yöntemi ile yapılmaktadır [33, 36]. En küçük kareler yöntemi bağımlı değişkeni ikili olan veri yapısına uygulandığında tutarlı sonuçlar vermemektedir. Lojistik regresyon modelini tahminlemede bu yöntem en temel yöntemdir. Bu yöntemi kullanabilmek için ilk önce olabilirlik fonksiyonunu belirlemek gerekmektedir. Bu fonksiyon gözlemlenen veri setini kullanır ve bilinmeyen parametrelerin fonksiyonudur. Bu parametrelerin maksimum olabilirlik tahmin edicileri fonksiyonu maksimum yapan değerleridir [4, 15]. Yaygın olarak kullanılan diğer tahminleme yöntemleri ise yeniden ağırlıklandırılmış iteratif en küçük kareler yöntemi, minimum logit ki-kare yöntemidir [36].

2.1.3. En Çok Olabilirlik fonksiyonu

İkili lojistik regresyon yönteminde parametre tahminlerini elde etmek için kullanılan en çok olabilirlik fonksiyonu

$$l(\boldsymbol{\beta}) = \prod_{i=1}^{n} F(\boldsymbol{x_i'\beta})^{y_i} \, [1 - F(\boldsymbol{x_i'\beta})]^{1-y_i} \qquad (2.13)$$

olarak ifade edilir.

En çok olabilirlik fonksiyonunu maksimum yapan parametre değerleri en çok olabilirlik tahmini şeklinde tanımlanır. Matematiksel olarak bu eşitliğin logaritması ile çalışmak çok daha kolaydır. Bu işlemin yapılmasındaki amaç, fonksiyondaki çarpımları toplamlara dönüştürerek parametrelere göre kısmi türevlerin alınmasını kolaylaştırmaktır. Logaritması alınmış en çok olabilirlik fonksiyonuna logaritmik en çok olabilirlik fonksiyonu denir ve *log olabilirlik* olarak isimlendirilir. Aşağıdaki eşitlikteki gibi gösterilir.

$$L(\boldsymbol{\beta}) = ln[l(\boldsymbol{\beta})] = \sum_{i=1}^{n} \{y_i \; ln[F(\boldsymbol{x_i'\beta})] + (1 - y_i) \; ln[1 - F(\boldsymbol{x_i'\beta})]\} \quad (2.14)$$

Parametre tahminleri log olabilirlik fonksiyonunu maksimum yapan veya *d* fonksiyonunu minimum yapan değerler olarak elde edilir.

$$\widehat{\boldsymbol{\beta}} = \max_{\boldsymbol{\beta}}\{L(\boldsymbol{\beta})\} = \min_{\boldsymbol{\beta}}\left\{\sum_{i=1}^{n} d(\boldsymbol{x_i'\beta}; y_i)\right\} \quad (2.15)$$

$$d(\boldsymbol{x_i'\beta}; y_i) = -y_i \; ln[F(\boldsymbol{x_i'\beta})] - (1 - y_i) \; ln[1 - F(\boldsymbol{x_i'\beta})] \quad (2.16)$$

L($\boldsymbol{\beta}$) fonksiyonunu maksimum yapan $\boldsymbol{\beta}$ değerinin bulunabilmesi için, L($\boldsymbol{\beta}$)'nın $\beta_0, \beta_1, \ldots, \beta_p$'e göre kısmi türevlerinin alınıp 0'a eşitlenmesi gerekir. Log olabilirlik fonksiyonunun birinci kısmi türevi;

$$\frac{\partial[L(\boldsymbol{\beta})]}{\partial \, \boldsymbol{\beta}} = \sum_{i=1}^{n} x_{ji}[y_i - F(\boldsymbol{x_i'\beta})] = 0 \quad (2.17)$$

şeklinde hesaplanır.

GLR yöntemi için parametre tahminleri iteratif yöntemler kullanarak elde edilen değerlerdir [32]. Doğrusal regresyonda olabilirlik eşitlikleri açık çözümü olan doğrusal denklemlerdir. Fakat lojistik regresyonda bu ifadeler $\boldsymbol{\beta}$ vektörüne göre doğrusal olmayan (nonlineer) denklemlerdir. Bundan dolayı bu denklemlerin çözümü için özel iterasyon yöntemi kullanan nümerik analiz yöntemler gerekir. Log olabilirlik fonksiyonundan elde edilen $\boldsymbol{\beta}$'ların değeri, en çok olabilirlik tahmini olarak adlandırılır ve $\widehat{\boldsymbol{\beta}}$ olarak gösterilir [40].

Parametrelerin varyans kovaryans matrisi, Fisher Information matrisinin tersi alınarak hesaplanır ve aşağıdaki şekilde ifade edilir.

$$Var(\boldsymbol{\beta}) = I^{-1}(\boldsymbol{\beta}) \tag{2.18}$$

Burada;

I Fisher information matrisini göstermektedir ve log olabilirlik fonksiyonunu ikinci kısmi türevi alınarak hesaplanır ve eşitlik 2.19'daki gibi gösterilir.

$$I(\boldsymbol{\beta}) = \frac{\partial^2 L(\boldsymbol{\beta})}{\partial\beta_j\,\partial\beta_m} = -\sum_{i=1}^{n} x_{ji}x_{mi}\,F(\boldsymbol{x_i'\beta})\big(1 - F(\boldsymbol{x_i'\beta})\big) \tag{2.19}$$

j, m = 0, 1, 2, …, p.

Parametre tahminlerinin varyans kovaryans matrisi ise aşağıdaki şekilde ifade edilir [9].

$$\widehat{Var}(\widehat{\boldsymbol{\beta}}) = \hat{I}^{-1}(\widehat{\boldsymbol{\beta}}) \tag{2.20}$$

$$\hat{I}(\widehat{\boldsymbol{\beta}}) = \boldsymbol{X'VX} \tag{2.21}$$

$\hat{F}(\boldsymbol{x_i'}\widehat{\boldsymbol{\beta}}) = \hat{\pi}_i$ olsun. Bu durumda;

$$\boldsymbol{X} = (\boldsymbol{x_1}, \dots, \boldsymbol{x_n})' = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix}_{n\times(p+1)} \tag{2.22}$$

$$\boldsymbol{V} = \begin{bmatrix} \hat{\pi}_1(1-\hat{\pi}_1) & 0 & \dots & 0 \\ 0 & \hat{\pi}_2(1-\hat{\pi}_2) & \dots & 0 \\ \vdots & 0 & \ddots & \vdots \\ 0 & \dots & 0 & \hat{\pi}_n(1-\hat{\pi}_n) \end{bmatrix}_{n\times n} \tag{2.23}$$

$\widehat{SE}\,(\hat{\beta}_j)$, $\widehat{Var}(\widehat{\boldsymbol{\beta}})$ matrisinin *j*. köşegen elemanının karekökünе eşittir.

2.1.4. Katsayıların Önemliliğinin Testi

Tahmin edilen $\boldsymbol{\beta}$ katsayılarının önemliliği, H_0: $\boldsymbol{\beta}$=0 hipotezinin test edilmesiyle belirlenir. Bu amaçla ileri sürülen ve yaygın olarak kullanılan testler; [36].

1. Olabilirlik oran (Likelihood Ratio) testi,
2. Wald test istatistiği,
3. Skor test istatistiği.

2.1.4.1. Olabilirlik Oran Testi

P tane bağımsız değişken içeren regresyon modelinde, oluşturulan iki farklı modelden, birinci modelde v tane bağımsız değişken, diğer modelde ise p = v + m tüm bağımsız değişkenler modelde olsun. İki regresyon modelinde de bağımsız değişkenlerin katsayıları gösteriminde, birinci model için $\boldsymbol{\beta_v}$ vektörü, ikinci model içinse $\boldsymbol{\beta_{v+m}}$ vektörü kullanılsın. $x_{v+1}, x_{v+2}, \dots, x_{v+m}$ bağımsız değişkenlerin katsayılarının sıfıra eşit olup olmadığını eşanlı (simültane) olarak test eden benzerlik oran test istatistiği;

$$LR = -2\left(ln\left[\frac{l(\boldsymbol{\beta_v})}{l(\boldsymbol{\beta_{v+m}})}\right]\right) = -2\left[\left[ln[l(\boldsymbol{\beta_v})] - ln[l(\boldsymbol{\beta_{v+m}})]\right]\right] \qquad (2.24)$$

şeklinde ifade edilir.

Burada;

$l(\boldsymbol{\beta_v})$: *v* tane bağımsız değişken içeren birinci model için en çok olabilirlik fonksiyonunu,

$l(\boldsymbol{\beta_{v+m}})$: *p* tane bağımsız değişken içeren ikinci model için en çok olabilirlik fonksiyonunu göstermektedir.

Olabilirlik oran istatistiği (v + m) – v = m serbestlik dereceli ki-kare dağılımı gösterir [11, 26].

2.1.4.2. Wald Testi

Wald test istatistiği, parametre tahminin ($\widehat{\boldsymbol{\beta}}$) standart hatasına [SE($\widehat{\boldsymbol{\beta}}$)] bölünmesi ile bulunur. j'ninci parametre için Wald test istatistiği aşağıdaki şekilde elde edilir.

$$w_j = \frac{\hat{\beta}_j}{\widehat{SE}\ (\hat{\beta}_j)} \qquad (2.25)$$

Wald test istatistiği standart normal dağılım gösterir (SND). Wald test istatistiğinin karesi 1 serbestlik dereceli kikare dağılımı gösterir. Parametrelerin önemliliklerinin test edilmesinde her iki dağılımdan da yararlanılır.

Wald testinin çok değişkenli gösterimi aşağıdaki şekilde hesaplanır.

$$w = \widehat{\boldsymbol{\beta}}'\left[\widehat{Var}\left(\widehat{\boldsymbol{\beta}}\right)\right]^{-1}\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}'(\boldsymbol{X}'\boldsymbol{V}\boldsymbol{X})\widehat{\boldsymbol{\beta}} \qquad (2.26)$$

Modelde yer alan her bir regresyon katsayının (p+1 tane) 0'a eşit olduğunu gösteren H_0 hipotezi altında p+1 serbestlik dereceli ki kare dağılımı gösterir.

2.1.4.3. Skor Testi

Skor testi log olabilirlik fonksiyonunun türevlerinin dağılım teorisine dayanmaktadır.

Log olabilirlik fonksiyonun birinci ve ikinci türevlerinden oluşan matrisden elde edilir. Aşağıdaki gibi hesaplanır.

$$U(\boldsymbol{\beta}) = \frac{\partial\, ln[l(\boldsymbol{\beta})]}{\partial\, \boldsymbol{\beta}} \qquad (2.27)$$

$$ST = \widehat{U}'\left(\widehat{\boldsymbol{\beta}}\right)\left[-\widehat{\boldsymbol{I}}^{-1}\left(\widehat{\boldsymbol{\beta}}\right)\right]\widehat{U}\left(\widehat{\boldsymbol{\beta}}\right) \qquad (2.28)$$

ST istatistiği p serbestlik dereceli ki kare dağılımı göstermektedir.

2.1.5. Güven Aralığının Tahmin Edilmesi

Parametre tahminlerinin önemliliğinde test etmekte kullandığımız yöntemler ile güven aralığı tahminlemesinde kullandığımız temel yapı aynı istatistik teorileri kullanmaktadır. j'ninci parametre için 100(1-α)% güven aralığı [39];

$$\hat{\beta}_j \mp z_{1-\alpha/2} SE\left(\hat{\beta}_j\right) \qquad (2.29)$$

Burada;

α : I. tip hatayı,

$z_{1-\alpha/2}$: I. tip hata değerine göre standart normal dağılıma ait kritik değeri göstermektedir.

2.1.6. En Çok Olabilirlik Fonksiyonunun Zayıf Sonuçlar Verdiği Durumlar

En çok olabilirlik fonksiyonunun parametre tahminlerinde zayıf sonuçlar verdiği durumlar genel olarak küçük, seyrek, tam ayrımsama, heterojen veri setlerinde karşımıza çıkmaktadır.

Mehta ve Patel'in (1995) çalışmalarında, en çok olabilirlik fonksiyonunun küçük ya da seyrek veri setlerinde yansız, tutarlı, etkili, yeterli ve minimum varyanslı parametre tahminleri vermediğini göstermişlerdir [10, 35].

Çok yaygın olarak bilinen diğer bir durum ise veri setinde tam ayrımsamanın meydana geldiği zamanlardır [10]. Basit bir örnek üzerinde gösterecek olursak, tek bağımsız değişken içeren lojistik regresyon modeli düşünelim. Burada Y=0 olduğu durumda bağımsız değişkenin X, $12 \leq X \leq 18$ aralığında; Y=1 olduğu durumda bağımsız değişkenin X, $22 \leq X \leq 29$ aralığında gözlendiğini varsayalım. Bu durum *Tam Ayrımsama* olarak nitelendirilir çünkü bağımsız değişkenin aldığı değerler bağımlı değişkene göre ayrılmış durumdadır ve üst üste çakışma (Overlapping) meydana gelmemektedir. Eğer bağımsız değişkenin Y=0 olduğu durumdaki aralığı $12 \leq X \leq 22$ şeklinde olursa bu durum da quasi-complete ayrımsama olarak tanımlanmaktadır. Tam ayrımsamanın veya quasi-complete ayrımsamanın meydana geldiği durumlarda en çok olabilirlik fonksiyonu parametre tahminleri vermemektedir [10].

Bağımlı değişkenin heterojen dağılım gösterdiği durumlar bağımlı değişkenin sapan değerler içerdiği durumlardır. Bir gözlemi sapan değer olarak tanımlamak için kullanılan iki yaklaşım vardır. Birincisi, geometrik bir yaklaşımdır ve diğer gözlemlerden oldukça uzakta yer alan bir değeri sapan bir değer olarak ifade eder. İkincisi, olasılıksal bir yaklaşımdır ve aşağıdaki gibi tanımlanır;

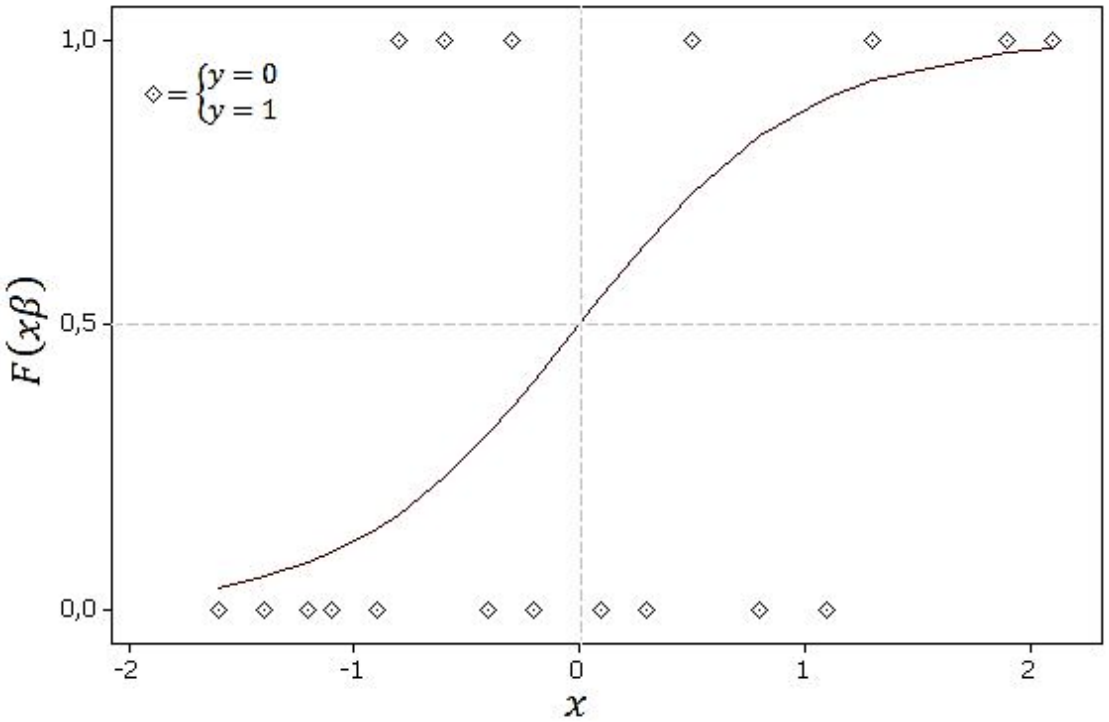
1. Kurulan model doğru ve bağımlı değişken Y=1 olarak gözlemlenmiş olsun. Fakat Y'nin 1 olarak gözlenme olasılığı kurulan modele göre oldukça düşük $P(Y_i = 1 \,|\boldsymbol{x_i}) \cong 0$ ise sapan bir değer gözlemlenmiş olur.
2. Kurulan model doğru ve bağımlı değişken Y=0 olarak gözlemlenmiş olsun. Fakat Y'nin 0 olarak gözlenme olasılığı kurulan modele göre

oldukça yüksek $P(Y_i = 0 \,|\boldsymbol{x_i}) \cong 1$ ise sapan bir değer gözlemlenmiş olur.

