T.C.
MARMARA ÜNİVERSİTESİ
SOSYAL BİLİMLER ENSTİTÜSÜ
EKONOMETRİ ANABİLİM DALI
İSTATİSTİK BİLİM DALI

DİSKRİMİNANT VE LOJİSTİK REGRESYON ANALİZLERİNDE BOOTSTRAP TEKNİĞİNİN KULLANIMI VE KREDİ RİSKİ MODELİ OLUŞTURULMASI.

(Doktora Tezi)

Murat ÇİNKO

Danışman: Prof. Dr. ŞAHAMET BÜLBÜL

İstanbul, 2003

İÇİNDEKİLER

Tablolar Listesi

ÖNSÖZ

Finansal sistemin önemli bir parçası olan kredi kartları ile ilgili olan bu çalışmada, kredi kartlarının verilmeden önce geri ödenmeme riskine karşılık bir model geliştirilmeye çalışılmıştır.

Bu çalışmayı sonuçlandırmamda katkısı olan hocam Prof. Dr. Şehamet Bülbül'e ve Yazdığım her satıra yorum yapan hocam Prof. Dr. Şule Özmen'e çok teşekkür ederim.

Tez çalışmam sırasında anlayışlı davranışlarını benden esirgemeyen anabilim dalında görevli öğretim görevlisi arkadaşlarım Dr. Serra Yurtkoru' ya, Dr. Beril Sipahi' ye ve bölümümüzde görevli diğer araştırma görevlisi arkadaşlarıma teşekkür ederim.

Ekonometri bölümü öğretim görevlilerine bu çalışma esnasında gösterdikleri ilgiden dolayı teşekkür ederim.

Giriş

Karar verme bireylerin, kurumların ve toplumların yüzyıllardır yaptığı bir işlem olmasına rağmen bunun için ihtiyaç duyulan verilerin toplanması işlenmesi ve elde edilen bulgular ışığında karar alınması uygulamada istatistik biliminin gelişmesini ve önemini artırmaktadır. Son yıllarda bilgisayar ve iletişim teknolojilerindeki gelişmeler sayesinde daha fazla verinin daha hızlı toplanması, işlenmesi ve karar vericilere sunulması eskisi kadar zor değildir. Özellikle çok sayıda müşterisi olan işletmelerin müşterileriyle ilgili verileri toplayıp bunları değerlendirerek onlara uygun ürün ve hizmet üretme ve sunma kararlarında bilgi ve iletişim teknolojileri ile birlikte kullanılan istatistiksel yöntemler kararların isabet derecesini anlamlı biçimde artırmaktadır. Örneğin mağazaların müşterilerine verdiği mağaza ve benzeri sadakat artırma yöntemi olarak kullandıkları kartlar sayesinde kişilerin ne tür ürünler aldığını takip etmeleri pek de zor değildir.

İstatistik tekniklere günümüzde hemen hemen bütün bilim dallarında ihtiyaç duyulmaktadır. Tıp, ekonomi, genetik, fizik, astronomi bunlardan bir kaçıdır. İstatistik teknikler işletmelerin satın alma, üretim, pazarlama, satış ve benzeri faaliyetlerini yerine getirebilmeleri ve isabetli kararlar almaları için de uygulamada vazgeçilmez tekniklerdir. İstatistik tekniklerin çok önem taşıdığı sektörlerden birisi de bankacılık ve finans sektörüdür. Bu sektörde kullanılmakta olan istatistik teknikler sayesinde karar vericiler kişisel değerlendirmeler yerine bilimsel yöntemlerle karar verme imkanına kavuşmaktadır.

Bankacılık sektöründe kredi verme kararının 1960'lara kadar kişisel kanaatlere dayandığı bilinmektedir. 1960'larda kredi kartlarının kullanılmaya başlanmasından sonra yöneticilerin kart talep eden müşterileriyle ilgili karar verme süreçleri karmaşıklaşmış, ihtiyaç duyulan değişken sayısı ve veri miktarı artmıştır. Dolayısıyla binlerce başvuru için sübjektif olmayan bazı yöntemler bulunması zorunluluk halini

almıştır. Bu yöntemlerin arasında istatistik yöntemler başta gelmektedir. Sadece müşterilerle ilgili kredi risklerinin hesaplanmasında değil birçok karar sürecinde doğrusal programlama, karar ağaçları, diskriminant analizi, lojistik regresyon, yapay sinir ağları gibi istatistik teknikler günümüzde yaygın olarak kullanılmaktadır.

İstatistiksel teknikler kullanarak her hangi bir problemin çözümü için ana kütleden alınacak olan bir örnek sayesinde ana kütle parametreleri için tahminde bulunmak, model oluşturmak, istatistik de önemli bir alandır. Tahmin edilecek parametrelerin seçilen her örnek için farklı değerler alacak olması doğaldır. Bu aşamada istatistik alanındaki büyük sayılar kanunu kullanılarak tahmin edilen değerlerle ilgili olarak bazı hesaplamalar yapmak mümkün olmaktadır. 1970'lerde Efron tarafından önerilen bir yeniden örnekleme tekniği olan Bootstrap tekniği istatistikte kullanılmaya başlanmıştır. Bu teknik, istatistikte tek örnekten hareketle karar vermek yerine örnekden elde edilen değerleri ana kütle değerleri olarak kabul eder ve iadeli olarak buradan seçilecek olan yeni örnekler sayesinde parametreyi tahmin eder.

Tez çalışmasında bootstrap tekniği diskriminant ve lojistik regresyon analizlerinde kullanılmıştır. Bilindiği gibi Diskriminant analizi ve Lojistik regresyon yöntemleri bağımlı değişkenin kategorik olduğu durumlarda kullanılan istatistik teknikleridir. Diskriminant analizinde bağımsız değişkenlerin sürekli olması ve kovaryans matrislerinin eşit olması gerektiğine ait bazı varsayımların uygulama alanında gerçekleşmesi oldukça zordur. Fakat kredi riskini belirlemeye çalışan araştırmacılar bu iki varsayımın bozulduğu durumlardaki modellerin tahmin amacıyla kullanılmasında hatalı sonuçlar vermediğini bulmuşlardır. Bu nedenle Diskriminant analizi uygulamalarında kukla değişkenlerin kullanılması mümkün olmuştur. Diskriminant analizi varsayımlarının yapılmadığı Lojistik regresyon tekniğinin kredi riski alanında kullanılması ise 1980'lerde başlamıştır.

Diskriminant ve lojistik regresyon analizleriyle bootsrap tekniği kullanılarak yapılan uygulamada kullanılan veri setinden birisi kredi kartı başvurusuyla ilgili veri

setidir. Kredi kartlarının kullanımının bu kadar yoğun olduğu günümüzde bankalar açısından verilen kredi kartlarının sahiplerinin borçlarını ödeyip ödemeyeceklerini tahmin etmesi herkese değil de sadece ödeme ihtimali yüksek insanlara kredi kartı vermesi kaynakların kullanımı açısından önemlidir. Günümüzde bankalar iki çeşit derecelendirme yapmaktadır. Birincisi kredi başvurusu yapan kişiye kredinin verilmesi veya verilmemesi, ikincisi ise kredi müşterisinin davranışlarına bakarak kredinin uzatılıp uzatılmayacağı kararlarıdır.

Bu tezin uygulama bölümü kredi başvurusu yapan müşterilere kredinin verilip verilmemesi konusunda karar verici için bir model oluşturmaktır. Bu model oluşturulurken Diskriminant analizi ve Lojistik regresyon teknikleri kullanılacaktır. Yeniden örnekleme tekniği kullanılarak eldeki örnek sayesinde alt örneklemler oluşturup her bir alt örneklem için parametre tahmininde bulunulacak ve bu tahmin edicilerin ortalamaları ana kütle parametresinin tahmin edicisi olacaktır. Yeniden örnekleme sayesinde karar verici bir örnekten elde ettiği tahmin edici yerine oluşturulan yeniden örnekler sayesinde örnek bazında elde edilecek tahmin edicinin bazı hatalarından arınmış olacaktır.

Bu tez beş bölümden oluşmaktadır. Birinci bölümde Diskriminant analizi anlatılacaktır. İkinci bölümde Lojistik regresyon tekniği anlatılacaktır. Üçüncü bölümde ise yeniden örnekleme teknikleri anlatılmaktadır. Dördüncü bölüm kredi riskinin hesaplanmasında kullanılan yöntemlerle ilgili literatürde yer alan çalışmaları açıklamaktadır. Beşinci bölümde ise uygulama sonucundaki bulgular yer almaktadır. Bu tezde iki ayrı veri setine yeniden örnekleme tekniği kullanılarak Diskriminant analizi ve Lojistik regresyon teknikleri uygulanmış. İlk veri seti Thomas (2000) tarafından sağlanan veri setidir, ikincisi ise bir finans kurumundan alınmış veri setidir.

I. DİSKRİMİNANT ANALİZİ

Diskriminant analizi bağımlı değişkenin kategorik, bağımsız değişkenlerin ise sürekli olduğu durumlarda kullanılan bir istatistik tekniktir[1]. Kategorik değişkenin iki veya daha fazla olduğu durumlarda kullanılan bu teknik iki gruplu veya çok gruplu diskriminant analizi olarak ikiye ayrılır.

Diskriminant analizinde öncelikle iki veya çok gruplu olmasına göre söz konusu bu gruplar belirlenir. Bu grupların belirlenme işlemi bağımsız değişkenler kullanılarak doğrusal bir model tarafından ayrıştırılması şeklinde yapılır. Bu ayrıştırma gruplar arasındaki varyansın grup içi varyansına oranın maksimum yapılması sonucunda elde edilir.

Diskriminant analizi bir çok araştırma ve tahmin problemlerinde kullanılabilmektedir. Bu analizin kullanıldığı alanlardan birisi bankacılık sektöründe söz konusu olan kredi verme kararıdır. Kredi talep eden bir kişi veya kuruma kredi vermeden önce onun geri ödemesini yapacak grup içerisinde yer aldığını önceden tespit etmek kredi tahsis kararı için önemli bir bilgidir. Bu analizin kullanılabileceği bir diğer örnek şirketlerin iflas edip etmeyeceğinin tahmin edilmesidir. Kredisini geri ödeyip ödemeyeceğinin çok çeşitli değişkenler ele alarak tahmin etmek ve bu bağımsız değişkenlerin bağımlı değişkenlere etkisini ölçmek iş hayatına sağlayacağı yararlar arasında gösterilebilir. Diskriminant analizi yukarıdaki örneklerde de görüldüğü gibi araştırmacılara çok değişkenli analiz olanağı sağlar ve aynı anda bir çok değişkenin kullanılması gruplar arasındaki farklılıkların çok yönlü ortaya çıkarılmasını mümkün kılar[2].

---

[1] Hair J. F. , Anderson R. E., Totham R. L., Grablowisky B. J., **Multivariate data analysis with readings**. Macmillan, 1984. Sy 81.

[2] Klecka, W. R. **Discriminant Analysis.** Sage Publication 1980, sy 7

Diskriminant analizi grupların bazı bağımsız değişkenler için ortalamalarının birbirine eşit olup olmadığının test edildiği bir hipotezdir[3]. Bunu yapmak için bağımsız değişkenler ağırlıklar ile çarpılarak her bir gözlem için bir diskriminant değeri elde edilir. Her bir grup içinde yer alan gözlemlere ait diskriminant değerlerinin ortalaması alınarak grupların ortalaması bulunur. Elde edilen grup ortalamalarına "merkez" (centroid) denir ve grup sayısı kadar merkez hesaplanır. Grup ortalamaları o gruba ait herhangi bir gözlem için beklenen değeri gösterir.

Diskriminant analizinin kullanım amacı araştırmacılara göre farklılaşabileceği gibi elde edilen bulguların da çok yönlü kullanılması mümkündür. Bazı araştırmacılar için hangi karakteristik özelliklerin gruplar arasındaki farklılığı ortaya çıkardıklarını ve ne kadar iyi ayırt edici olduklarını, bulmak önemli iken, diğer araştırmacılar için bir veya daha fazla matematiksel model ortaya koyarak ayrıştırma işlemini sağlamak daha önemlidir. Sonuç olarak Diskriminant analizinin başlıca iki amacı vardır. Birincisi gruplar arasındaki farkın hangi değişkenden kaynaklandığını ortaya koymak, ikincisi ise gözlemlerin hangi gruba ait olduklarını bulmaktır.

1.1 Diskriminant Fonksiyonu

Aşağıda, diskriminant analizinde kullanılan doğrusal model tanımlanmıştır:

$$Z = w_1X_1 + w_2X_2 + ... + w_nX_n \qquad (1)$$

Z diskriminant değerini

w diskriminant ağırlığını

X bağımsız değişkeni temsil etmektedir.

[3] J. F. Hair, R. E. Anderson, R. L. Totham, B. J. Grablowisky, **Multivariate data analysis**. Printice Hall 1998, Sy 245

Bu matematiksel modellere "diskriminant fonksiyonu" denir. Bu fonksiyon sayesinde herhangi bir gözlemin hangi grup tarafından en iyi temsil edildiği bulunmaya çalışılır. Bu modellerde yer alacak olan değişkenlere "ayrıştırıcı değişken" (discriminating variable) denir. Bu değişkenlerin ortalama ve varyanslarının hesaplanabilmesi için en azından aralık veya oran ölçekleme seviyesinde ölçülmüş olması gerekir. Bu fonksiyon Fisher[4] in doğrusal fonksiyonu olarak da tanımlanmaktadır. Bu fonksiyonun katsayıları hesaplanırken amaç gruplar arası varyansın grup içi varyansa oranını maksimum kılmaktır. **X** açıklayıcı değişkenlerinin oluşturduğu matrisin *p x 1* boyutlarında olduğu ve Σ bu matrisin varyans-kovaryans matrisi, γ *p x 1* ise boyutlarında katsayılar matrisidir. Diskriminant fonksiyonu aşağıdaki gibi tanımlanabilir[5].

$$\xi = X'\gamma \qquad (2)$$

diskriminant değerlerinin karelerinin Toplamı ise şöyle hesaplanır

$$\begin{aligned}\xi'\xi &= (X'\gamma)'(X'\gamma)\\ &= \gamma' XX'\gamma\\ &= \gamma' T\gamma\end{aligned} \qquad (3)$$

$T = X'X$ ve *p* değişken için kareler Toplamını göstermektedir. T varyasyonunu sırasıyla grup içi varyans (B) ve gruplar arası (W) varyansın Toplamı olarak ifade edebiliriz **T = W + B.** Bu Toplamı üç numaralı denklemde yerine koyarsak

$$\begin{aligned}\xi'\xi &= \gamma'(B+W)\gamma\\ &= \gamma' B\gamma + \gamma' W\gamma\end{aligned} \qquad (4)$$

---

[4] Fisher, R. A., "The use of multiple measurement in taxonomics", **Annals of Eugenics**, 7, 1936,179-188

[5] Sharma, S. **Applied Multivariate Techniques**. John Wiley & Sons, 1996,. Sy 277

dört nolu denklemde sırası ile gruplar arası, $\gamma' B\gamma$, ve grup içi varyansların toplamının, $\gamma' W\gamma$, olduğu görülmektedir. Diskriminant fonksiyonun amacının bu iki varyans arasındaki oranı maksimum yapmak olduğu hatırlanırsa,

$$\lambda = \frac{\gamma' B\gamma}{\gamma' W\gamma} \qquad (5)$$

olacaktır.

γ değerlenin bulunması için beş nolu denklemin γ'ye göre türevi alınıp sıfıra eşitlenmelidir.

$$\frac{\partial \lambda}{\partial \gamma} = \frac{2(B\gamma)(\gamma' W\gamma) - 2(\gamma' B\gamma)(W\gamma)}{(\gamma' W\gamma)^2} = 0 \qquad (6)$$

$$\frac{2(B\gamma - \lambda W\gamma)}{(\gamma' W\gamma)} = 0$$

$$(B - \lambda W)\gamma = 0$$

$$(W^{-1}B - \lambda I)\gamma = 0 \qquad (7)$$

Böylece problem artık öz değerleri ve öz vektörleri bulma haline dönüşmüş olur.

$$\left| W^{-1} B - \lambda I \right| = 0 \qquad (8)$$

Yedi numaralı denklem iki gruplu diskriminant analizi için daha da basitleştirebilir

$$B = \frac{n_1 n_2}{n_1 + n_2}(\mu_1 - \mu_2)(\mu_1 - \mu_2)'$$

$$= C(\mu_1 - \mu_2)(\mu_1 - \mu_2)' \qquad (9)$$

Dokuz nolu denklemdeki μ_1 ve μ_2 birinci ve ikinci gruplar için *p x 1* ortalama vektörünü, ***C*** katsayısı ise $n_1 n_2/(n_1+n_2)$ değerine eşittir. Yedi nolu denklem artık aşağıdaki gibi yazılabilir.

$$\frac{C}{\lambda}[W^{-1}(\mu_1 - \mu_2)(\mu_1 - \mu_2)'\gamma] = \gamma \quad (10)$$

$(\mu_1 - \mu_2)\gamma$ sabit bir sayı olması nedeni ile on numaralı denklem aşağıdaki gibi olacaktır.

$$\gamma = KW^{-1}(\mu_1 - \mu_2) \quad (11)$$

$K = C(\mu_1 - \mu_2)'\gamma / \lambda$ sabit olması, kovaryans matrislerinin eşit olması varsayımı altında, $\Sigma_1 = \Sigma_2 = \Sigma_W = \Sigma$ olduğundan, on bir numaralı denklem aşağıdaki şekle dönüşür.

$$\gamma = \Sigma^{-1}(\mu_1 - \mu_2) \quad (12)$$

veya

$$\gamma' = (\mu_1 - \mu_2)'\Sigma^{-1} \quad (13)$$

Diskriminant analizinde en kolay ve en çok kullanılan model "doğrusal" modeldir. Bu modelde bağımsız değişkenlerin doğrusal bir kombinasyonunu kullanmaktadır. Kolay olmasının sebebi ise grup korelasyon matrisinin eşit olduğunu kabul etmesiyle diskriminant fonksiyonundaki hesaplamaların basitleşmesidir.