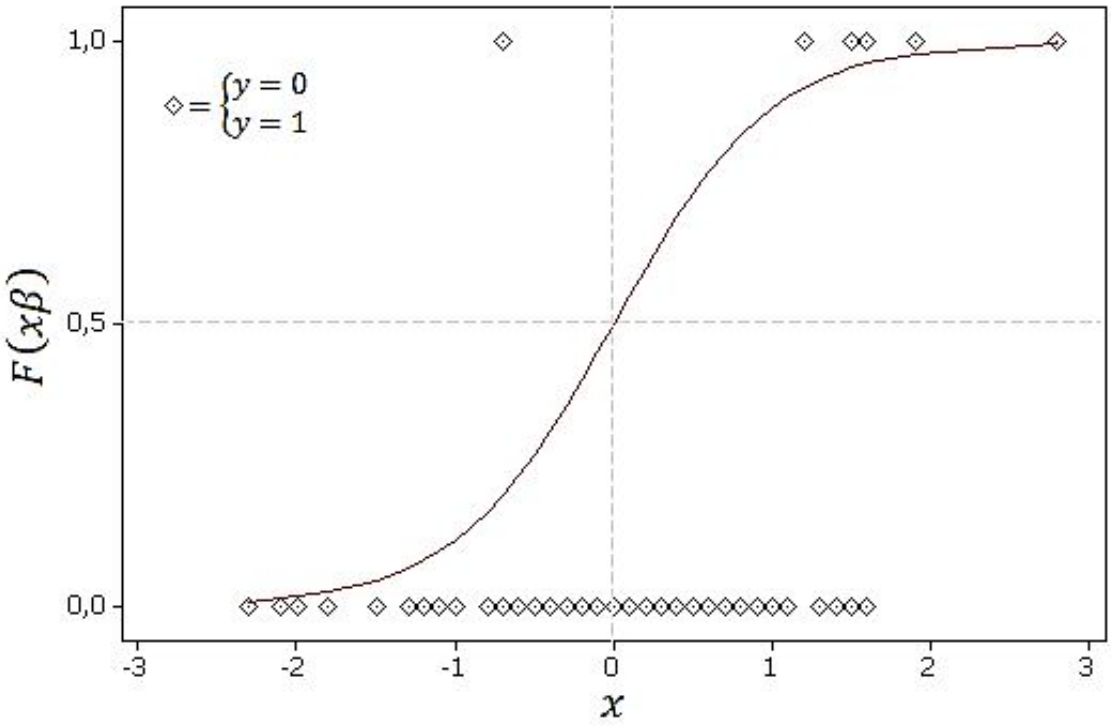
Yukarıda belirtilen tanımlarda bağımsız değişkenin değerinin yüksek olduğu durumlarda bağımlı değişkenin 1 olarak gözlemlendiği, bağımsız değişkenin değerinin düşük olduğu durumlarda ise bağımlı değişkenin 0 olarak gözlemlendiği kabul edilmiştir.

Heterojen bağımlı değişken içeren veri setleri için kurulan lojistik regresyon modelinin parametre tahminleri en çok olabilirlik fonksiyonu ile yapıldığında yansız, tutarlı, etkili, yeterli ve minimum varyanslı parametre tahminleri elde edilmeyebilir [7].

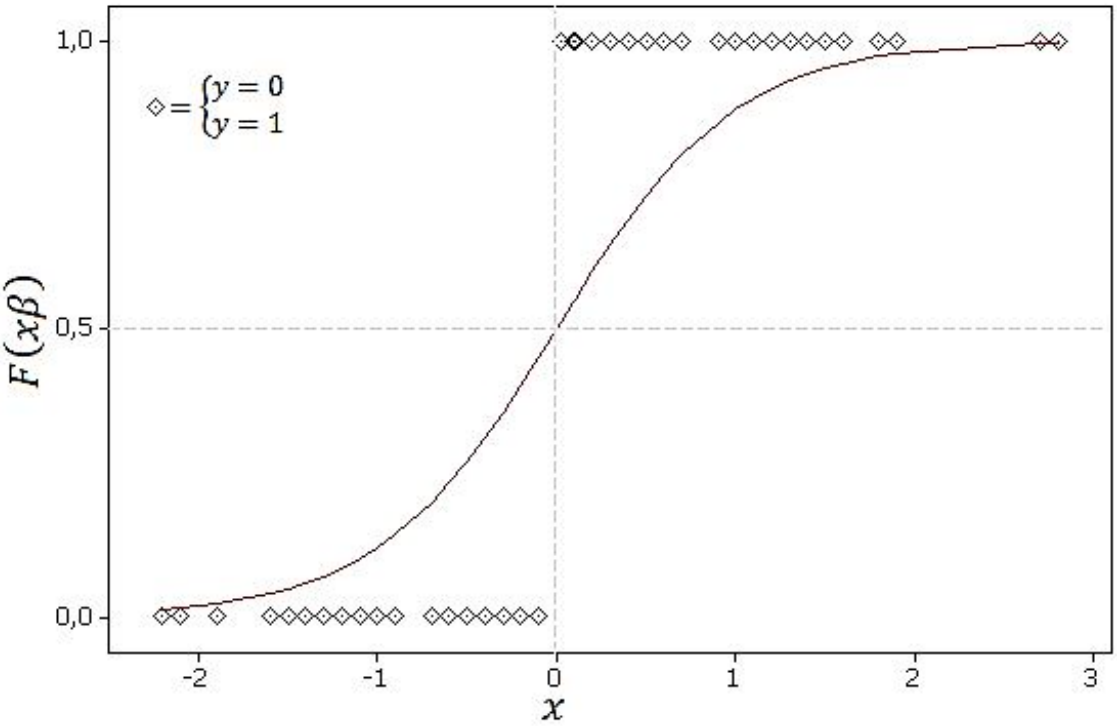
Yukarıda bahsedilen en çok olabilirlik fonksiyonunun zayıf sonuçlar verdiği durumlara örnek olarak verilen veri setlerine ilişkin olasılık dağılımı ve bağımlı değişkenin gözlenen değerleri şekil 2.1, 2.2, 2.3, 2.4 gösterilmiştir.



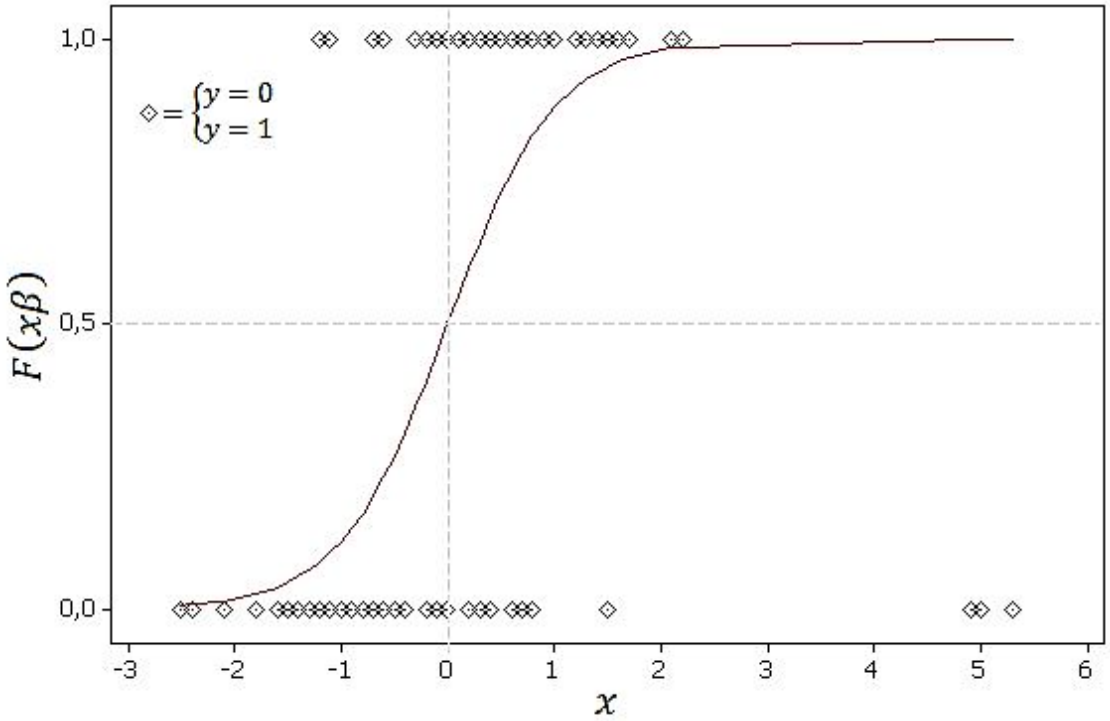
Şekil 2.1. Küçük örnek büyüklüğü olduğu veri seti.



Şekil 2.2. Bağımlı değişkenin seyrek gözlendiği veri seti



Şekil 2.3. Tam ayırsamanın meydana geldiği veri seti



Şekil 2.4. %5 bozularak heterojen hale gelmiş veri seti

2.2. Sağlam Lojistik Regresyon

Sağlam yöntemlerin temeli Huber (1981), Hampel (1986), Ronchetti, Rousseeuw ve Stahel (1986) tarafından yapılan çalışmalar sonucunda atılmıştır. Wilcox'ın (1998) sosyal bilimlerde ve psikoloji biliminde sağlam yöntemlerin kullanılmasında daha sistematik yollar açmıştır [13].

Croux ve Haesbroeck (2003), Bianco ve Yohai (1996) tarafından ortaya atılan SLR yöntemini modifiye ederek diğer SLR yöntemlerine göre hızlı ve stabil sonuç veren bir algoritma geliştirmişlerdir. Bundan dolayı bu tez çalışmasında, Croux ve Haesbroeck (2003) tarafından geliştirilen bu algoritma SLR yöntemi olarak kullanılmıştır [7].

Sağlam yöntem için parametre tahminleri Croux ve Haesbroeck (2003) tarafından modifiye edilerek geliştirilen en çok olabilirlik fonksiyonu ve iteratif algoritma kullanılarak elde edilir [32].

Sağlam yöntem için oluşturulan en çok olabilirlik fonksiyonu; [12]

$$\hat{\beta} = \min_{\beta} \sum_{i=1}^{n} \{H(d(\boldsymbol{x_i'\beta};\ y_i)) + Q(\boldsymbol{x_i'\beta})\} \qquad (2.30)$$

olarak tanılanır.

Burada Q fonksiyonu yanlılık düzeltme terimi olarak tanımlanmaktadır. Aşağıdaki şekilde hesaplanır.

$$Q(\boldsymbol{x_i'\beta}) = G(F(\boldsymbol{x_i'\beta})) + G(1 - F(\boldsymbol{x_i'\beta})) - G(1) \qquad (2.31)$$

Burada fonksiyonda yer alan H ve G fonksiyonları;

$$H(t) = \begin{cases} t\, e^{-\sqrt{k}} & eğer\ t \le k \\ -2\, e^{-\sqrt{t}}\left(1+\sqrt{t}\right) + e^{-\sqrt{k}}\left(2\left(1+\sqrt{k}\right)+k\right) & diğer\ durumlarda \end{cases} \qquad (2.32)$$

$$G(t) = \begin{cases} te^{-\sqrt{-\ln(t)}} + e^{1/4}\sqrt{\pi}\ \phi\left(\sqrt{2}\left(1/2+\sqrt{-\ln(t)}\right)\right) - e^{-1/4}\sqrt{\pi} & eğer\ t \le e^{-k} \\ e^{-\sqrt{k}}t - e^{-1/4}\sqrt{\pi}\ \phi\left(\sqrt{2}(1/2+\sqrt{k})\right) & diğer\ durumlarda \end{cases} \qquad (2.33)$$

Burada;

ϕ normal kümülatif dağılımı göstermektedir [1]. Croux ve Haesbroeck (2003) çalışmalarında *k* değeri olarak 0.5 kullanılmasını önermişlerdir.

2.3. Kesin Lojistik Regresyon

KLR analizinde parametre tahminleri için koşullu olabilirlik fonksiyonu kullanılır. Bu yaklaşım, koşullu lojistik regresyon yönteminde kullanılan olabilirlik fonksiyonu ile aynıdır. Fakat KLR analizinde parametrelerin önemlilikleri için elde edilen p değerleri genel yaklaşım ile değil tüm olası permütasyonlar hesaplanarak elde edilmektedir. Bu hesaplamaların nasıl yapıldığı "Kesin Koşullu Dağılım" bölümünde açıklanmıştır [25, 42].

Koşullu olabilirlik yaklaşımının uygulanması, öncelikle parametreler için yeterli istatistiklerin hesaplanmasına bağlıdır. Örneğin koşulsuz olabilirlik fonksiyonundaki her bir β_j parametresi için yeterli istatistik [27],

$$T_j = \sum_{i=1}^{n} y_i \, x_{ji} \tag{2.34}$$

şeklinde hesaplanmaktadır.

Her bir parametre için elde edilen yeterli istatistikler vektörü $\boldsymbol{T} = (T_0, T_1, \dots, T_P)'$ şeklinde elde edilir.

Elde edilen yeterli istatistikler için olasılık yoğunluk fonksiyonu,

$$P(\boldsymbol{T} = \boldsymbol{t}) = \frac{C(\boldsymbol{t})\exp(\boldsymbol{t}'\boldsymbol{\beta})}{\prod_{i=1}^{n}[1 + \exp(\boldsymbol{x}_i'\boldsymbol{\beta})]} \tag{2.35}$$

olarak belirlenir.

Burada,

$$C(\boldsymbol{t}) = \|\{\boldsymbol{y} : \ \boldsymbol{y}'\boldsymbol{X} = \boldsymbol{t}\}\| \tag{2.36}$$

t'yi oluşturan olası tüm y'lerin sayısını göstermektedir.

KLR yönteminde sabit β_0 önemsiz (nuisance) parametre olarak kabul edilir. Önemsiz parametre için elde edilen yeterli istatistik T_0 şeklinde gösterilir. Analizin gerçekleştirileceği diğer p değişken için parametre ve yeterli istatistik $\boldsymbol{\beta}_p = (\beta_1, \dots, \beta_p)'$ ve $\boldsymbol{T}_p = (T_1, \dots, T_p)$ şeklinde gösterilir.

Koşullu olabilirlik fonksiyonu, önemsiz parametrenin yeterli istatistikleri üzerine koşullanarak oluşturulur ve önemsiz parametre analizden çıkartılır. Bu durumda koşullu olabilirlik fonksiyonu,

$$P\big(\boldsymbol{T}_p = \boldsymbol{t}_p \big| \boldsymbol{T}_0 = \boldsymbol{t}_0\big) = \frac{P(\boldsymbol{T} = \boldsymbol{t})}{P(\boldsymbol{T}_0 = \boldsymbol{t}_0)} = \frac{C(\boldsymbol{t})\exp(\boldsymbol{t}_p'\boldsymbol{\beta}_p)}{\sum_u C(\boldsymbol{u}, \boldsymbol{t}_0)\exp(\boldsymbol{u}'\boldsymbol{\beta}_p)} = f_{\boldsymbol{\beta}_p}(\boldsymbol{t}_p|\boldsymbol{t}_0) \tag{2.37}$$

şeklinde elde edilir [24].