Diskriminant analizinin başarılı bir şekilde uygulanması için bazı konulara dikkat etmek gerekmektedir. Bu konuları şöyle sıralayabiliriz[6] :

- Bağımsız değişkenlerin seçimi,
- Modelin oluşturulması için gerekli olan örnek büyüklüğü,
- Veri setinin bölünerek tahmin edilen fonksiyonun test edilmesi.

[6]Hair ve diğerleri, **Multivariate data analysis**. Sy 256

- Bağımsız değişkenlerin seçimi. Değişkenlerin seçimi iki şekilde olabilir:

    1) daha önceden yapılmış olan çalışmalardan veya araştırılan konunun teorik çerçevesi içerisinde araştırma konusu ile ilgili olan değişkenler seçilir,
    2) araştırma ile ilgili olduğu düşünülen değişkenler seçilir.

- Örneklem büyüklüğü

Diskriminant analizinde[7] örnek büyüklüğü oldukça hassas bir konu olup, birçok çalışmada örnek büyüklüğünün bağımsız değişkenlerin her birine 20'şer tane gözlem düşecek şekilde ayarlanması tavsiye edilir. Örneklem büyüklüğünün yanı sıra dikkat edilmesi gereken ikinci bir konu ise her bir gruptan en az 20 gözlem olması gerektiğidir. Grupların büyüklüğünün 20'nin üzerinde olduğu durumda eğer bir grubun diğerine göre daha fazla gözlemi varsa bu tahmin edilen modelde gözlem sayısı çok olan grup lehine bir etki yapacağı için büyük olan gruptan rasgele örnek çekerek bu sorun giderilebilir. Bu orana ulaşmak mümkün değil ise örneklem büyüklüğü ve bağımsız değişkenler arasındaki oran azaldıkça sonuçların farklılaşmaya başlayacağının unutulmaması gerekir. Bağımsız değişken sayısı ile örneklem büyüklüğü arasındaki oran minimum beş olmalıdır.

- Modelin test edilmesi

Her hangi bir istatistik tekniği örnek üzerinde kullanarak elde edilen parametre tahmin değerlerinin, aynı veri seti kullanılarak doğruluğu test edilirse sonuçların yanlı olmasına sebep olur. Bu nedenden ortaya çıkan yanlılık sorununu ortadan kaldırmak için örnek iki parçaya ayrılarak birinci bölümü ile model tahmin edilirken ikinci veri seti kullanılarak da modelin tahmin gücü test edilebilir. Modeli tahmin etmek için

---

[7] Hair ve diğerleri, **Multivariate data analysis**, sy 258.

kullanılacak örneklem büyüklüğü ile test edilirken kullanılacak olan örneklem büyüklüğünün oranı konusunda görüş birliği yoktur. Bazıları veri setinin yarısı ile fonksiyonun tahmin edilmesini kalan yarısıyla da test etme işlemini önerirken, bazıları da veri setinin büyük kısmının tahmin işlemi için kullanılmasını az olan kısmının da test etmek için kullanılmasını önermektedir. Genelde önerilen ise veri setini rasgele iki bölüme ayırmak ve ilk bölümü ile diskriminant fonksiyonunu tahmin edip ikinci bölümü ile geçerliliğini test etmektir. Veri setini bölmek için kesin bir kural kullanılmamakla beraber bazı araştırmacılar %60-%40 veya %75-%25 oranlarında veri setini bölerek ilk bölümü kullanarak fonksiyonu tahmin etmekte ikinci veri setini kullanarak fonksiyonun doğruluğunu test etmektedirler.

Veri setinde bölme işlemi esnasında genellikle oransal zümrelere göre örneklem tekniği (proportionally stratified sampling) kullanılır. Eğer kategorik değişken ile ilgili olarak eşit sayıda gözlem var ise oluşturulacak olan iki veri setinde de aynı sayıda kategorik değişken olmalıdır. 50 erkek 50 kadın olan bir veri seti 25 erkek ve 25 kadın olan iki veri setine ayrılarak ilk grupla model kurulur ikincisi ile tahmin işlemi yapılır. Fakat bağımlı değişkenin grupları eşit dağılmadıysa 70 erkek 30 kadın gibi 35 erkek ve 15 kadın olarak iki veri seti elde edilir. Veri setinde böyle bir bölme işlemi yapılabilmesi için yeterli sayıda gözlem olmalıdır. Burada yeterli sayısı için kesin bir kural olamamakla beraber mantıklı olan rakam en az 100'dür. Araştırmacı veri setinin tamamını modeli tahmin etmek için kullanıp daha sonra veri setinin tamamıyla modelin uygunluğunu sınayabilir. Fakat bu daha öncede belirttiğimiz gibi sonuçların fazlaca doğru tahmin edilmesi sonucunu doğurur.

1.2 Diskriminat Analizi Varsayımları

Diskriminat analizinin sonuçlarının doğru olabilmesi için tüm istatistiksel analizlerde olduğu gibi bazı varsayımların yapılması gerekmektedir. Bu varsayımlar kısaca aşağıdaki gibi açıklanabilir.

(1) iki veya daha fazla grup olmalıdır.

(2) her gruba ait en az iki gözlem olmalıdır.

(3) bağımsız değişken sayısının Toplam gözlem sayısından en fazla iki eksiği olmalıdır.

(4) bağımsız değişkenler en azından aralık ölçekde ölçülmelidir.

(5) bağımsız değişkenlerden hiçbirisi diğerinin doğrusal kombinasyonu olmamalıdır.

(6) her grup için korelasyon matrisi eşit olmalıdır.

(7) bağımsız değişkenlerin çoklu normal dağılıma sahip olması gerekmektedir

Bağımsız değişken sayısının gözlem sayısının iki eksiğinden az olması durumunda matris hesaplamalarında matematiksel bir problem yaşanmayacaktır. Bağımsız değişkenlerle ilgili olarak bazı istatistiksel kısıtlar söz konusu olabilir. Hiçbir bağımsız değişken başka bir bağımsız değişkenin doğrusal bir fonksiyonu olamaz. Böyle bir durumda değişkenlerden birinin kullanılması yeterli olacaktır. Çünkü ikinci değişken modele ekstra bir bilgi getirmeyecektir. Tam korelasyona sahip iki değişkenin aynı anda kullanılması matematiksel modeldeki hesaplama yapılmasını imkansız hale getirir.

Çoklu normal dağılım varsayımının bağımsız değişkenlerin her birinin normal dağılıma sahip olması anlamına gelmektedir. Bu varsayım geçersiz ise hesaplanan ihtimal değerleri kesin olamaz. Ancak dikkatli analiz edildiği takdirde yararlı olabilir. Herhangi bir problem için bu kabullenme olmaz ise bulunan istatistiksel sonuçlar gerçeğe yakın olmaktan uzaklaşır.

Veri setinin çoklu normal dağılıma uymaması durumunda ise diskriminant fonksiyonun tahmininde hataya sebep olur. Kovaryans matrisinin eşit olmaması ise gruplama işlemini olumsuz etkileyecektir. Kovaryansların eşit olmadığı durumda tahmin işleminde büyük kovaryansa sahip grupta fazla tahmin değeri olacaktır. Bu durum o gruptaki gözlem sayısını artırarak veya tahmin işlemi esnasında grupların kendi kovaryans matrislerini kullanarak çözülebilir. Kuadratik sınıflama yöntemi

kullanılması da bir çözümdür. Veri seti içerisinde yer alan uç değerler (outlier), modeli etkileyeceği için bunların veri setinden çıkarılması sonuçların doğruluğu için yararlı olacaktır.[8]

Diskriminant analizinde en zor sağlanan varsayımlar; çoklu normal dağılım ve eşit kovaryans matrisi varsayımlarıdır. Çoklu normal dağılım varsayımı istatistiksel testlerin yapılabilmesi için gereklidir. Bu testlerde hesaplanan istatistiklerin teorik olarak düşünülen modelle karşılaştırılması söz konudur. Grup kovaryans matrisinin eşit olması durumu ise kanonikal diskriminant fonksiyonunda ve sınıflama fonksiyonunda hatalara sebep olacaktır[9].

1.3 Diskriminant Analizinde Bulunan Katsayılar

Diskriminant fonksiyonunda üç çeşit katsayı hesaplanır. Bunlar standart ağırlık, standart olmayan ağırlık ve yapısal katsayıdır. Standart olan ağırlıklar açıklamalarda anlamlıdır, standart olmayan ağırlıklar ise diskriminant değerlerinin hesaplanmasında kullanılır. Yapısal katsayı ise standart katsayıdan farklı olarak diskriminant fonksiyonu ile her bir değişken arasındaki ilişkiye bakar.

1.3.1 Standart Katsayısı

Standart katsayısı (discriminant coefficient) fonksiyondaki bağımsız değişkenlerin önünde yer alan sayıdır. Bağımsız değişkende meydana gelecek bir birimlik değişme sonrasında gözlemde meydana gelecek olan değişimi göstermektedir. İşareti göz ardı edersek standart katsayı ilgili bağımsız değişkenin fonksiyona ne miktarda katkı yaptığını gösterir. Büyük katsayıya sahip bağımsız değişken küçük katsayıya sahip değişkenden daha fazla katkı yapmaktadır. Katsayının işareti ise değişkenin pozitif veya negatif katkı yaptığını gösterir.

---

[8] Hair ve diğerleri, **Multivariate data analysis** a.g.e. sy 262.
[9] Klecka a.g.e sy 61

1.3.2 Standartlaştırılmış Katsayılar

Bağımsız değişkenlerin ölçüm birimlerindeki farklılık nedeniyle modele ne kadar etki edildiğini anlamak için standartlaştırılmış katsayılara bakılması gerekecektir. Çünkü bu katsıyılar değişkenin modele mutlak katkısını gösterecektir. Bu değişkendeki bir birimlik artışın veya azalışın modelde ne kadar değişim yaratacağını anlamımızı sağlayacaktır.[10] Her bir değişkendeki bir birimlik artış veya azalış aynı derecede önemli olmayacağından standardize edilmiş katsayıların kullanılması gerekmektedir. Aşağıdaki formülü kullanarak her hangi bir katsayıyı standardize etmek mümkün olacaktır.

$$C_i = u_i \sqrt{\frac{w_{ii}}{n_. - g}} \qquad (14)$$

u_i standardize edilmemiş diskriminant katsayısı,

W_{ii} i değişkeni için kareler Toplamını,

$n_.$ Toplam gözlem sayısını,

g ise grup sayısını göstermektedir.

1.3.3 Yapısal Katsayı (Structure Coefficient)

Diskriminant fonksiyonu ve bağımsız değişken arasındaki benzerliğe bakmak için bu katsayının kullanılması gerekecektir. Korelasyonda olduğu gibi bu sayı bağımsız değişken ve diskriminant fonksiyonu arasındaki açının kosinüsüdür. Bu sayının bilinmesi data düzleminde geometrik yapı hakkında bilgi vermesi açısından önemlidir. Yapısal katsayı bir değişken ile fonksiyonun yakınlığı hakkında bilgi verir. Yapısal katsayının mutlak değerinin büyük olması, 1'e yakın olması, fonksiyon ile değişkenin yapısının aynı olduğunu, mutlak değerin küçük olması, 0 yakın olması, değişken ile fonksiyon arasındaki ilişkinin az olduğunu göstermesi açısından önemlidir.

[10] Klecka, a.g.e. sy 29

Standartlaştırılmış katsayı ile yapısal katsayı birbirlerinden farklı yorumlanmaktadır. Standartlaştırılmış katsayı bize değişkenin modele ne kadar katkı yaptığını göstermektedir. İlk ciddi sorun eğer iki bağımsız değişken yüksek bir korelasyona sahip ise ortaya çıkar. Bu durumda her ikisinin etkisinin de modele katkısı önemli gözükebilir ancak birinin kullanılması yeterlidir. Standartlaştırılmış katsayılar birbirlerine yakın ancak ters işaretli de olabilir. Bu durumda birinin katkısı diğerinin katkısını yok edecektir. Fakat yapısal katsayı sadece ikili korelasyona yani, diskriminant fonksiyonu ve bağımsız değişken arasındaki korelasyona, baktığı için diğer değişkenler ilişkiyi etkilememektedir.

1.4 Diskriminant Analizinde Sınıflama İşlemi

Sınıflama işlemi[11] gözlemlerin ait olduğu gerçek grubunun ve modelin bu gözlemler için tahmin ettiği grubun gösterdiği bir tablodur. Sınıflama işlemi sonucu oluşan bu matrise sınıflama matrisi denilecektir. Örneğin iki gruba ayrıştırma söz konusu olduğu bir modelde sınıflama matrisi gerçek gözlem değeri birinci grupta olan gözlemlerin ne kadarının, yüzde kaçının, model tarafından da doğru olarak yani ilk gruba ayrıştırıldığı ne kadarının, yüzde kaçının, ise hatalı olarak ikinci gruba ayrıştırıldığını gösterir. Matriste bu sayı ve oranlar her grup için yer alır. Modelin doğru veya hatalı ayrıştırma oranını gösteren bu matris modelin tahmin gücünün anlaşılması açısından önemli bir göstergedir. Regresyon analizinde oluşturulan modelin tahmin gücü R^2 tarafından ölçülmekte aynı işlem diskriminant analizinde bağımlı değişkenin kategorik olması sebebiyle sınıflamanın doğru yapılmasıyla ölçülmektedir. Sınıflama matrisinde doğru tahmin edilme oranına doğru sınıflama oranı (hit ratio) denir. Herhangi bir gözlemi sınıflamak veya herhangi bir yeni gözlemin grubunu belirlemek için bir çok yöntem olmasına rağmen, genelde kullanılan üç tane yöntem vardır. Bu yöntemler:

[11] Sharma, a.g.e sy 254.

- Kritik değer.
- İstatistiksel karar teorisi.
- Mahalanobis uzaklığı metodudur.

1.4.1 Kritik Değer Metodu

Sınıflama her hangi bir deneğin hangi gruba ait olduğunun belirlenmesidir. Bu işlem için bağımsız değişkenler kullanılır, her bir deneğin grup ortalamasına olan uzaklıkları hesaplanarak deneğin hangi gruba ait olduğuna karar verilmeye çalışılır. Bu sınıflama işlemi esnasında bağımsız değişkenler veya kanonikal diskriminant fonksiyonu kullanılır. Birinci işlem sırasında diskriminant analizi kullanılmamakla beraber, her hangi bir testin yapılması da mümkün değildir. İkinci işlem olan kanonikal diskriminant fonksiyonu oluşturulacak olursa testlerin yapılması da mümkün olur.

Diskriminant değerinin hesabı:

Her bir gözlem için formül (1) kullanılarak bir diskriminant değeri hesaplanır. Hesaplanan değer sürekli bir değişken değeridir. Diskriminant değerlerinin birbirine yakın olması gözlemlerin aynı grup içerisinde olma olasılığını artırır. Sınıflama işleminin yapılabilmesi için bir kritik değerin hesaplanması gerekmektedir. Bu kritik değer gruplarda eşit sayıda gözlem var ise

$$Z_{KD} = \frac{Z_A + Z_B}{2} \qquad (15)$$

Z_{KD} kritik değer

Z_A A grubu için ortalama diskriminant değeri

Z_B B grubu için ortalama diskriminant değeri

olarak hesaplanır.

Gruplarda eşit sayıda gözlem olmadığı durumda ise kritik değer aşağıdaki şekilde hesaplanır

$$Z_{KD} = \frac{N_B Z_A + N_A Z_B}{N_A + N_B} \qquad (16)$$

Z_{KD} kritik değer

N_A A grubu için gözlem sayısı

N_B B grubu için gözlem sayısı

Z_A A grubu için ortalama diskriminant değeri

Z_B B grubu için ortalama diskriminant değeri

Gruplardaki gözlem sayılarının farklı olması durumunda çok gözlemi olan grup için diskriminant değerinin yükseleceği bilindiğinden gruplar için kendilerinin değil diğerlerinin gözlem sayısı ağırlık olarak kullanılmaktadır.

Sınıflama matrisinin oluşturulması:

Diskriminant fonksiyonun geçerliliğini test etmek için veri seti ikiye bölünür ve ilk veri seti ile fonksiyon elde edilir. Yukarıdaki formüllerden uygun olan (15) veya (16) kullanılarak kritik değer hesaplanır. İkinci veri setinde her bir gözlem değeri için diskriminant değeri hesaplanır.