Burada;

$C(\boldsymbol{u}, \boldsymbol{t}_0)$, $\boldsymbol{y}'\boldsymbol{X}_p = \boldsymbol{u}$ ve $\boldsymbol{y}'\boldsymbol{X}_0 = \boldsymbol{t}_0$ eşitliklerini sağlayan olası $\boldsymbol{y}$ vektörlerinin sayısını göstermektedir [22, 34]. Parametre tahminleri, koşullu olabilirlik fonksiyonunu maksimum yapan değerler olarak elde edilir. Bu yaklaşım koşulsuz olabilirlik

fonksiyonunu maksimize eden yaklaşım ile aynıdır. Burada da yine açık bir çözüm yapılamadığından koşulsuz olabilirlik fonksiyonunu maksimize eden parametre tahminleri nümerik analizler yardımı ile elde edilmektedir. Ancak parametre tahminlerinin anlamlılıkları için p değerlerinin hesaplanmasında genel yaklaşım kullanılmamaktadır. Kesin çıkarsama ya da parametre tahminlerinin anlamlılıkları için kullanılacak p değerlerinin hesaplanması, ilgilenilen parametreler için oluşturulan koşullu dağılıma bağlıdır. Bu dağılım kesin koşullu dağılım (Exact Conditional Distribution) olarak adlandırılır. Parametre tahminlerine ait standart hatalar kesin koşullu dağılım kullanıldığından dolayı hesaplanmamaktadır [24, 41].

2.3.1. Kesin Koşullu Dağılım

Kesin koşullu dağılım analizinin amacı, gözlenmesi olası 2^n tane bağımlı ikili değişken vektörlerinin dağılımını belirleyerek, $\boldsymbol{y}'\boldsymbol{X}_0 = \boldsymbol{t}_0$ eşitliğini sağlayanların sayısını elde etmektir [14, 34]. Örneğin, n=4 gözlem içeren bir veri seti aşağıdaki gibi olsun.

Tablo 2.1. n=4 gözlem içeren ikili yapıdaki bir veri seti

Gözlem	y_0	x_0	x_1
1	0	1	1
2	1	1	1
3	0	1	2
4	1	1	0

Burada gözlenen bağımlı ikili değişken $\boldsymbol{y}_0 = (0,1,0,1)'$, bağımsız değişkenler $\boldsymbol{x}_0 = (1,1,1,1)'$ ve $\boldsymbol{x}_1 = (1,1,2,0)'$ olarak görülmektedir. Gözlenen değerlerden hesaplanan yeterli istatistik vektörü ise,

$$\boldsymbol{t} = (t_0, t_1) = 0 \times (1,1) + 1 \times (1,1) + 0 \times (1,2) + 1 \times (1,0) = (2,1)$$

şeklinde hesaplanır. Bu durumda tüm dağılımdan $\boldsymbol{t_0 = 2}$ üzerine koşullanarak hesaplamalar yapılır. Örnek genişliğimiz n=4 olduğundan tüm olası ikili bağımlı değişken ve bu değişken değerlerinden hesaplanan $\boldsymbol{t_0}$ ve $\boldsymbol{t_1}$ değerleri aşağıda gösterildiği gibi elde edilir.

Tablo 2.2. n=4 olduğundan tüm olası ikili bağımlı değişken ve bu değişken değerlerinden hesaplanan $\boldsymbol{t_0}$ ve $\boldsymbol{t_1}$ değerleri

	y_1	y_2	y_3	y_4	t_0	t_1
1	0	0	0	0	0	0
2	0	0	0	1	1	0
3	0	0	1	0	1	2
4	0	0	1	1	2	2
5	0	1	0	0	1	1
6	0	1	0	1	2	1
7	0	1	1	0	2	3
8	0	1	1	1	3	3
9	1	0	0	0	1	1
10	1	0	0	1	2	1
11	1	0	1	0	2	3
12	1	0	1	1	3	3
13	1	1	0	0	2	2
14	1	1	0	1	3	2
15	1	1	1	0	3	4
16	1	1	1	1	4	4

Buradan koşullu dağılım, $\boldsymbol{t_0 = 2}$ olanları çıkartarak aşağıdaki gibi elde edilir.

Tablo 2.3. $\boldsymbol{t}_0 = 2$ olduğu durumlarda elde edilen koşullu dağılım

t_0	t_1	Frekans	Olasılık
2	1	2	2/6
2	2	2	2/6
2	3	2	2/6
Toplam		6	1

Burada β_j parametresi için kullanılacak p değeri yukarıdaki koşullu dağılım tablosundan elde edilir. Gözlenen değerlerden elde edilen yeterli istatistik $t_1 = 1$'e eşit ve küçük olan değerler için hesaplanan olasılıklar toplanır. Aynı hesaplama $t_1 = 1$'e eşit ve büyük olan değerler için de yapılır. Bu iki toplam olasılıklardan küçük olan toplam 2 ile çarpılarak β_1 parametresi için iki yönlü hipotez testinde kullanılacak p değeri elde edilir. Bu örnekte küçük olan toplam 2/6, büyük olan toplam ise 1 olduğundan β_1 parametresi için elde edilen kesin p değeri $p = 2 \times \frac{2}{6} = \frac{4}{6} = 0.667$olarak elde edilir.

Genel olarak önemsiz parametreler dışında kalan, analizlerin gerçekleştirileceği parametrelerden herhangi bir β_j parametresi için hipotez testleri ve bu testlere ilişkin kesin p değerlerinin hesaplanması aşağıda gösterilmiştir [24, 34].

$H_0: \beta_j = 0$

$H_1: \beta_j < 0 \quad P_L(t_j; 0) = \sum_{u \le t_j} f_0(u|\boldsymbol{t}_0)$

$H_1: \beta_j > 0 \quad P_G(t_j; 0) = \sum_{u \ge t_j} f_0(u|\boldsymbol{t}_0)$

$H_1: \beta_j \neq 0 \quad P(t_j; 0) = 2\, min\{P_L(t_j; 0), P_G(t_j; 0)\}$

3. GEREÇ VE YÖNTEM

İkili yapıda bağımlı değişken içeren heterojen veri setlerinin analizinde kullanılan GLR, SLR ve KLR yöntemlerinin karşılaştırılmasında 10,000 tekrarlı Monte Carlo simülasyon yöntemi kullanıldı. Yöntemlerin performans karşılaştırmaları aşağıdaki temel ölçütler kullanılarak yapıldı. Bu ölçütler,

- Parametre tahminleri ve yanlılıkları,
- Parametre tahminlerinin standart hataları

olarak belirlendi.

3.1. Simülasyon Çalışması

Simülasyonda kullanılmak üzere tek açıklayıcı değişken içeren lojistik regresyon modeli kullanıldı.

$$Logit[\,\pi(x_i)\,] = \beta_0 + \beta_1 x_i + e_i \qquad (3.1)$$

Burada β_0 sabiti, β_1 bağımsız değişken katsayısını göstermektedir. Modelde yer alan e_i rassal değişkeni, yer parametresi 0 ve ölçek parametresi 1 olan lojistik dağılıma sahip rasgele etkiyi göstermektedir ve x_i açıklayıcı değişkeninden bağımsızdır.

$$e_i \sim Logistic\ (0,1) \qquad (3.2)$$

3.1.1. Simülasyon Algoritması

Yöntemlerin karşılaştırılmasında kullanılan verilerin türetimi için kullanılan model (3.1)'dir. Türetim için uygulanan adımlar aşağıdaki gibidir;

1. β_0 ve β_1 için sabit birer değer atandı.
2. Heterojen veri seti elde etmek için %S bozulma durumları belirlendi. Burada S, veri setinin % kaçının heterojen yapıda olduğunu göstermektedir. Homojen veri seti için S=0 olarak belirlendi.
3. $n=n_1+n_2$ gözlem içeren veri seti için n_1=%(100-S)×n tane bağımsız değişken x_i, 0 ortalamalı ve 1 varyanslı standart normal dağılımdan türetildi $x_i \sim N(0,1)$.

4. $n=n_1+n_2$ gözlem içeren veri seti için n_2=%S×n tane bağımsız değişken x_i, minimum değeri 4.5, maksimum değeri 5.5 olan sürekli uniform dağılımdan türetildi. $x_i \sim U(4.5, 5.5)$. Bu sayede $5 \times \sqrt{p} = 5 \times \sqrt{1} = 5$ olacak şekilde aşırı yüksek değerler içeren bağımsız değişkenler türetildi. Burada $5 \times \sqrt{p}$ yüksek derecede bozulma sağlamak için kullanılmıştır. Bu eşitlik Croux ve Haesbroeck (2003) tarafından önerilmektedir.
5. n tane e_i rassal değişkeni, yer parametresi 0 ve ölçek parametresi 1 olan lojistik dağılımdan türetildi. $e_i \sim Logistic\ (0, 1)$
6. n_1 tane bağımlı değişken aşağıdaki şekilde türetildi.

$$y_i = \begin{cases} 0 & eğer\ \beta_0 + \beta_1 x_i + e_i \leq 0 \\ 1 & eğer\ \beta_0 + \beta_1 x_i + e_i > 0 \end{cases} \qquad (3.3)$$

   i=1, 2, …, n_1
7. n_2 tane bağımlı değişken ise

$$y_i = \begin{cases} 0 & eğer\ x_i > 4.4999 \\ y_i & eğer\ x_i \leq 4.4999 \end{cases} \qquad (3.4)$$

   i=1, 2, …, n_2
8. Bağımlı ve bağımsız değişkenler elde edildikten sonra bu değişkenler kullanılarak GLR, SLR ve KLR analizleri aynı verilerde uygulandı. Her üç analizden elde edilen parametre tahminleri ile GLR ve SLR yöntemlerinden elde edilen parametre tahminlerine ilişkin standart hatalar kaydedildi.

3., 4., 5., 6., 7. ve 8. adımlar 10,000 kez tekrarlandı. Böylece her bir yöntemden 10,000 tane parametre tahmini ve standart hata elde edildi.

3.1.2. Simülasyon Parametreleri

Simülasyon çalışmasında örneklem büyüklüğü n=100, 200, 300, 400, 500 olarak belirlendi.

Algoritmanın 1. adımında bahsedilen katsayılar için $\beta_0 = 0$, $\beta_1 = 2$ değerleri atandı ve yanlılıklar ile standart hataların elde edilmesinde kullanıldı.

Homojen veri setini oluşturmak için S bozulma durumları S=%0 olarak belirlendi.

Heterojen veri setini oluşturmak için S bozulma durumları S=%1, %2, %3, %4 ve %5 olarak belirlendi.

3.1.3. Karşılaştırma Ölçütleri

Yöntemlerin karşılaştırılmasında aşağıda bahsedilen karşılaştırma ölçütleri kullanılarak yapıldı.

Parametre tahminleri: Her bir yöntem için, 10,000 farklı parametre tahminin ortalaması alındı $(\overline{\hat{\beta}}_0, \overline{\hat{\beta}}_1)$. Bu ortalama tahmin değerlerinin, algoritmanın 1. adımda belirtilen $\beta_0 = 0$, $\beta_1 = 2$ parametreleri için verilen değere ne kadar yakınsadığı saptandı.

Parametre yanlılıkları: Parametre için verilen değer ile ortalama arasındaki fark alınarak her bir yöntem için yanlılıklar hesaplandı ve bu değerler kullanılarak yöntemler karşılaştırıldı $\left(\overline{\hat{\beta}}_0 - \beta_0\right), \left(\overline{\hat{\beta}}_1 - \beta_1\right)$.

Parametre tahminlerinin standart hataları: GLR ve SLR yöntemleri için, 10,000 farklı parametre tahmininin standart hatasının ortalaması alındı. Bu ortalama değerler kullanılarak hangi yöntemin nasıl bir varyasyonla parametre tahmininde bulunduğu karşılaştırıldı.

3.2. Simülasyon ve Analizlerde Kullanılan Programlar

Simülasyon çalışmalarında yer alan analizler R v2.13.2 ve SAS 9.0 paket programlarında gerçekleştirildi. Grafikerler Minitab 15.0 paket programında hazırlandı. Veri setlerinin türetilmesinde R ve SAS programlama dillerinden yararlanıldı.

GLR ve SLR yöntemlerinin analizleri R paket programında gerçekleştirildi. GLR analizleri için R paket programında GLM paketi kullanıldı. SLR analizleri için Bianco ve Yohai tarafından geliştirilen ve geçerliliği kabul edilmiş olan SLR yöntemi (BY tahmin edicileri) kullanıldı.

Örneğin aşağıdaki program heterojen veri yapılarında GLR ve SLR yöntemlerinin karşılaştırılması için R paket programında yazılmış olup,

n= 500

S= %5 olarak belirlendi.

β_0 parametresi ve β_1 parametresi için atanan değerler; $\beta_0 = 0$ ve $\beta_1 = 2$

olarak tanımlandı.

Yukarıdaki örneğe uygun olarak 10,000 farklı veri seti oluşturan ve bu veri setlerini kullanarak GLR ile birlikte Bianco ve Yohai tarafından geliştirilen SLR yöntemi sonuçlarının ortalamasını veren R paket programı kodları aşağıda yer almaktadır.

```
#  Computation of the estimator of Bianco and Yohai (1996) in logistic regression
#  -------------
#  Christophe Croux, Gentiane Haesbroeck, and Kristel Joossens
#
#  This program computes the estimator of Bianco and Yohai in
#  logistic regression. By default, an intercept term is included
#  and p parameters are estimated.
#  For more details we refer to
#     Croux, C., and Haesbroeck, G. (2003), ``Implementing the Bianco and Yohai estimator for
Logistic Regression'',
#    Computational Statistics and Data Analysis, 44, 273-295
#
#Input:
#-------
# x0= n x (p-1) matrix containing the explanatory variables;
# y= n-vector containing binomial response (0 or 1);
#
# initwml= logical value for selecting one of the two possible methods for computing
#        the initial value of the optimization process. If initwml=T (default), a
#        weighted ML estimator is computed with weights derived from the MCD estimator
#        computed on the explanatory variables. If initwml=F, a classical ML fit is perfomed.
#        When the explanatory variables contain binary observations, it is recommended
#        to set initwml to F or to modify the code of the algorithm to compute the weights
#        only on the continuous variables.
# const= tuning constant used in the computation of the estimator (default=0.5);
# kmax= maximum number of iterations before convergence (default=1000);
# maxhalf= max number of step-halving (default=10).
#
```

```
# Example:
# x0=matrix(rnorm(100,1))
# y0=numeric(runif(100)>0.5)
# BYlogreg(x0,y)
#
#Output:
#--------
# list with
# 1st component: T or F if convergence achieved or not
# 2nd component: value of the objective function at the minimum
# p next components: estimates for the parameters.
# p last components: standard errors of the parameters (if first component is T)

library(MASS)

BYlogreg<-function(x0,y,initwml=T,const=0.5,kmax=1e3,maxhalf=10)
{
 n=nrow(x0)
 p=ncol(x0)+1

 #Smallest value of the scale parameter before implosion
 sigmamin=1e-4

 x=as.matrix(cbind(rep(1,n),x0))
 y=as.numeric(y)

 # Computation of the initial value of the optimization process
 if (initwml==T)
 {
  hp=floor(n*(1-0.25))+1
  mcdx=cov.mcd(x0, quantile.used =hp,method="mcd")
  rdx=sqrt(mahalanobis(x0,center=mcdx$center,cov=mcdx$cov))
  vc=sqrt(qchisq(0.975,p-1))
  wrd=(rdx<=vc)
  gstart=glm(y~x0,family=binomial,subset=wrd)$coef
 }

else {gstart=glm(y~x0,family=binomial)$coef}

 sigmastart=1/sqrt(sum(gstart^2))
 xistart=gstart*sigmastart
 stscores=x %*% xistart
sigma1=sigmastart

 #Initial value for the objective function
 oldobj=mean(phiBY3(stscores/sigmastart,y,const))
 kstep=jhalf=1
```

```
  while ((kstep < kmax) & (jhalf<maxhalf)){


unisig <- function(sigma)

{ mean(phiBY3(stscores/sigma,y,const))}

optimsig=nlminb(sigma1,unisig,lower=0)
sigma1=optimsig$par

   if (sigma1<sigmamin) {print("Explosion");kstep=kmax
    } else {
    gamma1=xistart/sigma1
    scores=stscores/sigma1
    newobj=mean(phiBY3(scores,y,const))
    oldobj=newobj
    gradBY3=colMeans((derphiBY3(scores,y,const)%*%matrix(1,ncol=p))*x)
    h=-gradBY3+((gradBY3 %*% xistart) *xistart)
    finalstep=h/sqrt(sum(h^2))
    xi1=xistart+finalstep
    xi1=xi1/(sum(xi1^2))
    scores1=(x%*%xi1)/sigma1
    newobj=mean(phiBY3(scores1,y,const))

    ####stephalving
    hstep=jhalf=1
    while  ((jhalf <=maxhalf) & (newobj>oldobj)){
     hstep=hstep/2
     xi1=xistart+finalstep*hstep
     xi1=xi1/sqrt(sum(xi1^2))
     scores1=x%*%xi1/sigma1
     newobj=mean(phiBY3(scores1,y,const))
     jhalf=jhalf+1
    }

    if ((jhalf==maxhalf+1) & (newobj>oldobj)) {print("Convergence Achieved")
     } else {
     jhalf=1
     xistart=xi1
     oldobj=newobj
     stscores=x%*% xi1
     kstep=kstep+1
    }
   }
  }

  if (kstep == kmax) {
   print("No convergence")
```

```
   result=list(convergence=F,objective=0,coef=t(rep(NA,p)))
   } else {
   gammaest=xistart/sigma1
   stander=sterby3(x0,y,const,gammaest)
   result=list(convergence=T,objective=oldobj,coef=t(gammaest),sterror=stander)
  }
  return(result)
}

#####################################################################
#####################################################################
#Functions needed for the computation of estimator of Bianco and Yohai

phiBY3 <- function(s,y,c3)
{
  s=as.double(s)
  dev=log(1+exp(-abs(s)))+abs(s)*((y-0.5)*s<0)
  return(rhoBY3(dev,c3)+GBY3Fs(s,c3)+GBY3Fsm(s,c3))
}


rhoBY3 <- function(t,c3)
{
  (t*exp(-sqrt(c3))*as.numeric(t <= c3))+
    (((exp(-sqrt(c3))*(2+(2*sqrt(c3))+c3))-(2*exp(-sqrt(t))*(1+sqrt(t))))*as.numeric(t >c3))
}

psiBY3 <- function(t,c3)
{(exp(-sqrt(c3))*as.numeric(t <= c3))+(exp(-sqrt(t))*as.numeric(t >c3))}

derpsiBY3 <- function(t,c3)
{
res=NULL

  for (i in 1:length(t))

{
if (t[i] <= c3)

 { res=rbind(res,0) }

else

{res=rbind(res,-exp(-sqrt(t[i]))/(2*sqrt(t[i]))) }

}
res
```

```
}

sigmaBY3<-function(sigma,s,y,c3) {mean(phiBY3(s/sigma,y,c3))}

derphiBY3=function(s,y,c3)
{
 Fs= exp(-(log(1+exp(-abs(s)))+abs(s)*(s<0)))
 ds=Fs*(1-Fs)
 dev=log(1+exp(-abs(s)))+abs(s)*((y-0.5)*s<0)
 Gprim1=log(1+exp(-abs(s)))+abs(s)*(s<0)
 Gprim2=log(1+exp(-abs(s)))+abs(s)*(s>0)
 return(-psiBY3(dev,c3)*(y-Fs)+((psiBY3(Gprim1,c3)-psiBY3(Gprim2,c3))*ds))
}

der2phiBY3=function(s,y,c3)
{
 s=as.double(s)
 Fs= exp(-(log(1+exp(-abs(s)))+abs(s)*(s<0)))
 ds=Fs*(1-Fs)
 dev=log(1+exp(-abs(s)))+abs(s)*((y-0.5)*s<0)
 Gprim1=log(1+exp(-abs(s)))+abs(s)*(s<0)
 Gprim2=log(1+exp(-abs(s)))+abs(s)*(s>0)
 der2=(derpsiBY3(dev,c3)*(Fs-y)^2)+(ds*psiBY3(dev,c3))
 der2=der2+(ds*(1-2*Fs)*(psiBY3(Gprim1,c3)-psiBY3(Gprim2,c3)))
 der2=der2-(ds*((derpsiBY3(Gprim1,c3)*(1-Fs))+(derpsiBY3(Gprim2,c3)*Fs)))
 der2
}

GBY3Fs <- function(s,c3)
{
 Fs= exp(-(log(1+exp(-abs(s)))+abs(s)*(s<0)))
 resGinf=exp(0.25)*sqrt(pi)*(pnorm(sqrt(2)*(0.5+sqrt(-log(Fs))))-1)
 resGinf=(resGinf+(Fs*exp(-sqrt(-log(Fs)))))*as.numeric(s <= -log(exp(c3)-1))
 resGsup=((Fs*exp(-sqrt(c3)))+(exp(0.25)*sqrt(pi)*(pnorm(sqrt(2)*(0.5+sqrt(c3)))-
1)))*as.numeric(s > -log(exp(c3)-1))
 return(resGinf+resGsup)
}

GBY3Fsm <- function(s,c3)
{
 Fsm=exp(-(log(1+exp(-abs(s)))+abs(s)*(s>0)))
 resGinf=exp(0.25)*sqrt(pi)*(pnorm(sqrt(2)*(0.5+sqrt(-log(Fsm))))-1)
 resGinf=(resGinf+(Fsm*exp(-sqrt(-log(Fsm)))))*as.numeric(s >= log(exp(c3)-1))
 resGsup=((Fsm*exp(-sqrt(c3)))+(exp(0.25)*sqrt(pi)*(pnorm(sqrt(2)*(0.5+sqrt(c3)))-
1)))*as.numeric(s < log(exp(c3)-1))
 return(resGinf+resGsup)
}
```

```
sterby3 <- function(x0,y,const,estim)
{
  n=nrow(x0)
  p=ncol(x0)+1

  z=cbind(matrix(1,nrow=n),x0)
  argum=z %*% estim

  matM=matrix(data=0,nrow=p,ncol=p)
  IFsquar=matrix(data=0,nrow=p,ncol=p)
  for (i in 1:n)
{
myscalar=as.numeric(der2phiBY3(argum[i],y[i],const))
matM=matM+myscalar * (z[i,] %*% t(z[i,]))
IFsquar=IFsquar+myscalar^2 * (z[i,] %*% t(z[i,]))
}
  matM=matM/n
  matMinv=solve(matM)
  IFsquar=IFsquar/n
  asvBY=matMinv %*% IFsquar %*% t(matMinv)
  sqrt(diag(asvBY))/sqrt(n)
}

#######
r<-10000
n1<-475
n2<-25
n<-n1+n2
mu<-0
sigma<-1
location<-0
scale<-1
Beta0<-0
Beta1<-2
interceptmatrix_a<-rep(NA,r)
xmatrix_a<-rep(NA,r)
intercepthatamatrix_a<-rep(NA,r)
xhatamatrix_a<-rep(NA,r)

interceptmatrix_r<-rep(NA,r)
xmatrix_r<-rep(NA,r)
intercepthatamatrix_r<-rep(NA,r)
xhatamatrix_r<-rep(NA,r)

for(i in 1:r)
{
xn1=matrix(rnorm(n1,mu,sigma))
xn1<-round(xn1, 1)
```

```
xn2=matrix(runif(n2, min=4.5, max=5.5))
xn2<-round(xn2, 1)
x0=matrix(append(xn1,xn2))

e=matrix(rlogis(n, location, scale))
denk<-(2*x0)+e
y <- ifelse(denk<=0,0,1)

y <- ifelse(x0>4.49999,0,y)

asimp <- glm(y ~ x0, family=binomial("logit"))
asimpbeta<-summary(asimp)$coefficients
BYbeta<-BYlogreg(x0,y)$coef
BYhata<-BYlogreg(x0,y)$sterror

interceptmatrix_a[i]<-asimpbeta[1, 1]
xmatrix_a[i]<-asimpbeta[2, 1]
intercepthatamatrix_a[i]<-asimpbeta[1, 2]
xhatamatrix_a[i]<-asimpbeta[2, 2]

interceptmatrix_r[i]<-BYbeta[1, 1]
xmatrix_r[i]<-BYbeta[1, 2]
intercepthatamatrix_r[i]<-BYhata[1]
xhatamatrix_r[i]<-BYhata[2]

}

mean(interceptmatrix_a)
mean(xmatrix_a)
mean(intercepthatamatrix_a)
mean(xhatamatrix_a)

mean(interceptmatrix_r)
mean(xmatrix_r)
mean(intercepthatamatrix_r)
mean(xhatamatrix_r)
```

KLR yönteminin analizleri SAS paket programında gerçekleştirildi. Veri setlerinin türetilmesinde SAS programlama dilinden yararlanıldı. KLR analizleri için PROC LOGISTIC prosedürü kullanıldı.

Örneğin aşağıdaki program heterojen veri yapılarında KLR yönteminin karşılaştırılması için SAS paket programında yazılmış olup,

n= 500

S=%5 olarak belirlendi.

β_0 parametresi ve β_1 parametresi için atanan değerler; $\beta_0 = 0$ ve $\beta_1 = 2$

olarak kullanıldı. Yukarıdaki örneğe uygun olarak 10,000 farklı veri seti oluşturan SAS programlama kodları ve bu veri setlerini KLR analizinde kullanan LOGISTIC prosedürüne ait SAS kodları aşağıda yer almaktadır.

```
data kesindata;
n=500;

do j=1 to 10000;
do i=1 to n;

if i<=475 then x=Rand('NORMAL');
else x=Rand('UNIFROM')+4.5;

xr=round(x,0.1);
xu=Rand('UNIFORM');
e=LOG(xu / (1-xu));

denk=2*xr+e;
if denk<=0 then y=0;
else y=1;
if i>475 then y=0;

output;
end;
end;

ods listing close;
ods output ExactParmEst=pee;

proc logistic data = kesindata exactonly;
by j;
model y(event='1')=xr;
exact xr/estimate=parm;
run;

ods html body='c:\temp\output.htm';
proc sort data=pee;
by Parameter;
run;

proc means data=pee;
by Parameter;
title 'Exact';
run;
ods html close;
```

4. BULGULAR

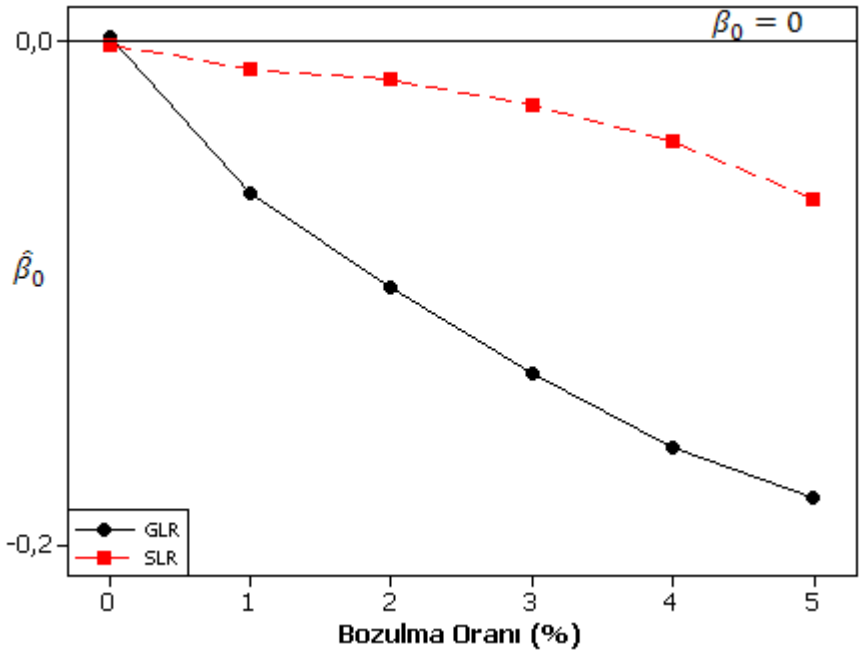
Yapılan simülasyon çalışmaları sonucunda GLR, SLR ve KLR yönteminden elde edilen parametre tahminleri, yanlılıkları (bias) ve standart hataları farklı örnek büyüklüğü ve farklı bozulma durumu düzeylerine göre tablolar ve şekiller aracılığıyla düzenlenerek açıklanmıştır. Bu üç yöntemin karşılaştırılmasında kullanılan ölçütlere göre de elde edilen bulgular sırasıyla aşağıda verilmiştir.

4.1. Parametre Tahminleri ve Yanlılıkları

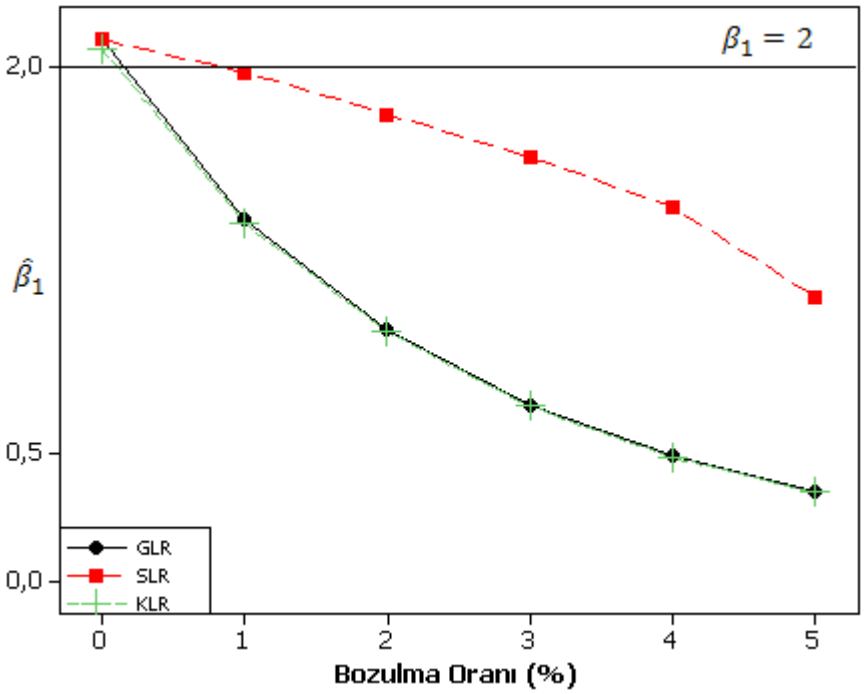
Yapılan simülasyonlar sonucunda her bir yöntemden elde edilen parametre tahminleri ve yanlılıkları farklı örnek büyüklüğü ve bozulma durumları düzeylerine göre Tablo 4.1, Tablo 4.2, Tablo 4.3, Tablo 4.4, Tablo 4.5'de verilmiştir. Tablo 4.1 – 4.5'de $\beta_0 = 0$, $\beta_1 = 2$ olup; Tablo 4.1'de sunulan bulgular n=100 örnek büyüklüğü, Tablo 4.2'de sunulan bulgular n=200 örnek büyüklüğü, Tablo 4.3'de sunulan bulgular n=300 örnek büyüklüğü, Tablo 4.4'de sunulan bulgular n=400 örnek büyüklüğü ve Tablo 4.5'de sunulan bulgular n=500 örnek büyüklüğü için elde edilen bulgulardır. Yanlılık değerleri $\hat{\beta}_j - \beta_j$ biçiminde hesaplandı.