Z_n n inci gözlem için hesaplanan diskriminant değeri

Z_{KD} kritik değer,

olmak üzere.

Eğer $Z_n < Z_{KD}$ ise gözlem A grubundadır

Eğer $Z_n > Z_{KD}$ ise gözlem B grubundadır

1.4.2 İstatistiksel Karar Teorisi

Bu metot[12] kullanılırken amaç hatalı sınıflama maliyetini ve önsel olasılıkları dikkate alarak hatalı sınıflama oranını minimize etmektir. Önsel olasılığı ve hatalı sınıflama maliyetini dikkate alan bu metot Bayes teorisini kullanmaktadır. Olasılıklar hakkında önsel bir bilgi varsa birinci grupta olma olasılığı p_1 ve ikinci grupta olma olasılığı ise p_2 olarak alınacaktır. Her hangi bir gözlemin birinci grupta tahmin edilmesi için

$$Z \geq \frac{\overline{Z}_1 + \overline{Z}_2}{2} + \ln\left[\frac{p_2}{p_1}\right] \qquad (17)$$

ikinci grupta tahmin edilmesi için ise

$$Z < \frac{\overline{Z}_1 + \overline{Z}_2}{2} + \ln\left[\frac{p_2}{p_1}\right] \qquad (18)$$

olması gerekecektir. Formülde yer alan Z değeri her hangi bir gözlem için hesaplanan diskriminant değerini, $\overline{Z}_j$ j grubu için hesaplanan ortalama diskriminant değerini ve p_j ise her grup için önsel olasılık değerini göstermektedir. Hatalı sınıflama maliyetlerini karar verme mekanizmasına dahil ederken aşağıdaki tablo kullanılacaktır.

Tablo 1. Hatalı Sınıflama Maliyeti

	Gerçek Durum	
Tahmin Edilen Durum	Grup 1	Grup 2
Grup 1	Sıfır Maliyet	C(1/2)
Grup 2	C(2/1)	Sıfır Maliyet

[12] Sharma, a.g.e. sy 256

Hatalı sınıflama maliyetlerini karar verme mekanizmasına dahil ettiğimizde 17 ve 18 numaralı denklemler aşağıdaki şekilde değişirler. Her hangi bir gözlemin birinci gruba atanması için

$$Z \geq \frac{\overline{Z}_1 + \overline{Z}_2}{2} + \ln\left[\frac{p_2 C(1/2)}{p_1 C(2/1)}\right] \qquad (19)$$

ikinci gruba atanması için ise

$$Z < \frac{\overline{Z}_1 + \overline{Z}_2}{2} + \ln\left[\frac{p_2 C(1/2)}{p_1 C(2/1)}\right] \qquad (20)$$

şeklinde olacaktır.

Eğer önsel olasılıklar ve hatalı sınıflama maliyetleri eşit kabul edilirse istatistiksel karar teorisindeki kritik değerler ilk metot olan kritik değer metodu ile aynı sonuçları verecektir.

1.4.3 Mahalanobis Uzaklığı Metodu

Gözlemlerin her birinin grup ortalamalarına olan uzaklıklarının hesaplanarak yakın olan gruba dahil edilmesi daha anlamlı olacaktır. Bağımsız değişkenler arasında korelasyon olduğunda standart sapmadaki bilgi uzaklık hesaplanmasıyla gözükmez, bu nedenle uzaklık tanımının iyi yapılması gerekecektir. Bu sorunu ortadan kaldıracak olan uzaklık formülü Mahalanobis[13] tarafından şöyle tanımlanmıştır:

$$D^2(X \mid G_k) = (n_. - g)\sum_{i=1}^{p}\sum_{j=1}^{p} a_{ij}(X_i - X_{ik.})(X_j - X_{jk.}) \qquad (21)$$

[13] Mahalanobis, P. C., "On the generalized distance in statistics", **Proceedings of the National Institute of Social Science**, India, ,12, 1963 49-55

Formüldeki D^2 ($X \mid G_k$) *k* nolu gruba *X* gözleminin uzaklığını göstermektedir. Her bir grup için D^2 ler hesaplandıktan sonra bu değerlere bakarak sınıflama işlemi yapılabilir. Bu formül için grup kovaryans matrislerinin eşitliği varsayımı yapılmıştır. Eğer bu varsayım gerçekleşmeyecek olursa Tatsuoka[14] tarafından geliştirilen formül kullanılmalıdır.

Sınıflama ihtimalleri:

Hesaplanan uzaklıklar, D^2, ki-kare dağılımına sahip olduklarından, uzaklıkların ki-kare cinsinden ölçüldüğü söylenebilir. Eğer gruplar çoklu normal dağılıma sahipse gözlemlerin grupların ortalaması etrafında toplanması gerekecektir. Uzaklıkların ki-kare cinsinden hesaplandığını düşünürsek istatistiksel anlamlılık testlerinin yapılması mümkün olacaktır. $P(X|G_k)$ herhangi bir gözlemin grup ortalamasına olan uzaklığın ihtimalini gösterecek olursa her hangi bir gözlemin hesaplanmış olan D^2 değerine göre en yakın gruba ait olduğu düşünülür. Herhangi bir gözlem hesaplanan değer nedeni ile iki gruba da ait veya hiç birine ait değilmiş gibi gözükebilir. Uzaklıkların ihtimalinin Toplamının bire eşit olması gerekmemektedir. Her gözlemin bir gruba ait olması gerektiğinden gruplardan birine ait olma olasılığını hesaplayabiliriz. Herhangi bir gözlem değeri, X, grup *k*'ya ait olma olasılığı şöyle hesaplanabilir.

$$P(G_k \mid X) = \frac{P(X \mid G_k)}{\sum_{i=1}^{g} P(X \mid G_i)} \qquad (22)$$

Bu ihtimale sonsal ihtimal denir ve bütün gruplar üzerinden Toplamı bire eşittir.

[14] Tatsuoka, M. M., **Multivariate Analysis** . New York: John Wiley. 1971.

1.5 Diskriminant Analizinde Seçme Kriterleri

Diskriminant analizi yapılırken bazı istatistiklerin incelenmesi sonucu modele girecek olan değişkenlerin seçilmesi mümkündür. Bu kriterler sırası ile Wilk's Lamda, Rao's V, Malanalobis uzaklığının karesi ve gruplar arası F oranıdır. Bu kriterler aşağıda sırasıyla açıklanmaktadır.

1.5.1 Wilk's Lamda Katsayısı

Wilk's lamda katsayısı birden fazla değişkenin olduğu durumlarda gruplar arasındaki farklılığın olduğunu ortaya koyan çok değişkenli bir katsayıdır. Seçilen örneğin anakütlenin bir parçası olması dolayısıyla yeni bir sorun gündeme gelmektedir. Aslında ana kütlede hiçbir grup farklılığı olmamasına rağmen örnek sonucunda gruplar arasında bir fark bulunması olasılığı nedir. Bu noktada istatistiksel anlamlılık şeçilen örneğin ne kadar anlamlı olduğunu gösterir. İstatistiksel anlamlılıktan söz edilebilmesi için örnekleme tekniğinin rastlantısal örnekleme tekniği olması gerekmektedir. Fonksiyonun kendisini incelemek yerine modelden elde edilen kalıntıların test edilmesi daha anlamlı olacaktır.

Bu katsayının hesaplanmasında birden fazla yöntem olmasına rağmen aşağıdaki yöntem kullanılacaktır[15].

$$\Lambda = \frac{SS_w}{SS_t} = \frac{SS_w}{SS_b + SS_w} \qquad (23)$$

SS_w grup içi Toplamların karesi, SS_b ise gruplar arası Toplamların karesidir.

Wilk's lamda katsayısının test edilebilmesi için öncelikle, bu katsayının F veya ki-kare dağılımına uygun hale getirilmesi gerekmektedir. Bu dönüşüm yapıldığı anda hesaplanmış tablolardan karşılaştırma yapmak kolaylaşacaktır. Bu dönüşüm ise aşağıdaki formül ile yapılacaktır:

---

[15] Sharma a.g.e. sy 266

$$\chi^2 = -\left[n_{\cdot} - \left(\frac{p+g}{2}\right) - 1\right] \log_e \Lambda_k \qquad (24)$$

değişkenin serbestlik derecesi ise $(p\text{-}k)$ $(g\text{-}k\text{-}1)$ olacaktır.

1.5.2 Rao's V

Rao's V[15] mahalanobis uzaklığını kullanarak her bir grubun ortalamasının bütün gözlemlerin ortalamalarına olan uzaklıklarını dikkate alarak hesaplanan bir istatistiktir. Bu istatistik gruplar arasındaki ayrıştırmayı maksimize ederken grubun homojenliğini dikkate almamaktadır.

1.5.3 Mahalanobis Uzaklığı

İlk iki yöntemde de gruplar arasındaki ayrıştırma maksimum yapılmaya çalışılmaktadır. Fakat iki gruptan fazla grup olması durumunda bütün grupların optimum uzaklıkta olması sağlanamayabilir. Mahalanobis uzaklığı bütün grupların birbirlerinden olan uzaklıklarının optimum olmasını sağlamaktadır.

1.5.4 Gruplar Arası F-Oranı

Mahalanobis uzaklığı kullanılarak hesaplanan değerde bütün grupların ağırlığının eşit olduğu varsayımı yapılır. Bu kısıtlamadan kurtulabilmek için Mahalanobis uzaklığı F oranına dönüştürülür. Bu dönüştürme işlemi sırasında büyük gruplar küçük gruplara oranla daha fazla ağırlık alırlar.

---

[15] Sharma, a.g.e. sy 266

II. LOJİSTİK REGRESYON

Regresyon analizi sosyal bilimlerde sıkça kullanılan istatistik yöntemlerden birisidir. Gauss-Markov varsayımları altında takdir ediciler sapmasız, etkin, tutarlı ve yeterli özelliklerine sahiptir. Bir çok istatistik paket programının içerisinde yer alması bu yöntemin bu kadar sık kullanılmasını teşvik etmiştir. Fakat bu kadar yaygın olmasına rağmen; varsayımların bazılarının geçerli olmadığı gibi durumlarda da araştırmacılar tarafından hatalı kullanıldığı da gözlenmektedir. Örneğin bağımlı değişkenin nicel olmayıp nitel olması durumun da kullanılması gerçekçi olmayan sonuçlar bulunmasına sebep olabilmektedir. Regresyon analizinin kullanılması esnasında bağımsız değişkenlerin nitel veya nicel olması sorun yaratmamaktadır. Ancak bağımlı değişkenin sürekli olduğu kabul edilir. Bağımlı değişken eksi sonsuz ve artı sosuz arasında değer alabilmektedir. Oysa sosyal bilimlerde bağımlı değişkenin nitel olduğu birçok araştırma söz konusu olmaktadır. Örneğin bağımlı değişkenin oy verip vermeme olarak, herhangi bir anlaşmaya girilip girilmemesi, gibi nitelendirildiği durumlarda regresyon analizi nitel bağımlı değişkenden dolayı işe yaramamaktadır. Bu durum sosyal bilimlerde birçok araştırma konusunun nasıl modellenmesi gerektiğine dair araştırmalar başlatmıştır.

Lojistik regresyonun kullanmasına karar verildiğinde bağımsız değişkenlerin dağılımı hakkında herhangi bir varsayımda bulunmaya gerek yoktur. Bu sebeple bağımsız değişkenler çoklu normal dağılıma uygun olmadığında lojistik regresyonun kullanılması tavsiye edilir[16].

Bağımlı değişkenin ölçüm şekli kesikli, kategorik, sıralı olduğu zaman kullanılan analiz yöntemleri "ikili veri analizi", "kategorik analiz" veya "lojit" olarak adlandırılır. Bu istatistiksel metotların ortak noktası olayın olma olasılığını vermektedirler.

---

[16] Sharma. a.g.e. sy. 317

Lojistik regresyon kullanılmaya başlanmadan önce doğrusal olasılık modeli kullanılmaya başlanmıştır.

2.1 Doğrusal Olasılık Fonksiyonu

Y'nin rasgele bir ikili (binary) değişken olduğu düşünülecek olursa. Y'nin iki şekilde yani başarılı veya başarısız şeklinde sonuçlanması mümkündür[17]. Y değişkeni Bernolli dağılımına sahiptir. Bu dağılımın ortalaması ve varyansı hesaplanacak olursa:

$$E(Y) = 1 * P(Y=1) + 0 * P(Y=0) = P(Y=1) = p \qquad (25)$$

$$V(Y) = E(Y^2) - [E(Y)]^2 = P(Y=1)[1-P(Y=1)]$$
$$= p * (1-p) \qquad (26)$$

olacaktır. Eğer Y'nin iki değer aldığı düşünülecek olursa hata terimleriyle ilgili aşağıdaki eşitlikler yazılabilir.

Eğer Y=1 ise;

$$1 = \sum b_k X_{ik} + u \text{ veya } u = 1 - \sum b_k X_{ik} \qquad (27)$$

eğer Y=0 ise;

$$0 = \sum b_k X_{ik} + u \text{ veya } u = -\sum b_k X_{ik} \qquad (28)$$

olacaktır.

---

[17] Aldrich, J. H. , Nelson, D. F, **Linear probability, Logit, and Probit Models** , Sage Pub, 1984 Sy.14

Hata terimleri ile ilgili yapılan varsayım beklenen değerinin sıfır olmasıdır. Bu varsayım gerçekleştiğinde,

$$E(u) = P(Y=0) * [-\sum b_k X_{ik}] + P(Y=1)* [1 - \sum b_k X_{ik}]$$
$$= -[1-P(Y=1)] * [-\sum b_k X_{ik}] + P(Y=1)* [1 - \sum b_k X_{ik}]$$
$$= 0. \qquad (29)$$

olacağından tahmin edilen katsayılar yansız olacak fakat hata terimlerinin sabit bir varyansı olacağına dair kabullenme doğrulanamayacağından tahmin edilen katsayılar en iyi olmayacaktır.

$$V(u) = [\sum b_k X_{ik}] * [1 - \sum b_k X_{ik}] \qquad (30)$$

Varyansın sabit olmaması nedeniyle hipotez testleri, güven aralıkları geçerliliklerini kaybedeceklerdir. Bu durumdan kurtulmak için Goldberger[18] tarafından önerilen ağırlıklı tahmin yöntemi kullanılabilir. Varyansın sabit hale getirilmesi için gözlenen değerler

$$W_i = ([\sum b_k X_{ik}] * [1 - \sum b_k X_{ik}])^{1/2} \qquad (31)$$

katsayısına bölünür ve yeni değerlerle regresyon tahminleri yapılır ise elde edilen katsayılar hem yansız hem de sabit varyanslı olurlar.

2.1.1 Doğrusallık Varsayımı

$$P(Y=1) = E(Y) = b_0 + b_1 X_1 \qquad (32)$$

[18] Goldberger, A. S., Econometric theory, 1964. New York: John Wiley.

denkleminin en küçük kareler yöntemine göre tahmincilerinin bulunabilmesi için modeldeki ilişkinin doğrusal olması gerekmektedir. En küçük kareler yöntemi kullanılırken varyansın sabit olmaması sorununun giderildiği düşünülmektedir. Oluşturulan modelin sol tarafının olasılık olduğu düşünülürse sol tarafın 0 ile 1 arasında olması gerekmektedir. Fakat bazı durumlarda mesela bağımsız değişkenin alabileceği büyük bir değerde katsayısının negatif olması durumunda sol tarafın sıfırdan küçük çıkması veya katsayının pozitif olması durumunda ise birden büyük çıkması durumu söz konusu olabilir. Bu durum doğrusal olasılık modelin önemli eksiklerindendir.

İkinci bir sorun ise bağımsız değişkenlerdeki değişimlerin modeldeki doğrusallık nedeni ile sonuca marjinal katkısının sabit olmasıdır. Mesela modeldeki bağımsız değişkenin gelir olması, bağımlı değişkenin ise ev sahipliğini göstermesi durumda aylık geliri 1 milyar olan bir kişinin ev sahibi olması olasılığı ile aylık geliri 50 milyar olan kişinin ev sahibi olması olasılığı hesaplanırken gelir değişkenin önündeki katsayı aynı ölçüde olasılığı artıracaktır.

2.2 Lojistik Regresyon Modeli

Yukarıda belirtilen problemler yüzünden doğrusal olasılık modelinin yerine doğrusal olmayan bir modelin oluşturulmasına ihtiyaç duyulmaktadır. Doğrusal olasılık modeline karşı popüler olan yöntemlerden biri lojistik regresyon, kısaca lojit[19], modeli anlatılacaktır. Diskriminant analizinde olduğu gibi bağımlı değişkenin nitel olması durumunda kullanılan lojistik regresyon yönteminin bazı avantajları vardır[20]. Diskriminant analizinde değerlerin hesaplanması, kritik değerin belirlenmesi ve daha sonra gruplama işlemi yapılırken, lojistik regresyonda doğrudan olasılık değerleri bulunmaktadır. Bulunan değer olasılık değeri olacağından dolayı 0 ile 1 arasında olmasını sağlayacak bir S-şekilli fonksiyon kullanılacaktır.