Tablo 4.1. n=100 örnek büyüklüğü, $\beta_0 = 0$, $\beta_1 = 2$ değerleri için 10,000 Monte Carlo Simülasyonundan elde edilen parametre tahminleri ve yanlılıkları

S	İstatistiksel Yöntemler									
	GLR				SLR				KLR	
	$\hat{\beta}_0$	Bias($\hat{\beta}_0$)	$\hat{\beta}_1$	Bias($\hat{\beta}_1$)	$\hat{\beta}_0$	Bias($\hat{\beta}_0$)	$\hat{\beta}_1$	Bias($\hat{\beta}_1$)	$\hat{\beta}_1$	Bias($\hat{\beta}_1$)
% 0	0,0017	0,0017	2,1081	0,1081	-0,0012	-0,0012	2,1064	0,1064	2,0729	0,0729
% 1	-0,0598	-0,0598	1,4107	-0,5893	-0,0112	-0,0112	1,9807	-0,0193	1,3909	-0,6091
% 2	-0,0975	-0,0975	0,9807	-1,0193	-0,0151	-0,0151	1,8173	-0,1827	0,9699	-1,0301
% 3	-0,1317	-0,1317	0,6871	-1,3129	-0,0253	-0,0253	1,6505	-0,3495	0,6801	-1,3199
% 4	-0,1614	-0,1614	0,4853	-1,5147	-0,0396	-0,0396	1,4588	-0,5412	0,4812	-1,5188
% 5	-0,1811	-0,1811	0,3516	-1,6484	-0,0623	-0,0623	1,1084	-0,8916	0,3469	-1,6531



Şekil 4.1. $\beta_0 = 0$ için n=100, veri setinden elde edilen parametre tahminleri



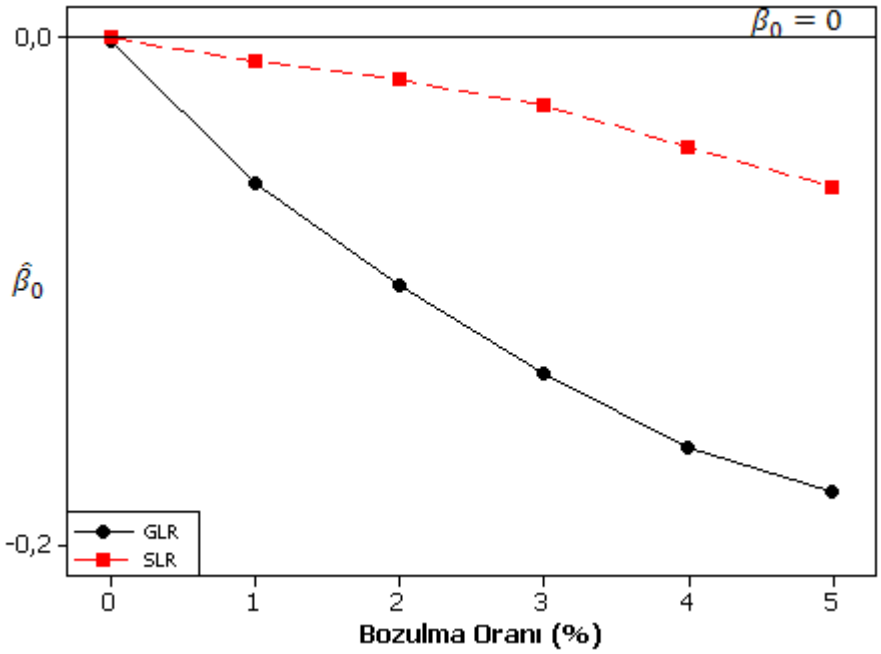
Şekil 4.2. $\beta_1 = 2$ için n=100, veri setinden elde edilen parametre tahminleri

Tablo 4.1'de her bir yöntem için verilen parametre tahminlerinin grafiksel gösterimleri şekil 4.1 ve 4.2'de verilmiştir.

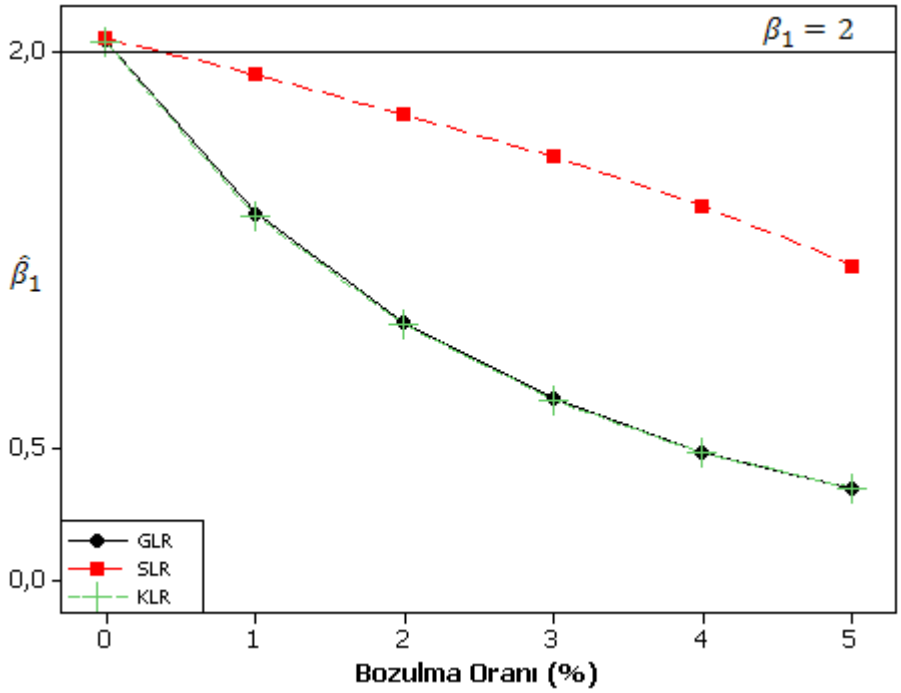
Tablo 4.1'de ve Şekil 4.1 – 4.2'de görüldüğü gibi bozulma oranı S= %0 olduğunda diğer bir anlatım ile homojen veri yapılarında üç yöntemin parametre tahminlerinde ve yanlılıklarında önemli düzeyde bir farklılık gözlenmedi. Ancak bozulma oranı arttıkça GLR ve KLR yöntemlerinin parametre tahminlerinin, SLR yönteminin parametre tahminlerine göre yüksek oranda yanlı olduğu saptandı.

Tablo 4.2. n=200 örnek büyüklüğü, $\beta_0 = 0$, $\beta_1 = 2$ değerleri için 10,000 Monte Carlo Simülasyonundan elde edilen parametre tahminleri ve yanlılıkları.

S	İstatistiksel Yöntemler									
	GLR				SLR				KLR	
	$\hat{\beta}_0$	Bias($\hat{\beta}_0$)	$\hat{\beta}_1$	Bias($\hat{\beta}_1$)	$\hat{\beta}_0$	Bias($\hat{\beta}_0$)	$\hat{\beta}_1$	Bias($\hat{\beta}_1$)	$\hat{\beta}_1$	Bias($\hat{\beta}_1$)
% 0	-0,0016	-0,0016	2,0477	0,0477	-0,0002	-0,0002	2,0519	0,0519	2,0359	0,0359
% 1	-0,0572	-0,0572	1,3900	-0,6100	-0,0096	-0,0096	1,9158	0,0842	1,3817	-0,6183
% 2	-0,0975	-0,0975	0,9737	-1,0263	-0,0165	-0,0165	1,7635	-0,2365	0,9674	-1,0326
% 3	-0,1323	-0,1323	0,6847	-1,3153	-0,0270	-0,0270	1,6029	-0,3971	0,6824	-1,3176
% 4	-0,1618	-0,1618	0,4859	-1,5141	-0,0431	-0,0431	1,4198	-0,5802	0,4832	-1,5168
% 5	-0,1787	-0,1787	0,3509	-1,6491	-0,0590	-0,0590	1,1923	-0,8077	0,3496	-1,6504



Şekil 4.3. $\beta_0 = 0$ için n=200, veri setinden elde edilen parametre tahminleri



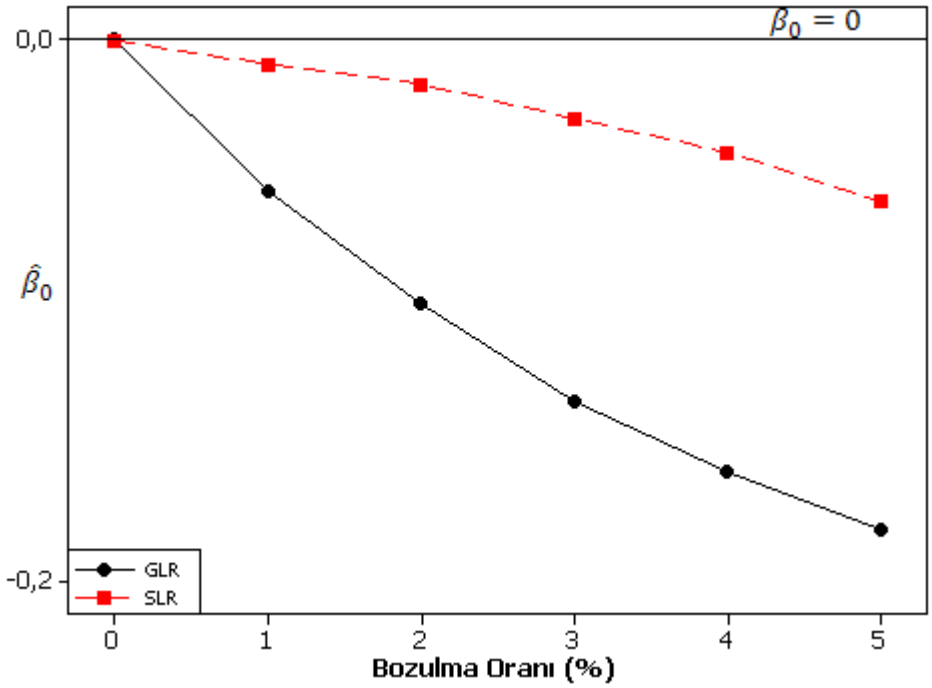
Şekil 4.4. $\beta_1 = 2$ için n=200, veri setinden elde edilen parametre tahminleri

Tablo 4.2'de her bir yöntem için verilen parametre tahminlerinin grafiksel gösterimleri şekil 4.3 ve 4.4'de verilmiştir.

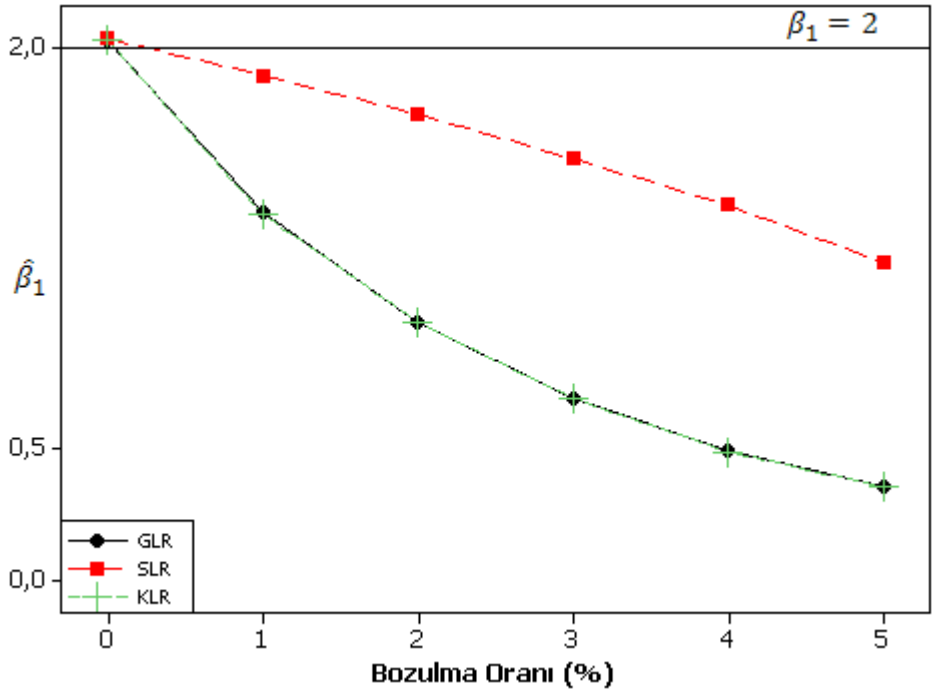
Tablo 4.2'de ve Şekil 4.3 – 4.4'de görüldüğü gibi bozulma oranı S= %0 olduğunda üç yöntemin parametre tahminlerinde ve yanlılıklarında önemli düzeyde bir farklılık gözlenmedi. Ancak S= %0 bozulma oranında elde edilen parametre tahminlerine ait yanlılıkların tablo 4.1'deki S= %0 bozulma oranında elde edilen parametre tahminlerine ait yanlılıklara göre daha küçük olduğu saptandı. Bozulma oranı arttıkça GLR ve KLR yöntemleri, SLR yöntemine göre oldukça yanlı parametre tahminleri vermiştir.

Tablo 4.3. n=300 örnek büyüklüğü, $\beta_0 = 0$, $\beta_1 = 2$ değerleri için 10,000 Monte Carlo Simülasyonundan elde edilen parametre tahminleri ve yanlılıkları.

S	İstatistiksel Yöntemler									
	GLR				SLR				KLR	
	$\hat{\beta}_0$	Bias($\hat{\beta}_0$)	$\hat{\beta}_1$	Bias($\hat{\beta}_1$)	$\hat{\beta}_0$	Bias($\hat{\beta}_0$)	$\hat{\beta}_1$	Bias($\hat{\beta}_1$)	$\hat{\beta}_1$	Bias($\hat{\beta}_1$)
% 0	0,0003	0,0003	2,0347	0,0347	-1,786e-05	-1,786e-05	2,0370	0,0370	2,0276	0,0276
% 1	-0,0560	-0,0560	1,3829	-0,6171	-0,0088	-0,0088	1,8977	-0,1023	1,3765	-0,6235
% 2	-0,0975	-0,0975	0,9733	-1,0267	-0,0162	-0,0162	1,7529	-0,2471	0,9697	-1,0303
% 3	-0,1335	-0,1335	0,6828	-1,3172	-0,0289	-0,0289	1,5851	-0,4149	0,6822	-1,3178
% 4	-0,1598	-0,1598	0,4854	-1,5146	-0,0414	-0,0414	1,4103	-0,5897	0,4839	-1,5161
% 5	-0,1811	-0,1811	0,3524	-1,6476	-0,0596	-0,0596	1,1927	-0,8073	0,3505	-1,6495



Şekil 4.5. $\beta_0 = 0$ için n=300, veri setinden elde edilen parametre tahminleri



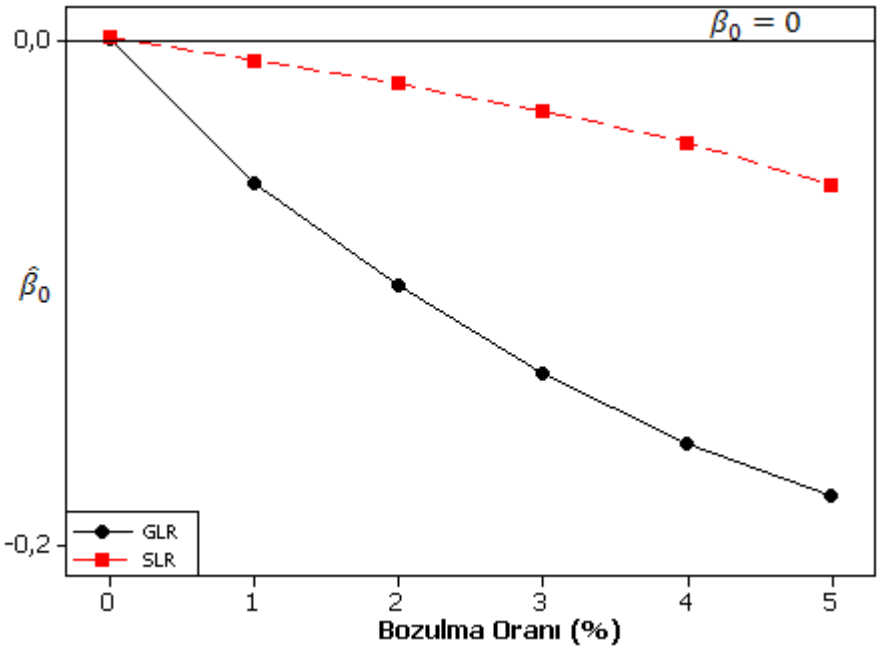
Şekil 4.6. $\beta_1 = 2$ için n=300, veri setinden elde edilen parametre tahminleri

Tablo 4.3'de her bir yöntem için verilen parametre tahminlerinin grafiksel gösterimleri şekil 4.5 ve 4.6'de verilmiştir.

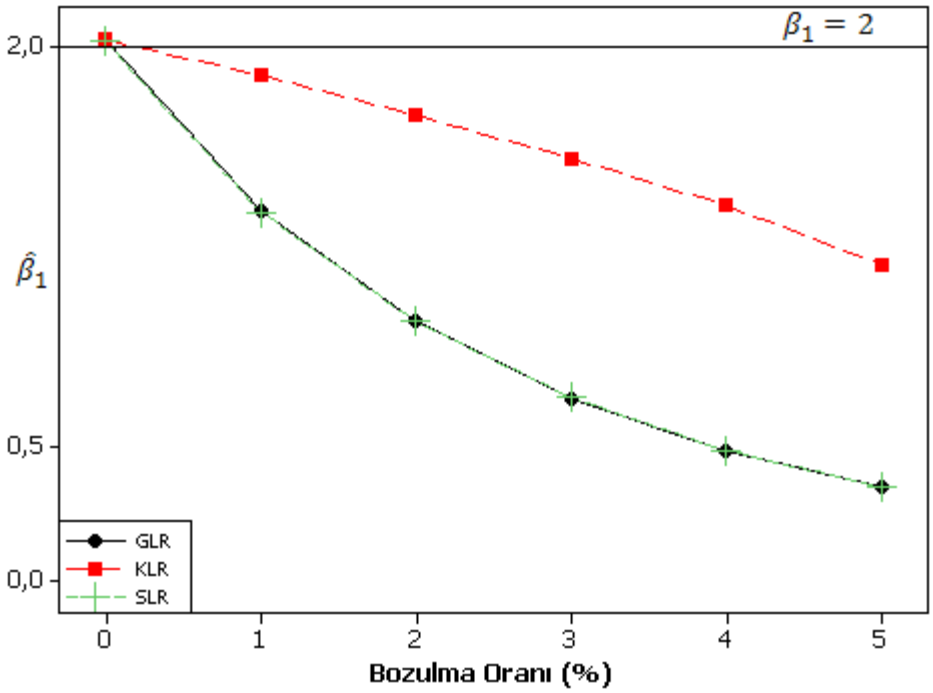
Tablo 4.3'de ve Şekil 4.5 – 4.6'de görüldüğü gibi bozulma oranı S= %0 olduğunda üç yöntemin parametre tahminlerinde ve yanlılıklarında önemli düzeyde bir farklılık gözlenmedi. Ancak S= %0 bozulma oranında elde edilen parametre tahminlerine ait yanlılıkların tablo 4.1 ve tablo 4.2'deki S= %0 bozulma oranında elde edilen parametre tahminlerine ait yanlılıklara göre daha küçük olduğu belirlendi. Fakat bozulma oranı arttıkça GLR ve KLR yöntemlerinin parametre tahminlerinin, SLR yönteminin parametre tahminlerine göre yüksek oranda yanlı olduğu gözlendi.

Tablo 4.4. n=400 örnek büyüklüğü, $\beta_0 = 0$, $\beta_1 = 2$ değerleri için 10,000 Monte Carlo Simülasyonundan elde edilen parametre tahminleri ve yanlılıkları.

S	İstatistiksel Yöntemler									
	GLR				SLR				KLR	
	$\hat{\beta}_0$	Bias($\hat{\beta}_0$)	$\hat{\beta}_1$	Bias($\hat{\beta}_1$)	$\hat{\beta}_0$	Bias($\hat{\beta}_0$)	$\hat{\beta}_1$	Bias($\hat{\beta}_1$)	$\hat{\beta}_1$	Bias($\hat{\beta}_1$)
% 0	0,00084	0,00084	2,0251	0,0251	0,0012	0,0012	2,0257	0,0257	2,0177	0,0177
% 1	-0,0563	-0,0563	1,3811	-0,6189	-0,008	-0,008	1,8899	-0,1101	1,3774	-0,6226
% 2	-0,0973	-0,0973	0,9711	-1,0289	-0,0169	-0,0169	1,7408	-0,2592	0,9688	-1,0312
% 3	-0,1323	-0,1323	0,6829	-1,3171	-0,0281	-0,0281	1,5807	-0,4193	0,6839	-1,3161
% 4	-0,1598	-0,1598	0,4854	-1,5146	-0,0403	-0,0403	1,4033	-0,5967	0,4845	-1,5155
% 5	-0,1801	-0,1801	0,3509	-1,6491	-0,0576	-0,0576	1,1832	-0,8168	0,3511	-1,6489



Şekil 4.7. $\beta_0 = 0$ için n=400, veri setinden elde edilen parametre tahminleri



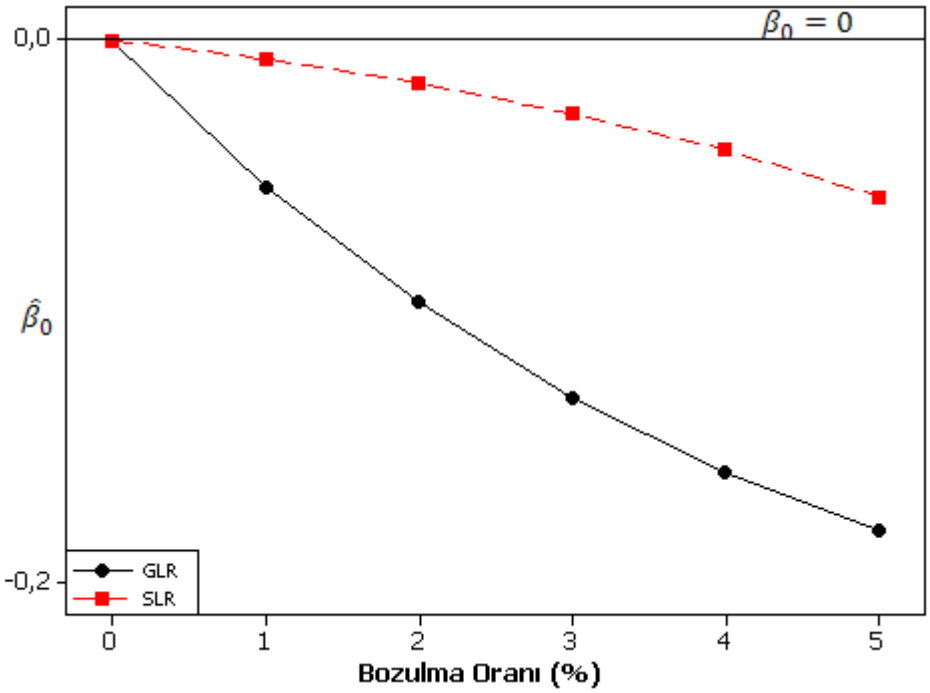
Şekil 4.8. $\beta_1 = 2$ için n=400, veri setinden elde edilen parametre tahminleri

Tablo 4.4'de her bir yöntem için verilen parametre tahminlerinin grafiksel gösterimleri şekil 4.7 ve 4.8'de verilmiştir.

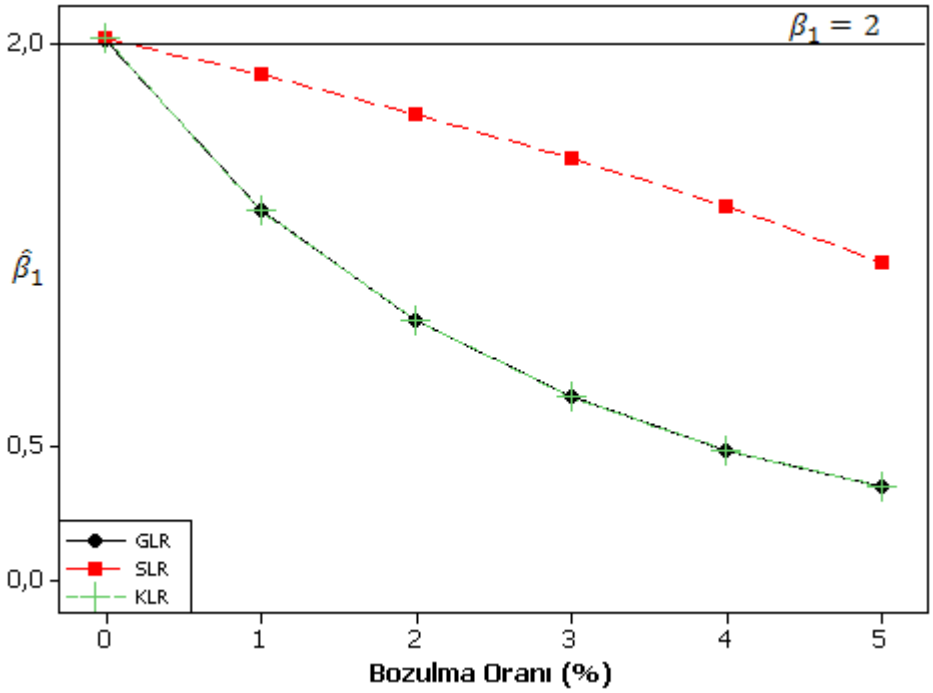
Tablo 4.4'de ve Şekil 4.7 – 4.8'de görüldüğü gibi bozulma oranı S= %0 olduğunda üç yöntemin parametre tahminlerinde ve yanlılıklarında önemli düzeyde bir farklılık gözlenmedi. Ancak S= %0 bozulma oranında tablo 4.4'de elde edilen yanlılıkların tablo 4.1, tablo 4.2, tablo 4.3'deki S= %0 bozulma oranında elde edilen yanlılıklara göre daha küçük olduğu saptandı. Bozulmanın var olduğu veri setlerinde GLR ve KLR yöntemlerinin parametre tahminlerinin, SLR yönteminin parametre tahminlerine göre yüksek oranda yanlı olduğu belirlendi.

Tablo 4.5. n=500 örnek büyüklüğü, $\beta_0 = 0$, $\beta_1 = 2$ değerleri için 10,000 Monte Carlo Simülasyonundan elde edilen parametre tahminleri ve yanlılıkları.

S	İstatistiksel Yöntemler									
	GLR				SLR				KLR	
	$\hat{\beta}_0$	Bias($\hat{\beta}_0$)	$\hat{\beta}_1$	Bias($\hat{\beta}_1$)	$\hat{\beta}_0$	Bias($\hat{\beta}_0$)	$\hat{\beta}_1$	Bias($\hat{\beta}_1$)	$\hat{\beta}_1$	Bias($\hat{\beta}_1$)
% 0	-0,0003	-0,0003	2,0164	0,0164	-0,00023	-0,00023	2,0187	0,0187	2,0199	0,0199
% 1	-0,0544	-0,0544	1,3803	-0,6197	-0,0072	-0,0072	1,8852	-0,1148	1,3771	-0,6229
% 2	-0,0968	-0,0968	0,9701	-1,0299	-0,0161	-0,0161	1,7368	-0,2632	0,9689	-1,0311
% 3	-0,1318	-0,1318	0,6827	-1,3173	-0,0275	-0,0275	1,5734	-0,4266	0,6821	-1,3179
% 4	-0,1594	-0,1594	0,4847	-1,5153	-0,0406	-0,0406	1,3978	-0,6022	0,4843	-1,5157
% 5	-0,1806	-0,1806	0,3518	-1,6482	-0,0579	-0,0579	1,186	-0,8140	0,3506	-1,6494



Şekil 4.9. $\beta_0 = 0$ için n=500, veri setinden elde edilen parametre tahminleri



Şekil 4.10. $\beta_1 = 2$ için n=500, veri setinden elde edilen parametre tahminleri

Tablo 4.5'de her bir yöntem için verilen parametre tahminlerinin grafiksel gösterimleri şekil 4.9 ve 4.10'de verilmiştir.

Tablo 4.5'de ve Şekil 4.9 – 4.10'de görüldüğü gibi bozulma oranı S= %0 olduğunda üç yöntemin parametre tahminlerinde ve yanlılıklarında önemli düzeyde bir farklılık gözlenmedi. Ancak S= %0 bozulma oranında tablo 4.5'de elde edilen yanlılıkların tablo 4.1, tablo 4.2, tablo 4.3, tablo 4.4'deki S= %0 bozulma oranında elde edilen yanlılıklara göre daha küçük olduğu saptandı. Diğer ilk 4 tablodaki elde edilen simülasyon sonuçlarına benzer olarak tablo 4.5'de bozulmanın var olduğu veri setlerinde GLR ve KLR yöntemlerinin parametre tahminlerinin, SLR yönteminin parametre tahminlerine göre yüksek oranda yanlı olduğu saptandı.

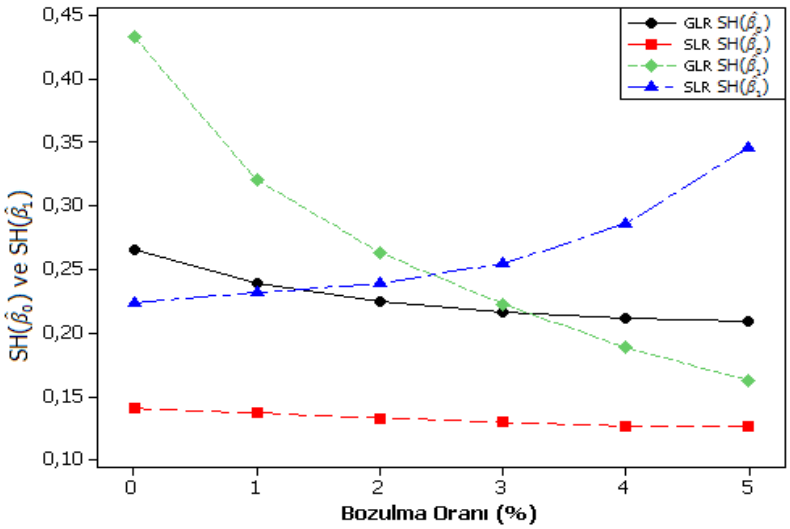
Genel olarak tablo 4.1 – 4.5 ve şekil 4.1 – 4.10'da verilen simülasyon sonuçları, heterojen ikili veri yapılarında diğer bir anlatım ile bozulmanın var olduğu veri yapılarında SLR yönteminin güvenilir ve tutarlı sonuçlar verdiğini göstermektedir.

4.2. Parametre Tahminlerinin Standart Hataları

Simülasyon sonucunda GLR ve SLR yöntemlerinden elde edilen parametre tahminlerine ait standart hatalar; farklı örnek büyüklüğünde ve bozulma durumlarına göre tablo 4.6 – 4.10'de verilmiştir. Tablo 4.6'da sunulan bulgular n=100 örnek büyüklüğü için standart hataları, tablo 4.7'de sunulan bulgular n=200 örnek büyüklüğü için standart hataları, tablo 4.8'de sunulan bulgular n=300 örnek büyüklüğü için standart hataları, tablo 4.9'da sunulan bulgular n=400 örnek büyüklüğü için standart hataları ve tablo 4.10'da sunulan bulgular n=500 örnek büyüklüğü için bulunan standart hataları verilmiştir.