[19] Nelson, Aldrich: a.g.e; sy. 48
[20] Hair ve diğerleri, **Multivariate data analysis** a.g.e: sy 277

2.2.1 Modelle ilgili varsayımlar

Bu modelde bağımlı değişken Y'nin ikişerli ayrık olduğu varsayımı yapılır. Modelde bağımlı değişkenin 1 ile ifade edilen sonucu, yani olumlu olması ile ilgilenilmektedir. Y bağımlı değişkeninin k tane bağımsız değişkene bağımlı olduğu varsayımı yapılır. Bağımsız değişkenlerin hiç birisi diğeri ile tam bir doğrusal bağlantıya sahip olmamalıdır. Bağımlı değişken değerleri birbirlerinden bağımsızdır. Lojistik regresyon fonksiyonu Lojit olarak adlandırılır ve aşağıdaki gibi ifade edilir.

$$P(Y_i = 1 \mid X_i) = \frac{\exp(\alpha + \beta x)}{1 + \exp(\alpha + \beta x)} \qquad (33)$$

2.2.2 Model parametrelerinin tahmini

Modelin parametrelerinin tahmini[21] için maksimum benzerlik yöntemi (Maximum Likelihood Estimation) kullanılmaktadır. Bağımlı değişken değerleri birbirlerinden bağımsız olduğundan gözlenecek N tane Y değişkenin olma olasılığının her birinin ayrı ayrı olma olasılıklarının çarpımı olacağı bilinmektedir. Bu olasılık ise

$$P(Y \mid X) = \prod_{i=1}^{N} P_i^{Y_i} (1 - P_i)^{1-Y_i} \qquad (34)$$

şeklinde ifade edilecektir. Benzerlik fonksiyonu kullanarak L(Y|X) şeklinde ifade edilebilir. L(Y|X) ise,

[21] Aldrich, Nelson: a.g.e. , sy 51

$$L(Y \mid X) = \prod_{i=1}^{N} \left[\frac{\exp(\sum b_k X_{ik})}{1+\exp(\sum b_k X_{ik})} \right]^{Y_i} \left[\frac{1}{1+\exp(\sum b_k X_{ik})} \right]^{1-Y_i} \quad (35)$$

şeklinde ifade edilebilir. Toplamlarla işlem yapmak daha kolay olduğundan benzerlik fonksiyonunun logaritması alınır ve

$$\log L(Y \mid X) = \sum_{i=1}^{N} [Y_i \log P_i + (1-Y_i)\log(1-P_i)] \quad (36)$$

eşitliğine dönüştürülür. Bu fonksiyon kullanılarak her bir değişkenine göre birinci dereceden türev alınarak sıfıra eşitlenir. Fakat bulunan denklemlerin analitik sonuçları bulunamayacağından "Newton-Rapson tekrarlı tekniği" kullanılarak katsayılar bulunur[22].

2.2.3 Sonuçların Yorumu

Lojistik regresyon sonuçları bulunduğunda iki sonuç önem taşımaktadır. Birincisi bütün olarak modelin anlamlılığını test eden logaritmik benzerlik katsayısı. İkincisi ise her bir değişkenin anlamlığını test eden Wald istatistiği[23].

Doğrusal regresyondaki F ve R^2 değerine yakın anlam veren logaritmik benzerlik (LL olarak kısaltılacaktır) katsayısı istatistik paket programlarında –2 ile çarpılarak verilir. Bunun sebebi –2LL değişkenin ki-kare dağılımına sahip olmasıdır.

-2LL katsayı modelin verilmiş olan katsayılar ile ne kadar iyi tahmin edilmiş olduğunu gösterir. Bu değer hesaplanırken modelde sadece sabit terim varken hesaplanacak daha sonrada bağımsız değişkenler sıra ile modele konacak ve "-2LL"

---

[22] Aldrich, Nelson: a.g.e., sy

[23] Menard, S. **Applied Logistic Regression Analysis**, Sage Pub, 1995, Sy17

değerinde belli oranda düşme sağlanır ise bağımsız değişken modele eklenmiş olacaktır. En iyi model ise en küçük –2LL değerine sahip değişkenlerden oluşacaktır. İlk elde edilen –2LL değeri ile en son elde edilen -2LL değeri arasındaki fark alınacak ve bu değer ki-kare dağılımında olacak ve serbestlik derecesi de iki modelde kullanılan değişken sayılarının farkı kadar olacaktır.

Modelin anlamlılığı –2LL katsayısı ile belirlendikten sonra her bir değişkenin katsayısının anlamlılığına bakılabilir. Modelde bulunacak olan bağımsız değişkenlerin anlamlılık testi ise Wald istatistiğini kullanacaktır. Wald istatistiği ki-kare dağılımına sahiptir[24].

2.2.4 Katsayıların Yorumu

Lojistik regresyon sonucunda iki katsayının yorumu önemlidir. Bunlar standart olmayan lojistik ve standart lojistik katsayılarıdır. Regresyon analizinde olduğu gibi standart olmayan katsayı bağımsız değişkendeki bir birimlik artışın bağımlı değişkende oluşturacağı artışı veya azalışı gösterecektir. Fakat lojistik regresyondaki katsayının yorumunda dikkat edilmesi gereken nokta değişkenler arasındaki ilişkinin doğrusal değil eğrisel bir ilişki olduğudur. Katsayılar regresyon analizinde olduğu gibi her bir bağımsız değişkendeki bir birimlik artışın bağımlı değişkende yarattığı değişimi gösterse de eğrinin eğiminin her noktada değiştiği unutulmamalıdır[25].

Standart lojistik katsayısı ise modelde kullanılan her bir bağımsız değişkenin ölçümü birbirinden farklı ise, mesela kredi almak için başvuran bir şirketin her hangi bir mali oranı ile aynı şirkette çalışan kişi sayısı modelimizdeki açıklayıcı değişkenler olsun. Mali orandaki bir birimlik artış ile çalışan sayısındaki bir birimlik artış aynı birimlerde ölçülmediği için anlamları farklı olacaktır. Doğrusal regresyonda standart katsayılar her bir değişkenin standartlaştırılması sonucu elde edilen, yani değişkenler

---

[24] Cizek. G.J, Fitzgerald. S. M., "An introduction to logistic regression", **Measurement&Evaluation in Counseling&Development**. 31, 1999, 223-245

[25] Menard a.g.e sy 43.

kullanılarak elde edilen katsayılardır. Standartlaştırma işlemi yapılırken her bir gözlem kendi ortalamasından çıkartılır ve standart sapmasına bölünür. Oluşturulan yeni değişkenler ile elde edilen katsayılar standart katsayılar olacaktır. Lojistik regresyonda standart katsayıların elde edilmesi biraz karışıktır. Bunun sebebi lojistik regresyon sonucunda Y nin değerinin hesaplanması değildir. Lojistik regresyonda Y'nin olasılık değeri hesaplanmaktadır. Lojistik regresyon sonucunda hesaplanan değer Y değil lojit değeridir ve lojit değeri $-\infty$ ile ∞ arasında değişmektedir. Dolayısıyla değişkenin ortalamasının yada standart sapmasının hesaplanması mümkün değildir. Direkt olarak standart katsayıyı hesaplamamız mümkün değilse de dolaylı olarak hesaplamak mümkündür. Bu hesaplamayı yaparken aşağıdaki formüller kullanılacaktır,

$$R^2 = \frac{SSR}{SST} \qquad (37)$$

SSR regresyonda sapma kareleri Toplamını, SST ise hata karelerinin Toplamıdır. Pay ve payda gözlem sayısına bölünerek

$$R^2 = \frac{SSR \ / N}{SST \ / N} = \frac{s^2_{\hat{Y}}}{s^2_Y} \qquad (38)$$

bu denklem üzerinden bazı hesaplamalar yapılarak standart katsayıları elde edilir. Standart katsayıların hesabı aşağıdaki formülle hesaplanır.

$$b^*_{YX} = \frac{(b_{YX})(s_X)(R)^2}{s_{lojit(\hat{Y})}} \qquad (39)$$

b^*_{YX} standart katsayıyı,

b_{YX} standart olmayan katsayıyı,

s_X bağımsız değişkenin standart sapmasını,

R^2 belirginlik katsayısı

$S_{Logit}\ (\hat{Y})$ ise lojit sonucu bulunan değerlerin standart sapmasını,

göstermektedir.

Standart katsayının yorumu ise şöyle yapılır. Bağımsız değişkendeki bir standart sapmalık artış, lojit (Y)'de b* 'lık bir artış veya azalışa sebep olur[26].

2.2.5 Sınıflama Kriteri

Katsayılar hesaplandıktan sonra bağımsız değişkenler otuz üç numaralı denklemde yerine konularak

$$P(Y_i = 1 | X_i) = \frac{\exp(\alpha + \beta x)}{1 + \exp(\alpha + \beta x)} \qquad (40)$$

bir olasılık değeri hesaplanır. Bulunan bu olasılık değeri 0,5 den büyük ise başarılı grup içerisinde yer alırken, hesaplanan olasılık değerinin 0,5 den küçük olması durumunda ise başarısız grup içerisinde olduğu kabul edilir.

[26] Menard a.g.e sy 45-47

III. YENİDEN ÖRNEKLEME TEKNİKLERİ

3.1 Yeniden Örnekleme Kavramı

İstatistik deneylerden veya gözlemlerden olaylar, durumlar hakkında bilgi edinme bilimidir. İstatistik biliminin kökenleri 1650'lere dayanmaktadır. Fakat istatistiğin önemi özellikle yaşadığımız son otuz yılda ortaya çıkmaya başlamış ve diğer bir çok bilim dalına yardımcı bilim dalı haline gelmiştir. Psikoloji, eğitim, iktisat, sosyoloji, genetik, fizik, astronomi gibi bilimler istatistiği yaygın olarak kullanmaya başlamışlardır. İstatistik teorisi temel olarak üç soruya cevap vermeye çalışır[27]:

(1) Veri seti nasıl toplanmalıdır?
(2) Toplanan bu veri seti nasıl analiz edilmeli ve sunulmalıdır?
(3) Özet istatistikler ne kadar doğrudur?

Yeniden örnekleme teknikleri üçüncü sorunun cevabını bulmak için geliştirilmiş bir tekniktir. Bootstrap tekniğinin son zamanlarda gelişmiş olmasının sebebi ise gelişmiş bilgisayarların istatistik alanında kullanılması ile olmuştur. İstatistik teorisi birikmiş verilerden en anlamlı bilgiyi çıkarmaya çalışmaktadır. İstatistikte vardamanın matematiksel alt yapısını ihtimal teorisi oluşturmaktadır. Gözlem yapılarak elde edilen verilerin üzerinden karar verme durumunda kalındığında elde edilen bilginin mutlak kesin bir bilgi olmadığı bilinir. Çünkü gözlemler üzerinden hareketle ana kütle parametresini takdir etmek gerekmektedir.

Uygulamalı istatistiğin[28] en önemli iki sorunu herhangi bir karakteristik için takdir değerinin ve bu takdir değeri için standart hata değerinin bulunması ve güven aralığının hesaplanmasıdır. Yeniden örnekleme teknikleri sadece takdir değerlerini ve

---

[27] Efron and Tibshirani, **Introduction to the Bootstap**. Chapman & Hall 1993. sy 1

[28] Cherney. **Bootstrap Methods: A Practitioner's Guide** . John Willey & Sons Inc, 1999. sy6

standart hataları belirlemenin dışında bir çok alan için kullanma imkanı vardır: Regresyon, Zaman Serileri Analizleri, Doğrusal Olmayan Regresyon, Kümeleme, Diskriminant Analizinde hatalı sınıflama matrisini oluşturma, Lojistik Regresyon ve her türlü hipotez sınamasında kullanılabilirler.

Yeniden örnekleme tekniği örneklem değerlerini kullanarak her hangi bir istatistiğin dağılımını bulur. İki çeşit yeniden örnekleme tekniğinden söz edilebilir: ilki bootstrap ikincisi ise çakı tekniğidir. Bootstrap tekniği ise iki farklı şekilde bilinir. İlki "parametrik bootstrap tekniği", ikincisi ise "parametrik olmayan bootstrap tekniğidir". Parametrik bootstrap tekniği kullanılmadan önce örnekleme dağılımı için varsayım yapılır. Bunlar örneğin normal dağılım varsayımı durumunda iki tane parametre gerekliyken, bernoulli dağılımı için bir parametre gerekmektedir. Parametrik olmayan tekniğin kullanımında ise örnekten yola çıkılarak iadeli örnekleme yöntemi ile istatistik takdir edilir ve istatistiğin dağılımı bulunmaya çalışılır. Çakı yönteminde ise her örnekten bir gözlem çıkarılarak kalan örneklem için parametre takdir edilir ve bu takdir değerlerinin dağılımına bakılır.

Yeniden örnekleme tekniklerinden olan çakı yönteminin başlangıcı 1949'a kadar dayanmaktadır. Ancak yeniden örnekleme tekniğinin istatistik ve diğer bilimsel çevrelerce genel kabul gören bir teknik olması için 1979 yılında Efron tarafından yazılan makale başlangıç olmuştur[29].

3.1.1 Çakı Tekniği (Jackknife)

Gözlemlenmiş değerlerin $\mathbf{X} = (x_1, x_2, \ldots x_n)$ olduğunu ve takdir edilecek olan istatistiğin $\hat{\theta} = s(x)$ olduğunu kabul edelim. Çakı yönteminde gözlemlenmiş veri setinden her seferinde bir gözlem dışarıda bırakılarak istatistik takdir edilir.

[29] Chernick. a.g.e. sy1

$x_i = (x_1, x_2, ...x_{i-1}, x_{i+1}, ..., x_n)$ i = 1,2,.....n. Örneklerinden her birine çakı örneği denir. Her bir çakı örneği için takdir edilecek olan istatistik ise $\hat{\theta}_{(i)} = s(x_{(i)})$ olarak gösterilir[30].

Takdir edilen istatistik değerinin standart hatası aşağıdaki gibi hesaplanır.

$$s\hat{e}_{çakı} = \left\{ \frac{n-1}{n} \sum (\hat{\theta}_{(i)} - \hat{\theta}_{(.)})^2 \right\}^{1/2} \qquad (41)$$

$$\hat{\theta}\ (.) = \sum_{i=1}^{n} \hat{\theta}_{(i)} / n \qquad (42)$$

Çakı yöntemi bootstrap yönteminin bir yakınsamasıdır[31]. Çakı yöntemi gözlem sayısının 100'den az olduğu durumlarda hesaplanması kolay olduğundan standart hata takdiri için kullanılabilmektedir. Çakı yönteminin bootstrap yöntemine göre zayıf kalmasının sebeplerinden biri şudur. Takdir edilmeye çalışılan istatistik örneğin ortanca gibi değiştirilmesi zor bir istatistik ise çakı yönteminde takdir edilen değerlerin değişmeden kalması söz konusu olacaktır[32].

3.1.2 Bootstrap Tekniği

Bootstrap metodu günlük hayattaki tanımını; gözlemlenmiş olan *n* tane veriyi kullanarak ana kütle karakteristiği hakkında hipotez testi yapmak veya bu karakteristik için güven aralığı takdirinde bulunmak olarak tanımlanabilir. Fakat *n* gözlem için bunu yapabilmek bazı varsayımlar yapılmasını gerektirecektir, bootstrap bu kabullenmeler yapılamadığında devreye girerek karakteristik hakkında işlemler yapılmasını sağlar.

---

[30] Efron. a.g.e sy 141
[31] Efron. a.g.e. sy145
[32] Efron. a.g.e. sy148

Tablo 2. Bootstrap Tekniğinin Kullanıldığı İlgili Çalışmalar

İşlem kapasite indeksi	Choi, Nam, Park	1996
Kalite kontrol	Liu, Tang	1996
Güvenirlilik analizi	Chao, Huwang	1987
Meteoroloji	Robeson	1995
İktisat	Tambour, Zethraecus	1998
Kimya	Roy	1994
Fizik	Das Peddala, Chang	1992
Ekoloji	Adams, Gurevitch Rosenberg	1997
Biyoloji	Brey	1990

Tablo 2'de görülebileceği gibi bootstrap bir çok alanda uygulanma imkanı olan bir istatistik tekniğidir. Tablo 2'de yer alan uygulamalar bootstrap tekniğinin kullanıldığı bazı çalışmaları göstermektedir. Diğer alanlarla ilgili Cherney'in Bootstrap Methods: A Practitioner's Guide isimli eseri yararlı bir kaynaktır.

Rasgele gözlemlenmiş olan n tane verinin ihtimal dağılımı F ise,

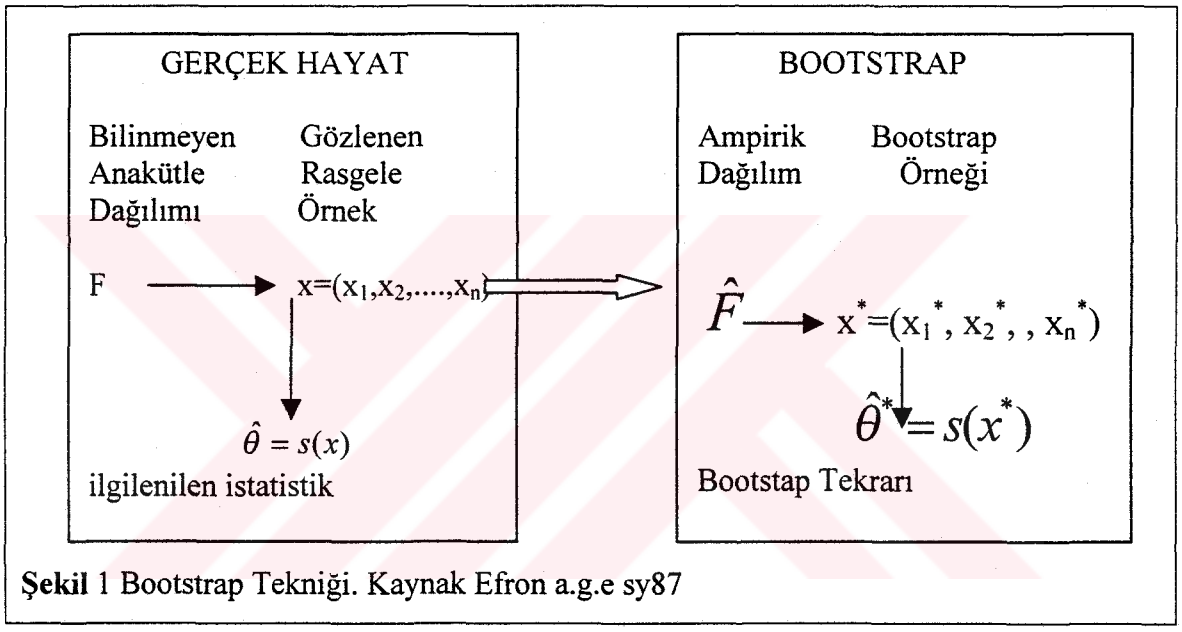
$$F \rightarrow (x_1, \quad x_2 \quad \ldots\ldots\ldots \quad x_n) \qquad (43)$$

olarak gösterilecek ve her bir gözlemin 1/n olasılığıyla seçilebileceği dağılıma ampirik dağılım denecek ve $\hat{F}$ olarak gösterilecektir.

Bootstrap yöntemi bootstrap örneklerine bağlıdır. $\hat{F}$ seçilmiş bir ampirik dağılım olduğu varsayılırsa bootstrap örneği gözlemlenmiş veri setinden iadeli olarak rasgele seçilmiş olan bir örnek olacak ve

$$x^* = \left(x_1^*, x_2^*, \ldots\ldots\ x_n^*\right) \qquad (44)$$

şeklinde gösterilecektir. Sol üst köşede olan yıldız işareti gözlemlenmiş gerçek değerlerin içerisinden iadeli seçim yapılarak oluşturulan örneği temsil etmektedir. Şekil 1 bootstap yönteminin mantığını göstermektedir.



Şekil 1 Bootstrap Tekniği. Kaynak Efron a.g.e sy87

Şekil 1'de tek örnekli bir problem için bootstrap tekniğinin nasıl uygulandığını gösterilmektedir. Sol tarafta gerçek hayatta karşılaşılan ve dağılımının ne olduğu bilinmeyen F dağılımı için gözlemlenmiş olan veri seti kullanılarak elde edilmeye çalışılan istatistik değeri gösterilmektedir. Sağ tarafta ise bootstrap tekniği kullanılarak gözlemlenmiş olan değerler kullanılarak yaratılan veri setlerinden elde edilen istatistik değeri gösterilmektedir. Gerçek hayatta tek bir veri seti ile karakteristik tahmin edilmeye çalışılırken, bootstrap tekniğinde iadeli rasgele seçim uygulanarak elde edilen istenilen sayıdaki veri setinden karakteristik tahmin edilmeye çalışılmaktadır.

3.2 Tek Değişken İçin Standard Hata Takdiri

Elde dağılımı bilinmeyen rasgele bir örnek olduğu düşünülürse. $\mathbf{X} = (x_1, x_2,....x_n)$ dağılımında bilinmeyen bu gözlemler için standart sapmanın hesaplanmasında bootstrap yöntemi kullanılırken aşağıdaki algoritma takip edilecektir[33].

(1) B tane birbirinden bağımsız bootstrap örneği , $\mathbf{x}^{*1}$, $\mathbf{x}^{*2}$,....$\mathbf{x}^{*B}$, her bir örneklem *n* tane gözlem içermektedir.

(2) Her bir örnek için standart sapma hesaplanacaktır.

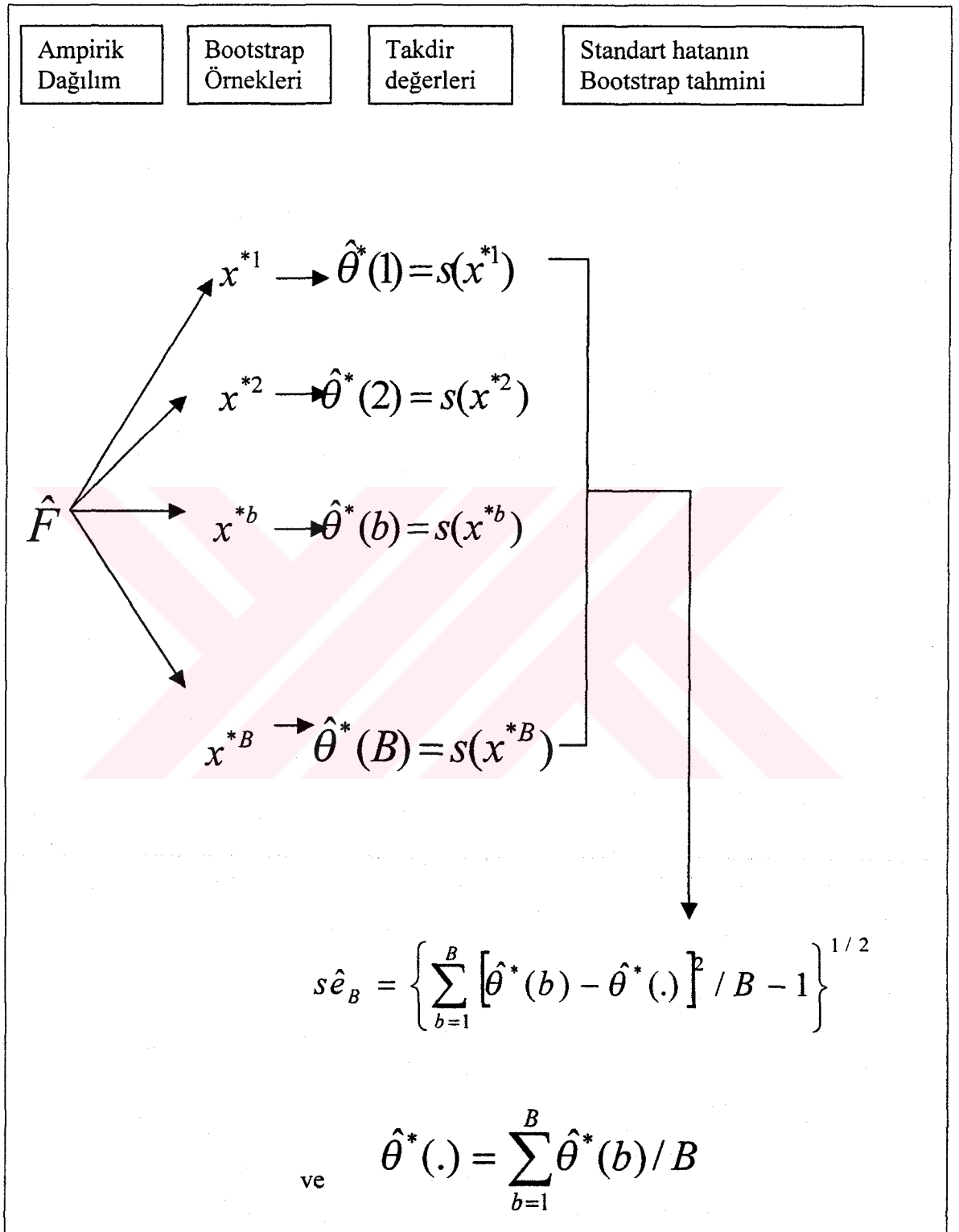
$$\hat{\theta}^*(b) = s(x^{*b}) \qquad b = 1,2,....B \qquad (45)$$

s standart sapmadır.

(3) Standart hata ise her bir standart sapma kullanılarak hesaplanacaktır.

$$s\hat{e}_B = \left\{ \sum_{b=1}^{B} \left[\hat{\theta}^*(b) - \hat{\theta}^*(.) \right]^2 / B - 1 \right\}^{1/2} \qquad (46)$$

[33] Efron. A.g.e sy 45



Şekil 2. Standart Hata Tahmini

Şekil 2 de bu algoritma daha detaylı standart hatanın bootstrap yöntemi kullanılarak hesaplanması anlatılmaktadır. İlk olarak gözlemlenmiş değerlerden bootstrap örnekleri oluşturulmuştur. Her bir örnek için bir takdir değeri hesaplanmıştır. Hesaplanan standart sapma değerlerinin ortalaması bulunarak standart sapmaların ortalaması hesaplanmıştır. Bu aşamadan sonra ise her bir takdir değerinden hesaplanan ortalama standart sapma değerinin farklarının kareleri alınarak sapmalar bulunur. Toplam sapmaların kareleri bootstrap örnek sayısının bir eksiğine bölünerek kare kökü alındığında standart hata terimi hesaplanmış olur.

3.3 İki Örnekli Veri Setinde Bootstrap Tekniği

Tek değişkenin olduğu durumda bootstrap tekniğini kullanmak oldukça kolay iken iki değişkenin olduğu veri setinde bootstrap tekniğini kullanırken daha dikkatli olunması gerekecektir. $P \rightarrow x$ gösteriminden bilinmeyen bir ihtimal modeli olan P 'nin gözlenen veri seti x anlaşılmalıdır. P nin F ve G gibi iki tane ihtimal dağılımından oluştuğu düşünülecek olursa. $P = (F,G)$ olarak ifade edilecek ve $P \rightarrow x$ ifadesindeki x veri setini (z,y) şeklinde düşünmemiz gerekecektir. Z veri seti F dağılımdan y veri seti ise G dağılımından gelecektir. Bootstrap tekniğini bu tür bir veri setinde kullanmak için veri setleri (z,y) ayrı ayrı düşünülüp kendi bootstrap örnekleri oluşturularak daha sonradan birleştirilmesi sonucu x veri seti elde edilecektir[34]. $\hat{P} = (\hat{F},\hat{G})$ olarak ifade edilecek ve

$$x^* = (z^*, y^*) \qquad (47)$$

böylece oluşturulacak her bir veri seti aşağıdaki gibi ifade edilecektir

[34] Efron. a.g.e sy88

$$x^* = (z^*, y^*) = (z_{i1}, z_{i2}, \ldots\ldots z_{in}, y_{j1}, y_{j2}, \ldots\ldots y_{jm}) \quad (48)$$

Toplam gözlem sayısı $n+m$ tane olmasına rağmen tek değişkenli bootstrap tekniğinden farklı olarak n gözlem kendi içerisinde, m gözlem ise kendi içerisinde iadeli olarak rasgele seçime tabi tutulacaktır.

3.4 Regresyon Analizinde Bootstrap Tekniğinin Kullanılması

Doğrusal regresyon analizi tüm bilim dallarında en çok kullanılan ve en yarar sağlayan istatistik yöntem olmuştur. Doğrusal regresyonda kullanılan veri seti n tane gözlem değeri içermektedir, $x_1, x_2, \ldots\ldots, x_n$. Her bir x_i gözlem değerinin iki tane değişkeni vardır.

$$x_i = (c_i, y_i) \quad (49)$$

y_i değişkeni bağımlı değişken olarak bilinirken, c_i değişkeni bağımsız değişkenlerin oluşturduğu matris olarak bilinir. Bağımsız değişkenlerin değerinin bilindiği durumda bağımlı değişkenin beklenen değeri μ_i ise,

$$\mu_i = E(y_i \mid c_i) \qquad (i = 1,2,\ldots\ldots,n) \qquad (50)$$

doğrusal regresyon modelinin en önemli varsayımı μ_i ın bağımsız değişkenlerin doğrusal bir fonksiyonu olduğudur.

$$\mu_i = c_i \beta = \sum_{j=1}^{p} c_{ij} \beta_j \qquad (51)$$

regresyon parametreleri, $\beta = (\beta_1, \beta_2, ..., \beta_p)^T$ bilinmemektedir ve analizin amacı bilinmeyen bu parametreler için bağımsız değişkenleri kullanarak takdirde bulunmaktır. Doğrusal modelin olasılık yapısı aşağıdaki gibi ifade edilir,

$$y_i = c_i \beta + \varepsilon_i \quad (i = 1,2,....,n) \qquad (52)$$

bu modeldeki parametreler için en küçük kareler yöntemi kullanılarak bulunacak olan takdir değerleri,

$$\hat{\beta} = (C^T C)^{-1} C^T y \qquad (53)$$

olarak bulunur.

Bootstrap tekniğinin regresyon analizinde kullanılması iki şekilde olabilir. Birinci yöntemde, doğrusal regresyonun olasılık modelini $P \rightarrow x$ olarak yazmak mümkündür. P ise iki parçadan oluşmaktadır,

$$\boldsymbol{P = (\beta, F)} \qquad (54)$$

β değerleri regresyon katsayılarını, **F** ise (52) nolu denklemdeki hata terimlerinin dağılımını gösterir. β değerleri bilinmediğinden en küçük kareler yöntemine göre takdir

edilen $\hat{\beta}$ değerlerini kullanarak hata terimlerini bulup, hata terimlerinin ampirik dağılımı elde edilebilir.

$$\hat{\varepsilon}_i = y_i - c_i\hat{\beta} \quad (i = 1,2,....,n) \qquad (55)$$

Hata terimleri bulunduktan sonra iadeli rasgele seçim yapılarak bootstrap örnekleri oluşturulur. Bağımsız değişkenler bilindiğinden (53) denklemde hata terimlerinin oluşturduğu bootstrap örnekleri yerine konularak bağımlı değişken değerleri bulunabilir.

$$y^*_i = c_i\beta + \varepsilon^*_i \quad (i = 1,2,....,n) \qquad (56)$$

Bağımlı değişkenlerin bulunması sırasında bağımsız değişkenlerin değişmediği kabul edilmiştir. β katsayılarının hesabında ise aşağıdaki formül kullanılır,

$$\hat{\beta}^* = (C^T C)^{-1} C^T y^* \qquad (57)$$

Her bir bootstrap örneğinde regresyon parametreleri için takdir değerleri hesaplanır.

İkinci yöntemde ise, 4.3'de anlatıldığı gibi x veri setinin bootstrap işlemine girmesi olabilir. Bu bölümde x veri setinin (c_i, y_i) şeklinde olduğu ve c'nin bağımsız değişkenlerin matrisi, y ise bağımlı değişkeni temsil etmekteydi.

$x_i = (c_i, y_i)$ ise, bootstrap tekniği sonrasında oluşacak olan her bir veri seti,

$$x^* = \{(c_{i1}, y_{i1}), (c_{i2}, y_{i2}), \ldots\ldots, (c_{in}, y_{in})\} \quad (58)$$

$i_1,\ldots,i_n$ 1'den n'e kadar olan rasgele seçilmiş olan örnekleri göstermektedir. Her bir bootstrap örneği için hesaplanacak olan regresyon katsayıları,

$$\hat{\beta}^* = (C^{*T}C^*)^{-1}C^{*T}y^* \quad (59)$$

şeklinde olacaktır.

Hangi bootstrap yönteminin daha uygun olduğu sorusunun cevabı, oluşturulan doğrusal regresyon modelinin ne kadar doğru olduğuna bağlıdır. (53) nolu denklemde oluşturulan regresyon modeli hata terimlerinin bağımsız değişkenlere göre değişmediğini varsaymaktadır. Bağımsız değişkenler ne olursa olsun hata terimlerinin dağılımı değişmemektedir. Bu varsayım oldukça güçlü ve gerçekleşmesi zayıf bir varsayımdır. Veri setinin bootstrap yöntemi ile seçilmesi hata terimlerinin seçilmesinden daha güvenilirdir. Hangi yöntemin kullanılacağına karar verilirken bağımsız değişkenlerin sabit olup olmadığına dikkat edilirse yöntem seçilmiş olacaktır. Bağımsız değişkenler sabit olarak kabul edildiği durumlarda hata terimleri yöntemini kullanmak daha iyi sonuçlar verirken, bağımsız değişkenler rasgele seçiliyor ise ikinci yöntem olan x veri setinin kullanılması daha iyi sonuç verecektir[35].