Tablo 4.6. n=100 örnek büyüklüğü için 10,000 Monte Carlo Simülasyonundan elde edilen standart hatalar

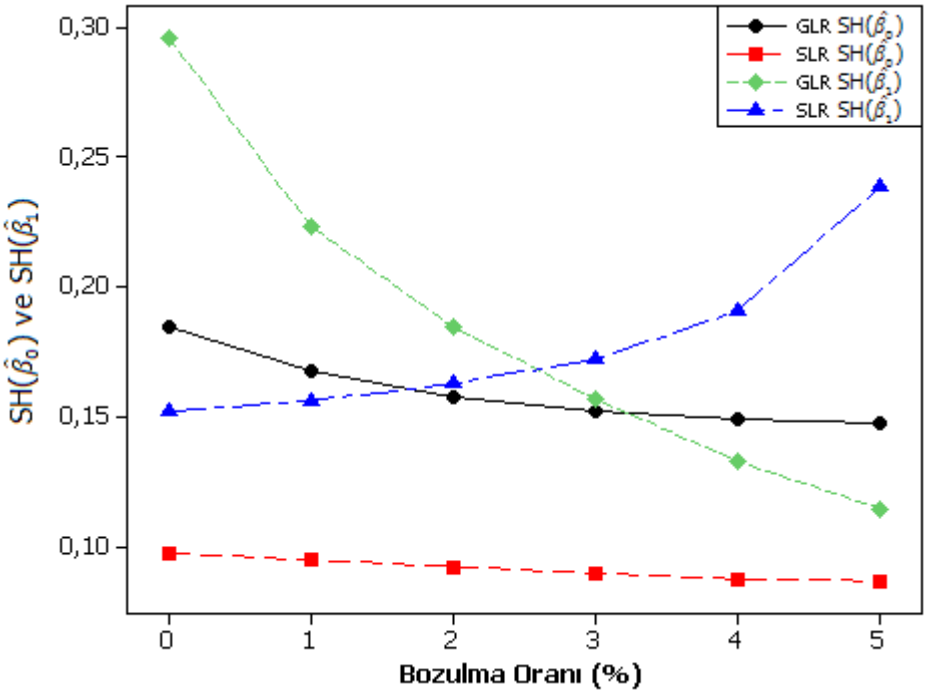
S	İstatistiksel Yöntemler			
	GLR		SLR	
	$SH(\hat{\beta}_0)$	$SH(\hat{\beta}_1)$	$SH(\hat{\beta}_0)$	$SH(\hat{\beta}_1)$
% 0	0,2654	0,4332	0,1401	0,2239
% 1	0,2388	0,3204	0,1367	0,2315
% 2	0,2243	0,2631	0,1328	0,2390
% 3	0,2161	0,2225	0,1293	0,2544
% 4	0,2116	0,1883	0,1266	0,2859
% 5	0,2096	0,1624	0,1263	0,3459



Şekil 4.11. Bozulma durumlarına göre n=100 için veri setinden elde edilen standart hatalar

Tablo 4.7. n=200 örnek büyüklüğü için 10,000 Monte Carlo Simülasyonundan elde edilen standart hatalar

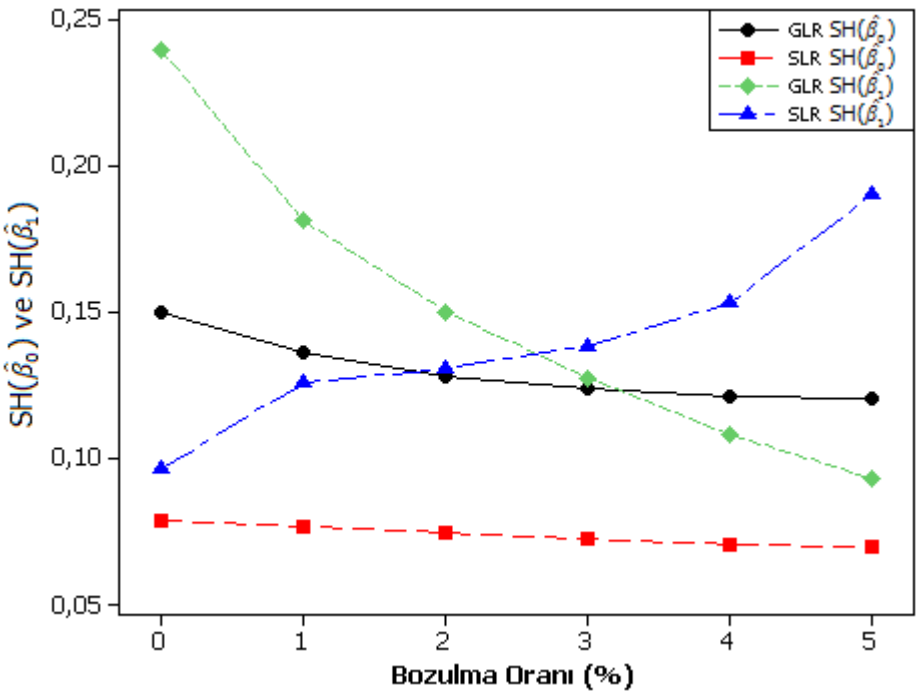
S	İstatistiksel Yöntemler			
	GLR		SLR	
	$SH(\hat{\beta}_0)$	$SH(\hat{\beta}_1)$	$SH(\hat{\beta}_0)$	$SH(\hat{\beta}_1)$
% 0	0,1844	0,2959	0,0971	0,1521
% 1	0,1672	0,2230	0,0946	0,1561
% 2	0,1575	0,1845	0,0919	0,1626
% 3	0,1519	0,1566	0,0895	0,1719
% 4	0,1490	0,1327	0,0872	0,1906
% 5	0,1476	0,1144	0,0864	0,2384



Şekil 4.12. Bozulma durumlarına göre n=200 için veri setinden elde edilen standart hatalar

Tablo 4.8. n=300 örnek büyüklüğü için 10,000 Monte Carlo Simülasyonundan elde edilen standart hatalar

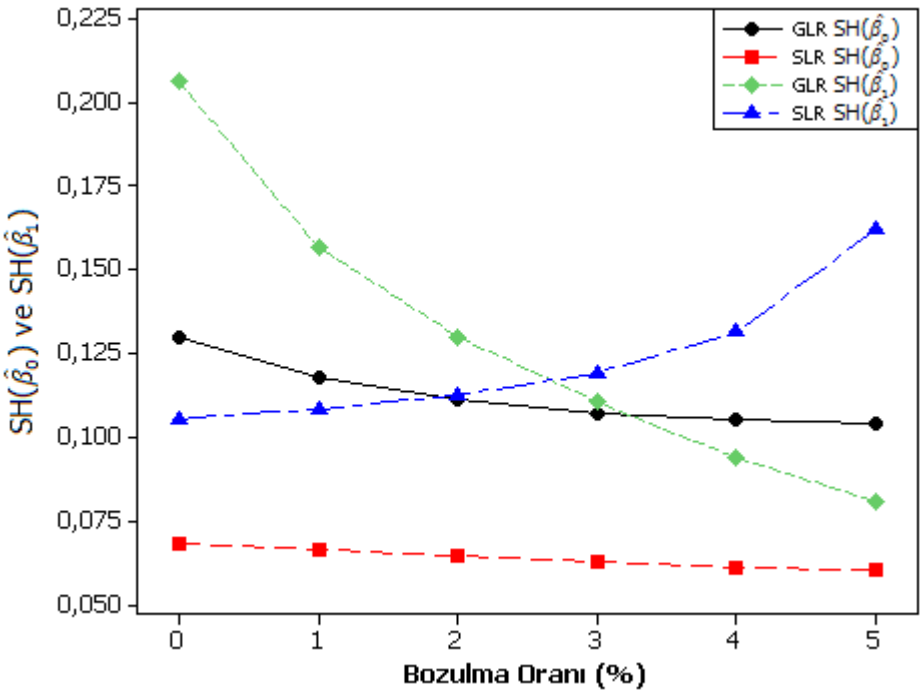
S	İstatistiksel Yöntemler			
	GLR		SLR	
	$SH(\hat{\beta}_0)$	$SH(\hat{\beta}_1)$	$SH(\hat{\beta}_0)$	$SH(\hat{\beta}_1)$
% 0	0,1499	0,2395	0,0788	0,0967
% 1	0,1361	0,1812	0,0768	0,1259
% 2	0,1284	0,1503	0,0747	0,1309
% 3	0,1238	0,1276	0,0725	0,1385
% 4	0,1214	0,1083	0,0706	0,1531
% 5	0,1204	0,0933	0,0698	0,1901



Şekil 4.13. Bozulma durumlarına göre n=300 için veri setinden elde edilen standart hatalar

Tablo 4.9. n=400 örnek büyüklüğü için 10,000 Monte Carlo Simülasyonundan elde edilen standart hatalar

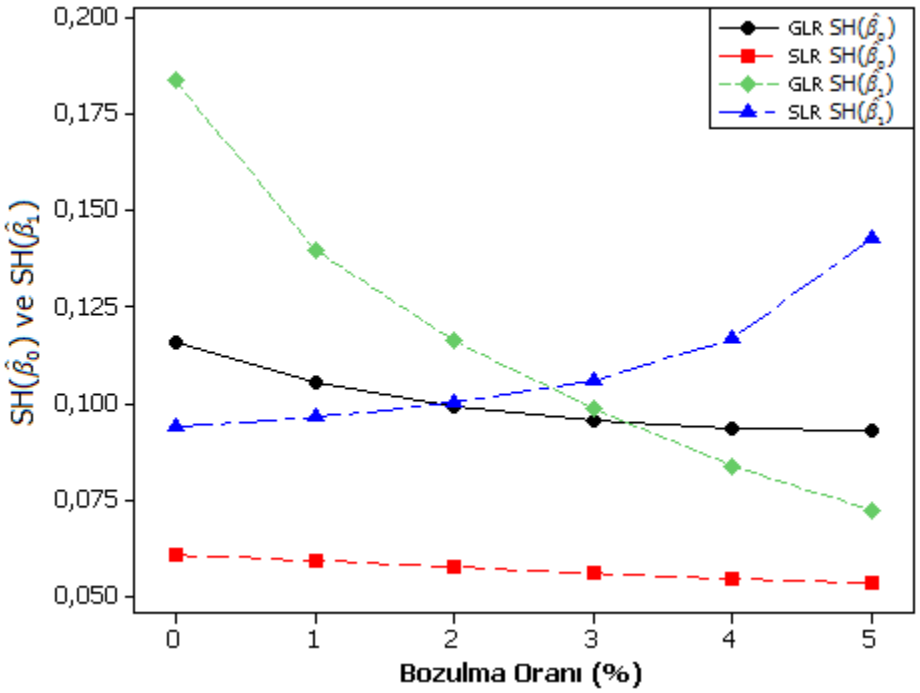
S	İstatistiksel Yöntemler			
	GLR		SLR	
	$SH(\hat{\beta}_0)$	$SH(\hat{\beta}_1)$	$SH(\hat{\beta}_0)$	$SH(\hat{\beta}_1)$
% 0	0,1295	0,2062	0,0681	0,1053
% 1	0,1177	0,1566	0,0663	0,1083
% 2	0,1110	0,1299	0,0644	0,1124
% 3	0,1071	0,1105	0,0626	0,1192
% 4	0,1051	0,0937	0,0609	0,1313
% 5	0,1041	0,0806	0,0600	0,1621



Şekil 4.14. Bozulma durumlarına göre n=400 için veri setinden elde edilen standart hatalar

Tablo 4.10. n=500 örnek büyüklüğü için 10,000 Monte Carlo Simülasyonundan elde edilen standart hatalar

S	İstatistiksel Yöntemler			
	GLR		SLR	
	$SH(\hat{\beta}_0)$	$SH(\hat{\beta}_1)$	$SH(\hat{\beta}_0)$	$SH(\hat{\beta}_1)$
% 0	0,1156	0,1837	0,0607	0,0938
% 1	0,1052	0,1399	0,0592	0,0964
% 2	0,0992	0,1161	0,0575	0,1002
% 3	0,0958	0,0987	0,0559	0,1059
% 4	0,0934	0,0837	0,0544	0,1169
% 5	0,0931	0,0722	0,0535	0,1428



Şekil 4.15. Bozulma durumlarına göre n=500 için veri setinden elde edilen standart hatalar

Tablo 4.6 – 4.10 ve şekil 4.11 – 4.15'de görüldüğü gibi $\beta_0 = 0$ parametresine ait tahminlerin standart hataları GLR yönteminde SLR yöntemine göre yüksek bulunmuştur. β_0 için her iki yöntemden elde edilen standart hatalar bozulma durumlarına göre önemli düzeyde değişiklik göstermemektedir. $\beta_1 = 2$ parametresine ait tahminlerin standart hataları bozulma oranlarına göre hem GLR yönteminde hem de

SLR yönteminde farklılıklar göstermektedir. GLR yönteminden elde edilen $\beta_1 = 2$ parametresine ait tahminlerin standart hataları bozulma oranları arttıkça önemli düzeyde küçülmektedir. Fakat bu durum SLR yönteminden elde edilen $\beta_1 = 2$ parametresine ait tahminlerin standart hataları için ters bir sonuç ortaya koymaktadır. Bozulma oranı arttıkça SLR yönteminden elde edilen $\beta_1 = 2$ parametresine ait tahminlerin standart hatalarının artış gösterdiği bulundu.

Örneklem büyülüğü arttıkça hem GLR hem de SLR yöntemlerinden elde parametresine ait tahminlerin standart hataların azaldığı gözlendi.

5. TARTIŞMA

Sağlık alanında yapılan araştırmalarda verilerin analizinde aynı amaç için farklı biyoistatistiksel yöntemler kullanılabilmektedir. Veri yapısına uygun olarak doğru yöntemin seçimi yansız, tutarlı, etkili, yeterli ve minimum varyanslı parametre tahminlerinin elde edilmesini sağlar [30].

Sağlık alanında kategorik yapıda bağımlı değişken içeren veri yapılarının analizlerinde kullanılan yöntemler arasında lojistik regresyon yöntemleri her zaman güncelliğini korumaktadır. Özellikle birden çok risk faktörünü ele alarak kategorik yapıdaki bağımlı değişken üzerinde etkili olan faktörlerin belirlenmesinde kullanılan GLR yöntemi hemen hemen tüm istatistiksel paket programlarında yer alan bir yöntem olduğu için araştırmacılar tarafından yaygın olarak tercih edilmektedir. Veri yapısının durumuna göre GLR yöntemine alternatif olarak geliştirilmiş KLR ve çeşitli SLR ve yöntemleri son yıllarda kullanılmaya başlanmıştır [5, 7, 12, 16, 20, 21, 24, 34].

GLR yöntemlerinden yansız, tutarlı, etkili, yeterli ve minimum varyanslı parametre tahminleri elde etmek, büyük örnek genişliğinde homojen, seyrek ve çarpık olmayan veri yapısına bağlıdır [15]. Ancak sağlık alanında yapılan çalışmalar sonucunda her zaman büyük örnek genişliğinde homojen ve dengeli dağılıma sahip veri setlerine rastlanılmamaktadır. Bu durumda SLR ve KLR yöntemleri alternatif yöntemler olarak karşımıza çıkmaktadır.

Yapılan literatür taramaları sonucunda KLR yöntemi olarak 1970 yılında Cox tarafından geliştirilen ve paket programlarında yaygın olarak kullanılan yöntem bu tez çalışmasında kullanıldı. SLR yöntemi olarak Bianco ve Yohai (1996) tarafından bulunan ve Croux ve Haesbroeck (2003) tarafından modifiye edilen SLR yöntemi bu tez çalışması için seçildi. Çünkü bu SLR yöntemi diğer SLR yöntemlerine göre hızlı ve stabil sonuç veren bir algoritma kullanmaktadır [7].

Sağlık alanındaki araştırmalardan elde edilen kategorik yapıdaki bağımlı değişken içeren veri setlerinde çoğunlukla bağımlı değişken ikili yapıda gözlenmektedir (ölü-sağ, hasta-sağlam, tedavi var-tedavi yok vb.) [31]. Dolayısıyla bu tez çalışmasında araştırmacılara yol göstermesi ve kullanacakları yöntemin seçiminde kılavuzluk etmesi

amacıyla ikili bağımlı değişken içeren veri setlerinde bu 3 yöntemin performansları karşılaştırıldı. Performansların karşılaştırmasında farklı örneklem büyüklükleri, homojen ve heterojen yapıdaki veri setleri kullanıldı.

Tez çalışmasında incelenen yöntemlerin karşılaştırmalarında Monte Carlo simülasyon yöntemi kullanılarak farklı örnek büyüklüğü ve bozulma oranlarında veri setleri oluşturuldu. Her bir yöntemde bu veri setleri kullanıldı. Yöntemlerin performanslarının değerlendirilmesinde, parametre tahminleri ve yanlılıkları, parametre tahminlerinin standart hataları kullanıldı.