[35] Efron, a.g.e sy 105-115

IV. KREDİ PUANININ BELİRLENMESİ

4.1 Kredi Puanlamasının Tarihsel Gelişimi

Her hangi bir anakütleden grupların ayrıştırılması tekniği ilk olarak bir istatistikçi olan Fisher (1936)[36] tarafından ortaya atılmıştır. Durand (1941)[37] ilk kez Fisher tarafından düşünülen bu yöntemin iyi ve kötü kredilerin ayrıştırılması için kullanıla bileceğini düşünmüştür. Durand tarafından önerilen yönteme kadar kararlar kişisel deneyimlere bakılarak alınıyordu. Ancak bu konuda bazı kuralların konulması gerekiyordu ve ilk uygulama San Fransisco' da Bill Fair ve Earl Isaac tarafından 1950'li yılların başında uygulanmaya başlandı[38]. Kredi kartlarının 1960'lı yılların başında ortaya çıkmasıyla birlikte bankalar ve diğer kredi veren kuruluşlar için kredi risklerinin hesaplanmasında ne kadar önemli olduğu anlaşılmıştır. Kredi kartı başvurusunun fazla olması ve kredi riskinin kişileriler tarafından değerlendirilme sürecini deneyimlere dayandığı için sonuçların güvenirliği konusunda soru işaretleri oluşturuyordu. Myers ve Frogy (1963)[39] yaptıkları çalışmada kredi riski puanlamasının kullanılması durumunda batık kredilerin %50 oranında azaldığını tespit etmişlerdir. Capon (1982)[40] bu çalışmalara ilk karşı çıkan kişi olmuştur. Capon kredi değerlendirilmesinde cinsiyet, din, etnik köken gibi bazı değişkenlerin kullanılmasının ayrımcılık olduğunu savunmuştur. 1980'lere gelindiğinde ise bankalar kredi kartlarında bu sistemi tam anlamıyla kullanmaya başlamışlardır, hatta ev, araba, ihtiyaç kredileri gibi bireysel ve küçük firmalara verilen krediler, bu sistem sayesinde karara bağlanmaya başlanmıştır.

Finansal riskin tahmini istatistik ve olasılık modellerinin son dönemlerde en önemli ilgi alanı olmuştur. Finansal risk dendiğinde portföy yönetimi, opsiyon

---

[36] Fisher R. A., The use of multiple measurements in taxonomic problems, **Annals of Eugenics** 7, 1936,179-188

[37] Durand D., Risk elements in consumer instalment financing, **National Bureau of Economic Research**, New York, 1941

[38] Thomas L. C., _A survey of credit and behavioural scoring: forecasting financial risk of lending to consumerd, **International Journal of Forecasting**, 16, 2000, 149-172

[39] Myers J.H., Forgy E. W., The development of numerical credit evaluation systems, **Journal of American Statistical Association** 58, 1963, 799-806

[40] Capon N., Credit scoring systems: a critical analysis, **Journal of Marketing**, 46,1982, 82-91

fiyatlaması, bono fiyatlaması ve daha az bilinen kredi riskinin hesaplanması kastedilmektedir. Kredi riskinin hesaplanması iki şekilde yapılabilmektedir. Kredinin tahsis edilip edilmeyeceği kararının verilmesi için başvuru değeri veya kredi başvuru değeri, ikincisi ise kredinin uzatılıp uzatılmaması kararının verilmesi, davranış değerinin hesaplanmasıdır (Behavioural Scoring). Bu tezin konusu başvuru değeri için bir model oluşturmaktır. Davranış değerinin hesaplanması başvuru değerinin hesaplanması ile aynı temele dayanmaktadır. Davranış değeri hesaplanırken bir grup müşteri örnek alınarak bu müşterilerin son dönemlerde yaptıkları ödeme davranışlarına göre kredinin geri ödemesinin mümkün olup olmadığı tahmin edilmeye çalışılır. Hopper ve Lewis (1992)[41] ile Wynn ve McNab (2000)[42] davranış değerinin uygulamada nasıl kullanılacağını açıklamışlardır.

Puanlama modelleri ikiye ayrılmaktadır: 1) başvuru puanı, 2) davranışsal puanlama. Kredi için başvuran müşteriye kredinin verilip verilmeyeceği başvuru puanı ile belirlenir, mevcut kredi müşterisinin kredisinin uzatılıp uzatılmayacağının kararı da davranışsal puanla belirlenir. Kredi başvurularındaki puanlama sisteminin mantığı farklı risk gruplarının birbirinden ayrıştırılmasını sağlamaya çalışmaktadır. Her iki yöntemin kullanılmasında da kredi verecek olan kuruluş kendi müşterilerinin bilgilerine dayanarak değerlendirme yapmaktadır. Ancak bu durumda her hangi bir müşteriye kredi verilmeyecek olursa onun nasıl davranacağı ve kredinin geri dönmesinin mümkün olup olmadığı sonucuna ulaşılamayacaktır. Bu da eldeki mevcut verilere göre incelendiğinde değersiz gözüken herhangi bir müşterinin kendini ispat etme şansını ortadan kaldırmaktadır.

---

[41] Hopper M. A., Lewis E.M., Behavior Scoring and Adaptive Control Systems, **In Credit Scoring and Credit Control**, ed. By L.C. Thomas, J.N.Crook, D.B.Edelman pp. 257-276. Oxford University Press, Oxford, 1992

[42] McNab H., Wynn A., Principles and Practice of consumer credit Risk Management, **CIB Publishing**, Canterbury, 2000

4.2 Kredi Puanının Belirlenmesinde Kullanılan Teknikler

Kredi puanlaması geçmiş veriyi ve istatistiksel teknikleri kullanarak başvuru yapan müşterinin kredi riskinin puanlanmasıdır. Puanlama oluşturulurken kredi verilmiş olan kişilerin verileri kullanılmaktadır. Her bir kredi sahibinin belli bir zaman aralığında 12, 18 veya 24 aylık dönemlerindeki ödemeleri incelenerek kredinin geri ödenip ödenmediği değerlendirilir. Kredi endüstrisinde kredi değerlendirilmesinin kötü olması kredi sahibinin arka arkaya üç ödeme yapmaması olarak tanımlanır[43].

Kredi puanlamasında bu güne kadar istatistik ve yöneylem araştırması yöntemleri kullanılmıştır. İstatistik yöntemler arasında diskriminant analizi, lojistik regresyon, sınıflama ağaçları (classification trees) bulunmaktadır. Yöneylem araştırması yöntemlerinden ise doğrusal programlama kullanılarak oluşturulan puan kartları (scoringcard) kullanılmıştır. Son dönemlerde ise bazı parametrik olmayan yöntemlerle beraber yapay zeka teknikleri de kredi puanlamasında kullanılmaktadır. Bunlar yapay sinir ağları, uzman sistemler, genetik algoritmalar ve en yakın komşuluk (nearset neighbour) teknikleridir.

Muhasebe kayıtlarını kullanarak oluşturulan modeller finansal kurumlar tarafından daha çok tercih edilmektedir. Belli oranların karşılaştırılması sonucunda şirketin piyasadaki durumu görülmektedir. Çok değişkenli istatistik teknikler ile hesaplanan kredi başvuru puanlama hesaplaması dört grupta değerlendirilebilir.

i) Diskriminant analizi
ii) Lojit modeli
iii) Probit modeli
iv) Doğrusal olasılık modeli

[43] Bridges S., Disney R. Modelling consumer credit and default:the research agenda. Experian Centre for Economic Modelling, [illegible], 2001

Bu teknikler arasında en yoğun kullanılan tekniklerden ilki diskriminant analizi ikincisi ise lojistik regresyondur[44].

Gruplar arasındaki farklılığın ortaya konulması çalışmalarının başlangıcı Fisher'e dayanmaktadır. Fisher'den sonra bir çok yöntem geliştirilmiştir. Bu yöntemleri başlıca üç ana başlık altında Toplamak mümkündür:

1) İstatistiksel yaklaşımlar
    a. Doğrusal Diskriminant analizi
    b. Lojistik Regresyon
    c. Sınıflama ve karar ağaçları
2) Yöneylem yaklaşımları
    a. Doğrusal Programlama
3) Parametrik olmayan yaklaşımlar
    a. Yapay sinir ağları
    b. Uzman sistemler
    c. Genetik algoritma
    d. En yakın komşuluk (nearest neighborhood)

Kredi puanlamasına bir karar mekanizması olarak bakmak gerekir. Başvuruyu yapan kişinin verdiği bilgiler girdi olarak, başvurunun olumlu veya olumsuz cevaplanması da çıktı olarak kabul edilir. Amaç yanlış sınıflama oranının minimum sayıda olmasını sağlamaktır. Kredi puanı hesaplanmasında istatistiksel modeller daha popüler olmuştur. Kredilerin ödenmesi veya ödenmemesi durumunda kredilerin çoklu normal dağılıma uygun olduğu ve kovaryans matrisinin eşit olduğu varsayımı altında model doğrusal karar verme kuralı olmaktadır.

---

[44] Altman E. I., Saunders A., "Credit risk measurement: development over the last 20 years", **Journal of Banking and Finance.** 21, 1998, 1721-1742

4.2.1 Diskriminant Analizi

Durand (1941)[45] diskriminant analizinin kredilerin ayrıştırılmasında kullanılabileceğini söyleyen ilk kişi olmuştur. Bu düşünceye Eisenbeis (1977)[46] yaptığı çalışma ile karşı çıkmış olsa da daha sonradan yapılan çalışmalar bu tekniğin kullanılabileceği göstermiştir. Eisenbeis (1977) çalışmasında dile getirdiği "diskriminant analizinin en önemli varsayımlarından olan değişkenlerin çoklu normal dağılıma sahip olması gerekmektedir" söylemi, aslında yanlış anlaşılan bir konudur. Eğer değişkenler çoklu normal dağılım varsayımına uygun ise doğrusal diskriminant kuralı optimum olmaktadır. Ancak eğer diskriminant analizi farklı grupları ayrıştırmak için kullanılan bir araç olarak değerlendirilirse çoklu normal dağılımla ilgili olan varsayımın gerçekleşmemesi durumunda da genel kabul görmüş olacaktır. Normallik varsayımı ise eğer parametreler hakkında testler yapılacak ise gerekli olmaktadır. Reichert ve diğerleri (1983)[47] yaptıkları ampirik çalışmalara dayanarak şu sonuçlara varmışlardır: "kredi bilgilerinin normal dağılması önemsendiği kadar önemli bir varsayım olmayabilir".

Diskriminant analizinin kredi puanlamasında kullanılmasında karşılaşılan problemleri Eisenbeis (1977) şu şekilde tanımlamaktadır.

a) Grupların tanımlanması
b) Ana kütle olasılıkları
c) Kovaryans matrisinin eşit olmaması
d) Diskriminant analizinin etkinliğinin ölçümü
e) Bağımsız değişkenlerin dağılımın normal olmaması
f) Bağımsız değişkenlerin test edilmesi
g) Boyut indirgenmesi

a) Eisenbeis'e göre en önemli problem grupların tanımlanması problemidir. İncelenilen grupların birbirinden farklı (ayrı) ve tanımlanabilir olması gerekmektedir.

---

[45] Durand a.g.e
[46] Eisenbeis R.A., "Pitfalls in the application of discriminant analysis in business, finance and economics" **Journal of Finance**, 32, 1977, 875-900
[47] Reichert A. K., Cho C.C., Wagner G.M., "An examination of the conceptual issues involved in developing credit scoring models", **Journal of Business and Ecomic Statistics**, 1, 1983, 101-114

Diskriminant analizinin bağımlı değişkenin kesikli olduğu durumlarda kullanılması uygundur. Eğer doğal bir gruplama olmaz ise belli bir değere göre gruplama yapmak gerekecektir. Bu da öznel olması gibi bir sorunu gündeme getirmektedir. Farklı kişiler farklı değerlerin belirlenmesine sebep olabilir.

b) Diskriminant analizi önsel anakütle olasılıklarının bilindiğini kabul etmektedir. Bir çok modelde önsel olasılıkların eşit olduğu kabul edilmekte veya seçilmiş olun örnek olasılıklarının anakütle önsel olasılığı olarak kullanılması söz konusudur. Anakütle önsel olasılıkları zaman içerisinde değişmesi de mümkündür dolayısıyla geleceğe ilişkin tahminlerin nasıl yapılacağı da bir sorun teşkil etmektedir.

c) Diskriminant analizinde başka bir varsayımın grupların kovaryans matrisinin eşit olması gerekmektedir. Eğer eşit değil ise doğrusal diskriminant analizi yerine kuadratik diskriminant analizi kullanılır.

d) Genellikle modelin geçerliliği, var olan gözlemlerin ikinci kısmı ile yapılmaktadır. Hatırlanacağı üzere veri seti ikiye bölünüp ilk kısmı ile model tahmin edilip ikinci kısmı ile modelin doğruluğu test edilebilmektedir. Şans oranında veri seti içerisinde rasgele seçim yapılıp gruplar tahmin edilmeye çalışılır. Diskriminant analizi ile elde edilen doğru sınıflama oranı ile rasgele sınıflama sonucu oluşan doğru sınıflama oranı karşılaştırılabilir.

e) Bağımsız değişkenlerin çoklu normal dağılıma sahip olması varsayımı genellikle sağlanamamaktadır. Özellikle kategorik değişken kullanıldığında hata oranının tahmininde yanlılığa sebep olacaktır. Bazı değişkenlerin çarpık olmasını logaritma alarak normale yaklaştırmak mümkün olmaktadır. Wagner ve diğerleri

(1983)[48] yaptıkları çalışmada değişkenlerin logaritmalarını alarak ve almadan sonuçları karşılaştırmışlar ve logaritması alınan modelin çok az bir üstünlüğü olduğunu göstermişlerdir. Bu çalışma, bu kabullenmenin bu tip modellemelerde çok da gerekli olmadığını göstermektedir.

f) Regresyon analizinde olduğu gibi bağımsız değişkenlerin katsayılarının test edilmesini sağlayan bir test diskriminant analizi için yoktur. Herhangi bir bağımsız değişkenin katsayısının sıfırdan veya her hangi bir sayıdan farklı olup olmadığını test etmek mümkün olmayacaktır.

g) Boyut indirgeme pratikte oldukça önemlidir. Çok sayıda bağımsız değişkenli modellerde boyut indirgenmesi, değişken sayısının azaltılması, oldukça önemlidir.

Finansal modellerin oluşturulmasında diskriminant analizinin bu gibi sorunlara sebep olacağını belirten Eisenbeis'e rağmen yapılan çalışmalar kredi puanlamasında bu sorunların önemli olmadığını göstermektedir.

Myers ve Frogy (1963)[49] diskriminant analizi ile regresyon analizi sonuçlarını karşılaştırmış. Lane (1972)[50], Apilado ve diğerleri (1974)[51], Moses ve Liao (1987)[52] diskriminant analizini kullanmışlardır. Grablowsky ve Talley (1981)[53] diskriminant analizi ve probit analizinin sonuçlarını karşılaştırmışlardır.

---

[48] Wagner G.M., Reichert A.K., Cho C.C, "Conceptual issues in credit credit scoring models", **Credit World**, 71, (May/June) 1983, 22-25

[49] Myers J.H., Frogy E.W., "The development of numerical credit evaluation systems", **Journal of American Statistical Association**, 58, 1963, 799-806

[50] Lane S., Submartingale credit risk classification, **Journal of Financial and Quantitative Analysis**, 7, 1972, 1379-1385

[51] Apilado V.P., Warner D.C., Dauten J.J., "Evaluation techniques in consumer finance", **Journal of Fianacial and Quantitative Analysis**, March, 1974, 275-283

[52] Moses D., Liao S.S., "On developing models for failure prediction", **Journal of Commercial Bank Lend**, 69, 1987, 27-38

[53] Grablowsky B.J., Talley W.K., "Probit and discriminant function for classifications for classifying credit applicants a comparison", **Journal of Economic and Business**, 33, 1981, 254-261

Titterrington (1992)[54] kuadratik puanlama fonksiyonun doğrusal diskriminant analizine göre daha kötü sonuçlar verdiğini göstermiştir.

Kolesar ve Showers (1985)[55] tarafından yapılan çalışma da doğrusal diskriminant analizi ile doğrusal programlama modeli karşılaştırılmıştır. Bu çalışmada 87,000 gözlem kullanılmıştır. Çalışmanın sonunda Kolesar ve Showers şu sonuçlara ulaşmışlardır. 1) Doğrusal diskriminant analizi varsayımlar gerçekleşmemiş olsa da kullanılabilir, 2) Basit ikili metotlar da doğru sonuçlar vermektedir. Bir başka bulgu ise yapılan çalışmanın örneklem büyüklüğü artıkça doğrusal programlama ve diskriminant analizinin sonuçlarının daha iyi sonuç verdiğidir.

Altman ve diğerleri (1994)[56] yaptıkları çalışmada 1982 ile 1992 yılları arasında 1000 firma kullanmışlardır. Doğrusal diskriminant analizinin yapay sinir ağlarına göre daha iyi sonuç verdiğini bulmuşlardır.

Muhasebe oranları ile şirketlerin iflas etmeleri arasında bir ilişki olup olmadığını ilk araştıran Altman (1968)[57] olmuştur. Altman ve diğerleri (1977)[58] ZETA diskriminant modelini geliştirmişlerdir. Kredi kullanan şirketlerin muhasebe oranlarını kullanarak geri ödeme yapanlar ile yapmayanları en iyi tahmin edecek doğrusal modeli kurmuşlardır. Altman tarafından ortaya atılan ilk modelde beş oran kullanılırken ZETA modelinde bu oranlar yediye çıkmıştır.

---

[54] Titterington D.M, "Discriminant analysis and related topics", In: Thomas L.C., Croock J.C and Edelman D.B. (eds), "Credit scoring and credit control", Oxford University Press, Oxford, 53-73,1992

[55] Kolesar P., Showers J.L., "A robust credit screening model using categorical data", **Management Science** 31(2), 1985, 123-133

[56] Altman E.I., Marco G., Varetto F., "Corporate distress diagnosis: comparison using linear disriminant analysis and nueral networks (the Italian experience)", **Journal of Banking and Finance**, 18, 1994, 505-529

[57] Altman E. I., "Financial ratios discriminant analysis and the prediction of corporate bankruptcy", **Journal of Finance**,1968, 589-609

[58] Altman E. I., Haldeman R., Narayanan P, "Zeta analysis: a new model to identify bankruptcy risk of corporations", **Journal of Banking and Finance**, 1977, 29-54

Lane (1972)[59] yaptığı çalışmada diskriminant analizini kullanmıştır. Bu çalışma 1964-1966 yılları arasındaki 274 batık kredi üzerine yapılmıştır. Çalışmada 17 bağımsız değişken kullanmıştır. Çalışma sonucunda şirketleri yüzde 93,2 doğru sınıflamıştır.

4.2.2 Lojistik Regresyon

Regresyon yöntemi de kredi puanlamasında kullanılan bir yöntem olmuştur. Regresyon analizinde diskriminant analizinde olduğu gibi değişkenlerin kukla değişken olarak kullanılması durumunda doğrusal bir kombinasyon elde edilir. Lachenbruch (1975)[60], Orgler(1970)[61], Orgler (1971)[62], Fitzpatrick (1976)[63], Lucas (1992)[64], Henley (1995)[65] bu yöntemin kullanıldığı çalışmalar yapmışlardır. Orgler (1971) yaptığı çalışmada yeni başvurular için, puanlama yerine verilmiş olan kredilerin durumunu değerlendirmektedir.

Teorik olarak bakıldığında doğrusal regresyon yöntemi yerine lojistik regresyon yöntemin daha uygun olduğu sonucuna varılmaktadır. Henley (1995) yaptığı çalışma lojistik regresyon sonuçlarının doğrusal regresyondan daha iyi sonuçlar vermediğini göstermiştir.

---

[59] Lane S., "Submarginal credit risk classification", **Journal of Financial and Quantitative Analysis**. January, 1972, 1379-1385

[60] Lachenbruch P.A., **Discriminant analysis**, New York: Hafner,1975

[61] Orgler Y.E., "A credit scoring models for commercial loans", **Journal of Money Credit Banking**, November, 1970, 31-37

[62] Orgler Y.E., "Evaluating of bank consumer loans with credit scoring models, **Journal of Bank Research**, 1, Spring, 1971, 31-37

[63] Fitzpatrick D.B., "An analysis of bank credit card profit", **Journal of Bank Research**., 7, 1976, 199-205

[64] Lucas A., "Updating scorecards: removing the mystique", **In credit scoring and credit control** (eds. L.C. Thomas, J.N. Crook, D.B. Edelman), pp 180-197, Oxford: Claredon, 1992

[65] Henley W.E., Statistical aspects of credit scoring. PhD Thesis. The Open University, Milton Keynes. 1995

Wiginton'un çalışması (1980)[66] kredi puanlaması alanında lojistik regresyonu kullanan ilk çalışmadır. Bu çalışmada 1967 ve 1969 yılları arasında bir petrol şirketin de yapılmış olan 1908 başvuru değerlendirilmiştir. Bu çalışmada bağımsız değişkenler iki gruba ayrılmıştır. Birinci grupta yer alan demografik değişkenler – ev sahibi olup olmadığı, evine bir önceki yıl taşınıp taşınmadığı, araba kullanım amacı-, ikinci grupta yer alan değişkenler ise ekonomik değişkenler olarak belirlenmiştir –meslek grubu, kaç yıldır bu işte çalıştığı, çalıştığı şirketin faaliyet alanı- yer almaktadır. Çalışmanın sonunda Wiginton lojistik regresyon yönteminin doğrusal diskriminant analizine göre üstün olduğunu bulmuştur. Bu sonucun dışında ulaşılan ikinci sonuç ise kredi puanın hesabında kullanılan yöntemlerde demografik değişkenlerin pek de yararlı olmadığını söylemektedir.

Srininivasan ve Kim (1987a)[67] lojistik regresyon ile diğer teknikleri karşılaştırmışlardır. Çalışmalarında şirket kredilerini incelemişlerdir. Leonard (1993a)[68] ticari kredilere lojistik regresyon ile diğer teknikler arasında karşılaştırma yapmıştır.

Boyes ve diğerleri (1989)[69] yaptıkları çalışmada beklenen getiriyi esas almışlar ve probit analizi kullanmışlardır. Çalışmada kullanılan veri seti 1977 ve 1980 yılları arasında başvuru yapmış olan 4632 kişiyi kapsamaktadır. 4632 kişinin 3711'ine kredi verilmişken 921'inin kredi başvurusu reddedilmiştir. Kredi veren firma kredi verdiği müşterilerinin %52,2'sini iyi ve %47,8'ini de kötü kredi olarak sınıflamıştır. Yapılan çalışmada bağımsız değişkenler üç gruba ayrılmıştır:

i) Kişisel özellikler (yaş, kaç aydır mevcut adresinde oturduğu, gibi)

---

[66] Wiginton J.C., "A note on the comparison of logit and discriminant models of consumer credit behavior", **Journal of Financial and Quantitative Analysis** XV,(3), 1980, 757-770

[67] Srinivasan V. Kim Y.H., "Credit granting: a comparative analysis of classification procedures", **Journal of Finance**, 42, 1987a, 665-683

[68] Leonard K.J., "Empirical bayes analysis of the commercial loan evaluation process", **Statistical Probability Letters**, 18, 1993a, 289-296

[69] Boyes W.J., Hoffman D.L., Low S.A., "An econometric analysis of the bank credit scoring problem", **Journal of Econometrics** 40, 1989, 3-14

ii) Ekonomik değişkenler (kira mı ev sahibi mi olduğu, çalıştığı meslek grubu)

iii) Finansal değişkenler (başka bir kredi kartının olup olmadığı, çek hesabının olup olmadığı, yatırım hesabının olup olmadığı).

Çalışmanın sonunda daha önceden kredi verilenlerin %94'üne, kötü olarak belirlenenlerin %61.4'üne ve reddedilen kredilerin %62.9'una kredi verilebileceği bulgusu elde edilmiştir.

4.2.3 Doğrusal Programlama

Amaç fonksiyonu başvuruları doğru sınıflama oranının optimizasyonu olarak belirlendiğinde sorunun doğrusal programlama ile çözmesi mümkün olmaktadır.

Mangasarian (1965)[70] doğrusal programlamanın kredi puanının hesabı için kullanılabileceğini gören ilk kişi olmuştur. Ancak Freed ve Glover (1981a)[71] , (1981b)[72] tarafından yapılan çalışmalar bu konuyla ilgili ilk eserler olmuştur.

Showers ve Chakrin (1981)[73] , Kolesar ve Showers (1985)[74] tam sayılı doğrusal programlama yöntemini kullanarak telefon müşterilerinin telefona yatırılan depozitin bırakılmasını tahmin etmeye çalışmışlardır.

---

[70] Mangasarian O.L., "Linear and nonlinear separation separation of patterns by linear programing" **Operations Research** 13, 1965, 444-452

[71] Freed N., Glover F., "A linear programing approach to the discriminant problem", **Decision Science**, 12, 1981a, 68-74

[72] Freed N., Glover F., "Simple but powerful goal programming formulations for the discriminant problem", **European Journal of Operational Research**, 7, 1981b, 44-60

[73] Showers J.L., Chakrin L.M., "Reducing uncollectable revenue from residental telephone customers, **Interfaces**, 11, 1981, 21-31

[74] Kolesar P., Showers J.L., "A robust credit screening model using categorical data", **Management Science**, 31, 1985, 123-133

Doğrusal programlama yönteminin kredi puanlamasında kullanılmasına katkı ise Joachimsthaler ve diğerleri (1990)[75] tarafından yapılmıştır. Nath ve diğerleri (1992)[76] ile Hardy ve Adrian (1985)[77] doğrusal programlama ve istatistik metotları arasında yaptığı karşılaştırmada istatistik metotlarının daha iyi sonuçlar verdiğini bulmuş olsalar da daha sonra bir çok çalışma yapılmıştır. Glen (1997)[78] tamsayılı doğrusal programlama bunlardan biridir.

Ziari ve diğerleri (1997)[79] yaptıkları çalışmada doğrusal programlama tekniğini kullanarak ayrıştırma işlemini yapmışlardır. Bu teknikle beraber lojistik regresyon analizini kullanmışlardır. Bu çalışmada yeniden örnekleme tekniğini kullanarak iki teknik arasında karşılaştırma yapmışlardır. Çalışmalarında kullandıkları veri setini yakın zamandaki başvurular ve eski başvurular olarak iki gruba ayırmışlardır. Toplam gözlem sayısı 1999 dur. Bütün gözlemlerin yarısını -1000 adet- modelin kurulması için , kalan kısmını ise modelin doğruluğunu test etmek için kullanılmışlardır. Başvuru yapan firmaların finansal oranları ise bağımsız değişkenler olarak kullanılmıştır. Bağımsız değişkenler; Likidite oranı, Aktif karlılık oranı, Aktif borç oranı, katkı payı, geri ödeme oranı ve yeniden finansman ihtiyacı var ise bir yok ise sıfır olarak tanımlanmıştır. Yapılan çalışmanın sonucunda bulunan sonuçların karşılaştırılmasında lojit modeli ile doğrusal programlama arasında ciddi bir farklılık görülememiştir.

---

[75] Joachimsthaler E. W., Stam A., "Mathematical programming approaches for the classification problem in two-group discriminant analysis", **Multivariate Behavioural Research** 25, 1990, 427-454

[77] Hardy W.E., Adrian J. L., "A linear programming alternative to discriminant analysis in credit scoring" **Abribus** 1, 1985, 285-292

[78] Glen J.J., "Integer programming models for normalisation and variable selection in mathematical programming models for discriminant analysis", **Procedings of Credit Scoring and Credit Control V**, Credit Research Center, University of Edinburgh, 1997

[79] Ziari H.A., Leatham D.J., Ellinger P.N., "Development of statistical discriminant mathematical programimg model via resampling estimation techniques", **American Journal of Agricultural economics**, 79, 1997, 1352-1362

Bugera ve diğerleri (2002)[80] yaptıkları çalışmada Yunanistan Ulusal Bankası'na 1995-1996 yılında yapılmış olan 150 başvuruyu değerlendirmişlerdir. Doğrusal programlama yöntemini kullanmışlardır. Bağımsız değişken olarak medeni durum (boşanmış,bekar, evli-dul), meslek, iş telefonu, ikamet durumu (kira, yurt, kendi evi), bankadaki hesabının olup olmadığı, yaş, çalışma süresi olarak belirlemiş ve bütün değişkenleri kategorik olarak modele koymuşlardır. Sonuç olarak fayda fonksiyonun doğrusal veya kuadratik olmasının farklılık yarattığını bulmuşlardır. Bununla beraber veri setinin küçük olması dolayısıyla bu sonuçların kendi veri setleri için böyle bir sonuç verdiğini de söylemektedirler.

4.2.4 Karar Ağaçları

Karar ağaçları, istatistik, yapay zeka gibi bir çok disiplinde kullanılmaktadır. Karar ağaçları ile ilgili en önemli kaynaklar Breiman ve diğerleri (1984)[81] ve Safavian ve Landgrebe (1981)[82] olmuştur.

Sınıflama ağacı ve uzman sistemler puan kartlarından farklı müşterileri gruplamak yerine, her bir grubun riskini hesaplamaya dayanmaktadır. Bu gruplar kendi içinde homojen iken gruplar birbirleri arasında heterojendir. Sınıflama ağacı Breiman ve diğerleri (1984)[83], uzman sistemler ise Safavain ve Landgrebe(1991)[84] tarafından kullanılmaya başlanmıştır. Makowski (1985)[85] sınıflama ağacını kredi puanlamasında kullanan ilk kişi olmuştur. Coffman (1986)[86] sınıflama ağacı ile diskriminant arasında karşılaştırma yapan ilk kişi olmuş ve karar ağaçlarının değişkenler arasında etkileşim

---

[80] Bugera V., Konno H., Uryasev S., "Credit cards scoring with quadratic utulity function", **Research Report**, University of Florida 2002-1

[81] Breiman L., Friedman J.H:, Olshen R.A., Stone C.J., **Classification and regresiion trees,** Belmont:Wadsworth, 1984

[82] Safavian S.R., Landgrebe D., "A survey of decision tree classifier methodology", **IEEE Tranportation System Man Cyb**. 21, 1991, 660-674

[83] Breiman L., Friedman J.H., Olshen R. A., Stone C.J., **Classification and regression trees,** Wadsworth, Belmont, California, 1984

[84] Safavain S. F., Landgrebe D., "A survey of decision tree classifier methodology", **IEEE Trans. On Systems, Man and Cybernetics**, 21, 1991, 660-674

[85] Makowski P. "Credit scoring branches out", **The Credit World**, 75, 1985, 30-37

[86] Coffman J. Y., "The proper role of tree analysis in forecasting the risk behaviour of borrowers", MDS Reports, **Manangement Decision Systems**, Atlanta, 3,4,7 and 9, 1986

varsa, diskriminant analizinin ise karşılıklı ilişki varsa daha iyi sonuç verdiğini bulmuştur. Mehta (1968)[87], Carter ve Catlett (1987)[88], Boyle ve diğerleri (1992)[89] kredi puanlamasında sınıflama ağacının sonuçlarını tartışmaktadırlar.

Bierman ve Hausman (1970)[90] kredi puanının hesabında diskriminant analizi yerine Bayesian yöntemini kullanarak karar ağaçlarından yararlanılmasını önermiştir. Modelinde çok aşamalı dinamik programlama kullanarak kredi verilmesi kararının nasıl alınması gerektiğini açıklamıştır. Kredi yönetimini kredi verilmesi sonucu toplanacak olan beklenen getiri ile toplanamayacak olan ödemeler sonucu doğacak zarar arasındaki denge olarak kabul etmektedir.

4.2.5 Uzman Sistemler

Teknolojideki değişikler kredi riskinin tahmini ile ilgili tahmin metotlarının ilerlemesine sebep olmuştur. Zocco (1985)[91], Davis (1987)[92] ve Leonard (1993b)[93], (1993c)[94] bu yöntemle yapılmış çalışmalar arasındadır. Bu sistemin iyi tarafı kredi başvurusu reddedilen kişilere açıklama yapılabilmesidir.

---

[87] Mehta D., "The formulation of credit policy models", **Management Science** 15, 1968, 30-50

[88] Carter C., Catlett J. "Assesing credit card applications using machine learning", **IEEE Expert** 2, 1987, 71-79

[89] Boyle M., Crook J. N., Hamilton R., Thomas L.C., "Methods for credit scoring applied to slow payers in credit scoring and credit control", ed. L. C. Thomas, D.B. Edelman, Oxford University Press, Oxford, 75-90, 1992

[90] Bierman H., Hausman W.H., "The credit granting decision", **Management Science**,16,8, 1970, 519-532

[91] Zocco D.P., "A framework for expert systems in bank loan management", **Journal of Commercial Bnak Lending**, 67, 1985, 47-54

[92] Davis D.B., "Artificial intelligence goes to work", **High Technology April**, 1987, 16-17

[93] Leonard K.J., "Detecting credit fraud using expert systems", **Computer Industrial Engineer**, 25, 1993b, 103-106

[94] Leonard K.J., "A fraud alert model for credit cards during the authorization process", **IMA Journal of Mathematical Applied Business Industry**, 5, 1993c, 57-62

4.2.6 Sinir Ağları

Yapay sinir ağları son on yıldır kredi puanlamasında kullanılan başka bir tekniktir. Cheng ve Titterington (1994)[95] kredi puanlamasında yapay sinir ağlarını kullanmışlardır. Yapay sinir ağları ile yapılan çalışmaların bir çoğu şirketlerle ilgili puan hesabında kullanılmaktadır. Altman ve diğerleri (1994)[96], Tam ve Kiang (1982)[97], Desai ve diğerleri (1996)[98], Desai ve diğerleri (1997)[99] kredi kartlarındaki uygulamaları dışında tüketici kredilerinde yapay sinir ağlarının performansı ile diskriminant, genetik algoritma, tekniklerini karşılaştırmışlardır.

4.2.7 Parametrik Olmayan Yöntemler

Kredi puanının hesaplanmasında ilk defa parametrik olmayan bir yöntem kullanan Chatterjee ve Barcun (1970)[100] bütün bağımsız değişkenleri kukla değişkene dönüştürmüştür. Kullandığı değişkenler ise gelir, iyi bir bölgede yaşamak, borcunun 300$'dan az olması, telefonunun olması, evinin olması, 3 yıldan fazla süredir şu anki işinde çalışıyor olması, memur olması, bekar olmasıdır. Kurduğu modelde maliyeti de düşünmüş ve beş farklı maliyet oranına göre değişik bağımsız değişkenleri kullanarak modeller kurmuş ve yanlış sınıflama oranlarını bulmuştur.