Yapılan simülasyon çalışmaları sonucunda bozulma oranının yöntem seçiminde en etkin faktör olduğu belirlendi. Örneklem büyüklüğünün yöntemlerin performansları üzerine etkilerinin önemli düzeyde olmadığı, örneklem büyüklüğü arttıkça her üç yöntemden de elde edilen parametre tahminlerinin yanlılıkları ve bu tahminlere ait standart hataların aynı oranda azaldığı saptandı. Bozulma durumunun olmadığı (S=%0) diğer bir anlatımla homojen yapıda ikili bağımlı değişken içeren veri setlerinin analizlerinde üç yöntemin de benzer sonuçlar verdiği ancak bozulma oranı %1'den %5'e doğru arttığında yöntemlerin performansları arasında farklılıkların ortaya çıktığı gözlendi.

Yöntemlerin parametre tahminleri ve yanlılıklara göre karşılaştırılmasında, bozulma oranının arttığı durumlarda üç yönteminde yanlı parametre tahminleri vermeye başladığı belirlendi. Ancak SLR yönteminin GLR ve KLR yöntemlerine göre daha yansız parametre tahminleri verdiği gözlendi. Bu sonuçlar bozulma oranının parametre tahminlerine olan etkisinin çok önemli olduğunu göstermektedir. Bozulma oranı %1 bile olsa yaygın olarak kullanılan GLR ve KLR yöntemlerinin parametre tahminlerinin oldukça yanlı olduğu belirlendi. Croux ve Haesbroeck (2003) yaptıkları çalışmada geliştirdikleri ve bu tezde kullandığımız SLR yönteminin performansını %5 bozulma durumunda en çok olabilirlik yöntemi ile parametre tahmini veren GLR yönteminin performansı ile karşılaştırmışlardır. Yaptıkları simülasyon çalışmasında SLR yönteminin GLR yöntemine göre daha az yanlı parametre tahminleri verdiğini göstermişlerdir.

Parametre tahminlerine ait standart hatalar yönünden yöntemlerin karşılaştırılması yapıldığında örneklem büyüklüğü arttıkça hem GLR hem de SLR yöntemlerinden elde edilen parametre tahminlerine ait standart hataların azaldığı gözlendi. KLR yöntemi parametre tahminlerine ait standart hataları hesaplamamaktadır. Dolayısıyla KLR yöntemi diğer yöntemlerle standart hatalar yönünden karşılaştırılamamıştır. Çünkü KLR yöntemi parametre tahminlerinin anlamlılıklarının belirlenmesinde test istatistiği yerine koşullu dağılımdan yararlanarak p değeri hesaplanmaktadır [24].

β_0 parametresi için elde edilen standart hatalar SLR yönteminde GLR yöntemine göre tüm bozulma oranlarında küçük gözlendi. Bozulma oranı arttıkça her iki yöntemden de β_0 parametresi tahmini için elde edilen standart hataların önemli düzeyde değişmediği belirlendi. Hem GLR hem de SLR yönteminde bozulma oranı arttıkça β_0 parametresi tahmini için elde edilen standart hataların düşük oranda azaldığı saptandı.

β_1 parametresi için elde edilen standart hataların bozulma oranlarına göre hem GLR yönteminde hem de SLR yönteminde farklılıklar gösterdiği belirlendi. GLR yönteminden elde edilen standart hataların bozulma oranı arttıkça önemli düzeyde küçüldüğü gözlendi. Fakat SLR yönteminden elde edilen standart hataların bozulma oranı arttıkça önemli düzeyde yükseldiği gözlendi. Bu durum, SLR yönteminin bağımlı değişkende meydana gelen bozulmanın etkisini modele katarak parametre tahminlerine ilişkin standart hatalarda bir düzeltme yaptığını göstermektedir. Aksi taktirde bozulma oranına bağlı olarak standart hataların küçülmesi parametre tahminlerinin anlamlılıklarında kullanılan test istatistiklerini önemli ölçüde etkiler ve parametre tahminlerini önemlilik testlerinde kullanılan p değerlerinin küçük çıkmasını sağlar [3, 7, 15, 24, 32].

Kullanılan ölçütlere göre yöntemlerin performans karşılaştırılmasında genel olarak aşağıdaki sonuçlar gözlenmiştir.

1. Bozulma oranının 0 olduğu homojen veri setlerinde üç yöntemde benzer yansız ve tutarlı parametre tahminleri vermiştir.
2. Bozulma oranının var olduğu heterojen veri setlerinde SLR yöntemi GLR ve KLR yöntemlerine göre daha yansız parametre tahminleri

vermiştir. Ayrıca veri yapısındaki bozulmanın etkisini modele katarak güvenilir tahminler elde etmeyi sağlamaktadır.

3. Seyrek, küçük ve çarpık veri setlerinin analizlerinde KLR yöntemi yaygın olarak kullanılan bir yöntemdir. Ancak aynı şartlarda SLR yöntemi de kullanılabilir. Bu şartlarda SLR yönteminin kullanılması aynı zamanda parametre tahminlerine ilişkin standart hataları düzelterek daha güvenilir sonuçlar vermeyi sağlamaktadır.

6. SONUÇ VE ÖNERİLER

İkili yapıda bağımlı değişken içeren veri setlerinin analizinde kullanılan lojistik regresyon yöntemlerinin performansları veri yapısına göre önemli düzeyde etkilenmektedir. Bozulma oranının %0 olduğu homojen, örnek genişliğinin büyük ve seyrek olmayan (dengeli) veri setlerinin analizlerinde GLR, SLR ve KLR yöntemleri yansız, etkili, yeterli, tutarlı, küçük varyanslı parametre tahminleri vermektedir. Bu açıdan üç yöntem arasında önemli düzeyde bir farklılık yoktur. Fakat GLR yöntemi hemen hemen tüm istatistiksel paket programlarında yer almakta ve yaygın olarak kullanılmaktadır. Bu nedenle homojen, büyük örnek genişliğine sahip, seyrek olmayan veri setlerinin analizlerinde GLR yönteminin kullanılması önerilmektedir.

Küçük örnek büyüklüğüne sahip ya da seyrek yapıdaki veri setlerinin analizlerinde yaygın olarak kullanılan lojistik regresyon yöntemi KLR yöntemidir. Ancak bu şartlar altında da SLR yönteminin kullanılması, β_0 parametresinin tahmin edebilmesi ve tüm parametrelere ait standart hataları hesaplayabilmesinden dolayı bir avantaj sağlamaktadır.

Bozulma durumunun var olduğu heterojen veri setlerinde SLR yöntemi GLR ve KLR yöntemlerine göre güvenilir ve daha az yanlı parametre tahminleri vermektedir. Dolayısıyla veri yapısında meydana gelebilecek bozulmalar söz konusu ise SLR yönteminin kullanılması önerilmektedir. Veri setlerindeki bozulmaları önceden belirlemek, kullanılacak lojistik regresyon analizinin seçiminde önemli rol oynamaktadır. Bu nedenle, analize başlamadan önce bağımlı değişkenin bağımsız değişkenlerle olan ilişkisini gösteren ilişki (Scatterplot) grafiklerinin çizilmesi, bağımlı değişkende meydana gelebilecek bozulmaların saptanması ve bu bozulma oranına göre uygulanacak lojistik regresyon analizinin seçilmesi önerilmektedir.

Bu tez çalışmasında veri setinde meydana gelen bozulma durumlarında üç yöntemin karşılaştırılması yapılmıştır. Gelecekte araştırmanın devamı açısından yöntemlerin karşılaştırılması küçük örnek büyüklüğündeki veri setlerinde, seyrek yapıya sahip bağımlı değişken içeren veri setlerinde ve tam ayrımsamanın var olduğu veri setlerinde bu üç yöntemin karşılaştırılması yapılabilir.

KAYNAKLAR DİZİNİ

1. Adewale, A.J. and Wiens, D., P., 2009, Robust designs for misspecified logistic models, Journal of Statistical Planning and Inference, Canada, 139, 3-15 s.
2. Adimari, G. and Ventura, L., 2001, Robust inference for generalized linear models with application to logistic regression, Elsevier, Italy, 0167-7152/01.
3. Agresti, A., 2002, Categorical data analysis, John Wiley & Sons, Second Edition, Canada, 0-471-36093-7.
4. Albert, A., and Anderson, J.A., 1984, On The existence of maximum likelihood estimates in logistic models, Biometrika, 71, 1–10 s.
5. Bianco, A.M. and Martinez., E., 2009, Robust testing in the logistic regression model, Computational Statistics and Data Analysis, Elsevier, 53, 4095-4105 s.
6. Copas, J.B., 1988, Binary regression models for contaminated data, Journal of the Royal Statistical Society, United Kingdom, 50, 225-265 s.
7. Croux, C. and Haesbroeck, G., 2003, Implementing the bianco and yohai estimator for logistic regression, Elsevier, Belgium, 0167-9473/03.
8. Çolak, E., 2002, Koşullu ve sınırlandırılmış lojistik regresyon yöntemlerinin karşılaştırılması ve bir uygulama, Eskişehir Osmangazi Üniversitesi, Yüksek Lisans Tez Çalışması, Eskişehir.
9. Dietz, K., Gail, M., Krickeberg, K., Samet, J., Tsiatis, A., 2005, Statistics for biology and health, Springer, United States of America, 0-387-20275-7.
10. Elizabeth, N.K., and Thomas, P.R., 2002, A preliminary ınvestigation of maximum likelihood logistic regression versus exact logistic regression, American Statistical Association, 56.
11. Feinstein, A.R., 1993, Multivariate analysis: an introduction, Yale University Press, New Haven, 318, 325-330, 297-312 s.
12. Gervini, D., 2005, Robust adaptive estimators for binary regression models, Journal of Statistical Planning and Inference, Switzerland, 131, doi:10.1016/j.jspi.2004.02.006.

KAYNAKLAR DİZİNİ (devam ediyor)

13. Hampel, F., 2001, Robust statistics: A brief introduction and overview, Seminar for Statistics, Zurich.
14. Hirji, K.F., Mehta, C.R., Patel, N.R., 1987, Computing distributions for exact logistic regression, JASA, 82, 1110-1117 s.
15. Hosmer D.W. and Lemeshow, S., 2000, Applied logistic regression, Second Edition, John Wiley & Sons Inc., Canada.
16. Hosseinian, S. and Morgenthaler, S., 2010, Robust binary regression, Elsevier, Switzerland, 0378-3758.
17. Hüdaverdi, B., 2004, Lojistik regresyon analizi: Tıp verileri üzerine bir uygulama, Kocaeli, Kocaeli Üniversitesi Sosyal Bilimler Enstitüsü Dergisi, 185-208 s.
18. King, G. and Zeng, L., 2001, Logistic regression in rare events data, Society for Political Methodology Cambridge.
19. Kleinbaum, D.G., and Klein, M., 2010, Logistic regression: A self-learning text, 3th Edition, Springer, London.
20. Komarek, R.P. and Moore, A.W., 2003, Fast robust logistic regression for large sparse datasets with binary outputs, Artificial Intelligence and Statistics, Pittsburgh.
21. Kordzakhia, N., Mishra, G.D., Reiersolmoen, L., 2001, Robust estimation in the logistic regression model, Journal of Statistical Planning and Inference, Australia, 0378-3758(00)00312-8.
22. Langholz, B. and Goldstein, L., 2001, Conditional logistic analysis of case control studies with complex sampling, Biostatistics, 1-22.
23. Manning, W.D., Longmore, M.A., Giordano, P.C., 2000, The relationship context of contraceptice use at first ıntercourse, Family Planning Perspectives. 32,104-110 s.
24. Mehta, C.R. and Patel, N.R., 1995, Exact logistic regression: Theory and examples, Harvard School of Public Health, Department of Biostatistics, Cambridge USA, (14): 0277-6715.

KAYNAKLAR DİZİNİ (devam ediyor)

25. Mehta, C.R., Patel, N.R., Senchaudhuri, P., 2000, Efficient monte carlo methods for conditional logistic regression, Journal of the American Statistical Association, Cambridge, 95, 449.
26. Menard., S., 1995, Applied logistic regression analysis, Quantitative Applications in the Social Sciences, Institute of Behavioral Science, University of Colorado, 07-106 s.
27. Myres, J., Huang, S.F., Tsay, J., 2007, Exact conditional ınference for two-way randomized bernoulli experiments, Journal of Statistical Software, 21.
28. Ocakoğlu, G., 2006, Lojistik regresyon analizi ve yapay sinir ağları tekniklerinin sınıflama özelliklerinin karşılaştırılması ve bir uygulama, Yüksek Lisans Tezi, Uludağ Üniversitesi, Bursa.
29. Özdamar, K., 2009, Paket programlar ile istatistiksel veri analizi 1, 7. Baskı, Kaan Kitabevi, Eskişehir.
30. Özdamar, K., 2010, Paket programlar ile istatistiksel veri analizi 2, 7. Baskı, Kaan Kitapevi, Eskişehir.
31. Özdamar, K., 2010, PASW ile biyoistatistik, 8. Baskı, Kaan Kitapevi, Eskişehir.
32. Özdamar, K., Çolak, E., Bal C., Elmalı F., Öztürk, A., Musmul, A., 2010, Seyrek yapıya sahip ikili cevap değişken içeren veri setlerinde asimptotoik, exact ve robust lojistik regresyon yöntemlerinin karşılaştırılması üzerine bir çalışma, XII. Ulusal Biyoistatistik Kongresi, Van, Bildiri Özet Kitapçığı, 47, 28 Haziran-1 Temmuz, 2010.
33. Pampel, F.C., 2000, Logistic regression a primer, Sage University Papers, London 07, 132.
34. Robert, E.D., 2005, Performing exact logistic regression with sas system, NC, SAS Institute Inc., 254-25 s.
35. Santner, T.J. and Duffy, E.D., 1986, A note on A. Albert and J.A. Anderson's conditions for the existence of maximum likelihood estimates in logistic regression models, Biometrica, 73, 755–758 s.

KAYNAKLAR DİZİNİ (devam ediyor)

36. Şahin, M., 1999, Lojistik regresyon ve biyolojik alanlarda kullanımı, Yüksek Lisans Tezi, Kahramanmaraş Sütçü İmam Üniversitesi, Kahramanmaraş.
37. Ürük, E., 2007, İstatistiksel uygulamalarda lojistik regresyon analizi, İstanbul, Cilt Yüksek Lisans Tezi, Marmara Üniversitesi.
38. Victoria, F. and Maria, P., 2000, Robust logistic regression for binomial responses, Geneva : University of Geneva 4, CH-1211.
39. Wilson, P.D. and Langenberg, P., 1999, Usual and shortest confidence ıntervals on odds ratios from logistic regression, The American Statistical, 53, 332-335 s.
40. Yıldırım, A., 1999, Lojistik regresyon analizi ve tıp alanında kullanımına ilişkin bir uygulama, Ankara Üniversitesi Tıp Fakültesi Mecmuası, Ankara 52, 119 - 199 s.
41. Zamar, D., 2006, Monte carlo markov chain exact ınference for binomial regression models, Simon Fraser University, British Columbia.
42. Zamar, D., McNeney, B., Graham, J., 2007, elrm: Software implementing exact-like inference for logistic regression models, Journal of Statistical Software, (2l).

ÖZGEÇMİŞ

Bireysel Bilgiler

Adı-Soyadı	: Muzaffer BİLGİN
Doğum tarihi ve Yeri	: 1986 Aydın
Uyruğu	: Türkiye Cumhuriyeti

Eğitim Durumu

Lisans	Eskişehir Osmangazi Üniversitesi Fen Edebiyat Fakültesi İstatistik Bölümü	2005 - 2009
Yüksek Lisans	Eskişehir Osmangazi Üniversitesi Sağlık Bilimleri Enstitüsü Biyoistatistik ve Tıbbi Bilişim Anabilim Dalı	2009 -

Yabancı Dil **:** İngilizce