---

[95] Cheng B., Titterington D.M., "Neural networks: a review from a statistical perspective", **Statistical Science** 9, 1994, 2-30

[96] Altman E.I., Marco G., Varetto F., "Corporate distress diagnosis: comparison using linear discriminant analysis and nueral networks (the Italian experience)", **Journal of Banking and Finance**, 18, 1994, 505-529

[97] Tam K.Y., Kiang M.Y., "Managerial applications of neural networks: the case of bank failure prediction", **Management Science**, 38, 1992, 926-947

[98] Desai V.S., Crock J. N., Overstreet G.A., "A comparison of neural networks and linear scoring models in the credit environment", **European journal of Operational Research**, 95, 1996, 24-37

[99] Desai V.S., Convay D. G., Crock J. N., Overstreet G.A., "Credit scoring models in credit union environment using neural networks and genetic algorithms", **IMA Journal of Mathematics Applied in Business and Industry**, 8, 1997, 323-346

[100] Chatterjee S. Barcun S., "A nonparametric approach to credit screening", **Journal of American Statistical Association**, March, 65, 1970, 150-154

Parametrik olmayan yöntemlerden en yakın komşuluk (nearest neighbour) yöntemi kredi puanlaması için geliştirilmiştir. Hand (1986)[101], Henley ve Hand (1996)[102] bu yöntemi kullanmıştır. İlk çalışma ise en yakın komşuluk ile karar ağaçlarının karşılaştırmasını yapmıştır. İkinci çalışma ise en yakın komşuluk yönteminin derinlemesine araştırmıştır.

Genetik algoritma kredi puanlamasında kullanılan yöntemlerden biridir. Sınıflama işlemi yapılırken değişik puan kartları karıştırılarak karar verilmeye çalışılır. Fogerty ve Ireson (1993)[103] ile Albright (1994)[104] bu yöntemi ilk kullananlar arasındadır. Yobas ve diğerleri (1997)[105] genetik algoritma, yapay sinir ağları ve sınıflama ağacı arasında karşılaştırma yapmışlardır.

Salas ve Saurina (2002)[106] İspanyadaki iki grup banka arasındaki kredi rejimini incelemiş ve makro ekonomik değişkenlerin de modelde yer alması gerektiğine karar vermişlerdir. Bağımlı değişken olarak problemli kredilerin oranını almışlar, bağımsız değişkenler olarak Gayri Safi Milli Hasıla büyüme oranını, şirketin borçlarının piyasa değerine oranını, 1988 yılında kredi rejiminde değişiklik olduğu için bu yılı kukla değişken olarak ve her şirket için kredilerdeki büyümeyi almışlardır. Bu değişkenlerden bazılarını gecikmeli olarak modele koymuşlardır. Bu çalışma sonunda Salas ve Saurina tasarruf bankalarındaki kredi riskinin ticari bankalardaki kredi riskine göre büyük oranda mikro ekonomik değişkenler tarafından açıklanabildiğini söylemektedirler.

---

[101] Hand D.J., "New instruments for identifying good and bad credit risk: a feasibly study report", **Trustee Saving Bank**, London, 1986

[102] Henley W.E., Hand D.J., "A k-nearest neighbour classifier for assessing consumer credit risk" **Statistician**, 45, 1996, 77-95

[103] Fogarty T.C., Ireson N.S., "Evolving bayesian classifiers for credit control-a comparison with other machine learning methods", **IMA Journal of Mathematics Applied in Business and Industry**, 5, 1993, 63-76

[104] Albright H.T., "Construction of a polynomial classifier for consumer loan applications using genetic algorithms", Working Papers, Department of Systems Engineering, University of Virginia, 1994

[105] Yobas M.B., Crook J.N., Ross P., "Credit scoring using neural and evolutionary techniques", Working Paper 97/2, Credit research Centre, University of Edinburg, 1997

[106] Salas V., Saurina J., "Credit risk in two institutional regimes:Spanish commercial and savig banks", **Journal of Financial Service Reserach**, 22:3, 2002, 203-224

Ticari bankalardaki problemli kredilerin tasarruf bankalarındaki problemli kredilere göre iktisadi dalgalanmalardan daha fazla etkilendikleri sonucuna ulaşmışlardır.

Srinivasan ve Kim (1987)[107] yaptıkları çalışmada diskriminant, lojistik, doğrusal programlama ve karar ağaçları modellerini kullanarak bulunan sonuçları karşılaştırmışlardır. Yapılan çalışmada 215 müşteri değerlendirilmiştir. Bağımsız değişkenler olarak cari oran, likidite oranı, net varlığın borçlara oranı, Toplam varlığın logaritması, net kar marjı, aktif karlılık oranı kullanılmıştır. Çalışma sonuçları aşağıdaki tabloda verilmiştir.

Tablo 3. Doğru Sınıflama Oranları (%)

	Bootstrap	Eldeki veri
Doğrusal Diskriminant Analizi	88,89	85,05
Quadratik Diskriminant Analizi	90,74	86,92
Lojit	92,59	87,85
Doğrusal Programlama 1	87,96	84,11
Doğrusal Programlama 2	88,39	86,92
Karar Ağaçları 1	94,44	92,56
Karar Ağaçları 2	93,52	92,29

Tablodan görülebileceği gibi lojistik modelleme, diskriminant (DDA), kuadratik diskriminant (QDA) analizinden ve doğrusal programlama (DP) modellerinden daha iyi sonuç vermesine rağmen karar ağaçları (KA) bütün modellere üstünlük sağlamıştır.

[107] Srinivasan V., Kim Y.H., "Credit granting :a comparative analysis of classification procedures", **Journal of Finance**, 42, 3, 1987, 665-681

V. UYGULAMA

5.1 Giriş

Çalışmanın bu bölümünde Lojistik Regresyon ve Diskriminant Analizlerine Bootstrap tekniği uygulanacak ve bulunan sonuçlar tartışılacaktır. Uygulama da MATLAB programı kullanılmıştır. Uygulama iki farklı veri seti üzerinde yapılmıştır. Birinci veri seti Thomas ve diğerleri (2002)[108] tarafından yazılan kitaptan alınan yurtdışına ait olan veri setidir. İkinci veri seti ise yurtiçindeki bir finans kurumundan sağlanan veri setidir. Her iki veri seti de ikiye ayrılmış birinci kısmı ile fonksiyon katsayıları elde edilmiş, ikinci kısmı ile elde edilen katsayıları kullanılarak doğru sınıflama oranları karşılaştırılmıştır. Her iki veri setinin de ilk kısmına Bootstrap uygulanarak Lojistik Regresyon ve Diskriminant Analizi sonuçları elde edilmiş ve doğru sınıflama oranları ve katsayıların dağılımları incelenmiştir.

Bu çalışmanın amacı kredi derecelendirmesinde kullanılan istatistik tekniklerin sonuçlarının bootstrap tekniği kullanılarak doğru sınıflama oranında etkili olup olmadığını tespit etmektir. İki farklı veri setinin kullanılma amacı ise elde edilen sonuçların bir birleriyle tutarlı olup olmadığını göstermektir.

5.2 Thomas'ın Veri Seti

Thomas ve diğerleri (2002) tarafından sağlanan kredi verisinin kullanıldığı birinci uygulamada. Veri seti 1225 gözlemden oluşmaktadır. Veri setinin %80'i diskriminant ve lojistik regresyon katsayılarının tahmini için %20'si ise bulunan katsayıların test edilmesi için kullanılacaktır. Veri setinde yer alan değişkenler arasından aşağıda belirtilenler seçilmiştir.

- Başvuran kişinin yaşı.

[108] Thomas, L. C, Edelman, D. B., Crook J. N., **Credit scoring and its applications**, Society for Industrial and Applied Mathematics, Philadelphia, 2002

- Çocuk sayısı.
- Evinde telefon olup olmadığı.
- Başvuran kişinin mesleği.
- Başvuran kişinin aylık geliri.
- Başvuran kişinin ikamet durumu.

Yas değişkeni başvuran kişinin yaşını göstermektedir.

Çocuk sayısı değişkeni ise başvuran kişinin çocuk sayısını göstermektedir.

Tel: Başvuruyu yapan kişinin evinde telefon olup olmadığını gösterir (sıfır evde telefon olmadığını, bir ise telefonun olduğunu göstermektedir).

İş1: Başvuruyu yapan kişi devlette çalışıyor ise bir diğer durumlarda sıfır değerini alıyor.

İş2: Başvuruyu yapan kişi ev kadını ise bir diğer durumlarda sıfır değerini alıyor.

İş3: Başvuruyu yapan kişi asker ise bir diğer durumlarda sıfır değerini alıyor.

İş4: Başvuruyu yapan kişi özel sektörde çalışıyor ise bir diğer durumlarda sıfır değerini alıyor.

İş5: Başvuruyu yapan kişi kamu kesiminde çalışıyor ise bir diğer durumlarda sıfır değerini alıyor.

İş6: Başvuruyu yapan kişi emekli ise bir diğer durumlarda sıfır değerini alıyor.

İş7: Başvuruyu yapan kişi esnaf ise bir diğer durumlarda sıfır değerini alıyor.

İş8: Başvuruyu yapan kişi öğrenci ise bir diğer durumlarda sıfır değerini alıyor.

İş9: Başvuruyu yapan kişi işsiz ise bir diğer durumlarda sıfır değerini alıyor.

İş10: Başvuruyu yapan kişi yukarıdaki seçenekler dışında bir işe sahip ise bir diğer durumlarda sıfır değerini alıyor.

Gelir: Başvuruyu yapan kişinin aylık gelirini göstermektedir.

Ev1: Başvuruyu yapan kişinin oturduğu ev kendinin ise bir diğer durumlarda sıfır değerini alıyor.

Ev2: Başvuruyu yapan kişinin oturduğu ev eşyalı olarak kiralandı ise bir diğer durumlarda sıfır değerini alıyor.

Ev3: Başvuruyu yapan kişinin oturduğu ev eşyasız olarak kiralandı ise bir diğer durumlarda sıfır değerini alıyor.

Ev4: Başvuruyu yapan kişinin ailesi ile oturuyor ise bir diğer durumlarda sıfır değerini alıyor.

Aşağıdaki tabloda her bir kukla değişken için krediyi ödeyen ve ödemeyen kişilerin sayılarını gösteren frekans tablosu verilmiştir.

Tablo 4. Bağımsız Değişkenlerin Frekans Tablosu.

Bağımsız Değişken	Krediyi Ödeyen Kişi Sayısı		Krediyi Ödemeyen Kişi Sayısı	
Telefon				
var	689	76%	219	24%
yok	61	66%	31	34%
Devlette Çalışanlar	153	82%	34	18%
Ev Kadını	19	63%	11	37%
Asker	14	74%	5	26%
Özel Sektör	351	80%	88	20%
Kamu Sektörü	18	75%	6	25%
Emekli	48	55%	40	45%
Esnaf	73	72%	28	28%
Öğrenci	64	68%	30	32%
İşsiz	3	50%	3	50%
Bunların dışında	5	83%	1	17%
Ev Sahibi	390	76%	122	24%
Eşyalı Evde Kiracı	75	75%	25	25%
Eşyasız Evde Kiracı	94	75%	31	25%
Ailesi ile Kalanlar	163	77%	49	23%

Telefonu olan başvuru sahiplerinden 689'u (%76) kredisini ödemişken 219'u (%24) kredisini ödememiştir.

Sürekli değişkenlerden yaş, gelir ve çocuk sayısına bakıldığında ise, kredisini geri ödeyenlerin yaş ortalaması 51.16, kredisini geri ödemeyenlerin yaş ortalaması 46.84 olarak hesaplanmıştır. Aldıkları kredileri geri ödeyenlerin gelir ortalamalarına bakıldığında kredilerini geri ödemeyenlerden daha yüksek olduğu görülmektedir.

Tablo 5. Sürekli Bağımsız Değişkenlerin Tanımlayıcı İstatistikleri.

Sürekli Değişkenlerin Ortalamaları		
YAS	Krediyi Ödeyenler	51.16
	Krediyi Ödemeyenler	46.84
ÇOCUK SAYISI	Krediyi Ödeyenler	0.62
	Krediyi Ödemeyenler	0.54
GELİR	Krediyi Ödeyenler	23198.2 $
	Krediyi Ödemeyenler	16282.7 $

Çalışmanın amacı veri setinden elde edilecek doğru sınıflama oranı ile yeniden örnekleme tekniğini kullanarak elde edilecek olan doğru sınıflama oranlarının karşılaştırılmasıdır. Bu amaçla veri seti öncelikle %80'lik ve %20'lik iki bölüme ayrılacak. Veri setinin %80'lik kısmına Diskriminant Analizi ve Lojistik Regresyon uygulanarak fonksiyon katsayıları bulunacaktır. Veri setinin ayrılan %20'lik kısmında ise bulunan katsayılar kullanılarak doğru sınıflama oranları bulunacaktır. İkinci aşamada yeniden örnekleme tekniği kullanılarak 250, 500 ve 1000 adet veri seti türetilecek Diskriminant Analizi ve Lojistik Regresyon tekniklerindeki katsayıları elde edilecek ve bu katsayıların ortalamaları hesaplanacak. Bu ortalama katsayılar kullanılarak %20'lik veri setinde doğru sınıflama oranı bulunacaktır.

Bağımsız değişkenler; telefonun olması, meslek durumu ve ikamet durumu değişkenleri kategorik olduğu için modele kukla değişken olarak girecektir. Çoklu bağlantı sorunuyla karşılaşmamak için her kategorik değişken durumunun bir eksiği kadar kukla değişken modele girecektir. Modele giremeyen kukla değişkenler sabit terim tarafından modelde temsil edilecektir. Kategorik değişkenlerden evde telefonun

olmaması durumu, meslek değişkeninde cevap vermemesi durumu ve ikamet değişkeninde ise diğer durumlar modelde sabit terim tarafından temsil edilecektir.

5.2.1 Diskriminant Analizi Uygulaması ve Sonuçları

Veri setine uygulanan Diskriminant analizi sonucunda bulunan Diskriminant fonksiyonun katsayıları tablo 6'da gösterilmektedir. Tablo 6'dan görülecegi üzere kredinin geri ödenme riski çocuk sayısı ile doğru orantılıdır telefonun olması durumunda ise ters orantılıdır.

Tablo 6. Diskriminant Analizi Sonucunda Bulunan Katsayılar.

	Katsayılar		Katsayılar
Yaş	0.020		
Çocuk Sayısı	-0.071	Öğrenci Olanlar	1.253
Telefonu Olanlar	0.546	İşsiz Olanlar	0.026
Devlette Çalışanlar	1.899	Mesleği Farklı Olanlar	2.320
Ev Kadını	1.314	Gelir	0.000038
Asker	1.305	Ev sahibi Olanlar	1.220
Özel Sektörde Çalışanlar	1.781	Eşyalı Evde Kiracılar	1.359
Kamu Kesiminde Çalışanlar	1.126	Eşyasız Evde Kiracılar	1.512
Emekli	0.711	Ailesi İle Kalanlar	1.335
Esnaf	1.276	Sabit	-5.060

Tablo 6'da bulunan katsayılar veri setinin %80'lik kısmında kullanılırsa Tablo 7'deki sonuçlar bulunacaktır. Doğru sınıflama oranı (138+519)/(250+750)=0.657 olarak hesaplanmaktadır. Katsayıların hesabında %80'lik veri seti kullanıldığından bu oran yanlıdır.

Tablo 7. Diskriminant Fonksiyonu Doğru Sınıflama Oranı (Veri Setinin %80) .

			Tahmin Edilen Grup		Toplam
			Krediyi Ödemeyenler	Krediyi Ödeyenler	
Gerçek Veri	Adet	Krediyi Ödemeyenler	138	112	250
		Krediyi Ödeyenler	231	519	750
	%	Krediyi Ödemeyenler	55.2	44.8	100
		Krediyi Ödeyenler	30.8	69.2	100

Modelin ne kadar güvenilir olduğunu bulmak için test amacıyla ayrılmış olan %20'lik, 225 adet, kısmı kullanılacaktır. Bu veri seti için bulunan diskriminant ağırlıklarının kullanılarak sınıflama yapılırsa (40+54)/225=%42'lik doğru sınıflama yapılacaktır.

Tablo 8. Diskriminant Fonksiyonu Doğru Sınıflama Oranı (Veri Setinin %20).

			Tahmin Edilen Grup		Toplam
			Krediyi Ödemeyenler	Krediyi Ödeyenler	
Gerçek Veri	Adet	Krediyi Ödemeyenler	40	33	73
		Krediyi Ödeyenler	98	54	152
	%	Krediyi Ödemeyenler	0.55	0.45	100
		Krediyi Ödeyenler	0.64	0.36	100

5.2.2 Lojistik Regresyon Uygulaması ve Sonuçları

Lojistik regresyon sonucunda bulunan katsayılar Tablo 9'da verilmiştir. Tablo 9'dan görüleceği üzere çocuk sayısı burada da kredinin ödenmeme riskini artırırken telefonun olması kredinin ödenmeme riskini azaltmaktadır.

Tablo 9. Lojistik Regresyon Katsayı Sonuçları.

	Katsayılar		Katsayılar
Yaş	0.01		
Çocuk Sayısı	-0.04	Öğrenci Olanlar	0.46
Telefonu Olanlar	0.31	İşsiz Olanlar	-0.05
Devlette Çalışanlar	0.83	Mesleği Farklı Olanlar	1.11
Ev Kadını	0.53	Gelir	0.000024
Asker	0.45	Ev sahibi Olanlar	0.61
Özel Sektörde Çalışanlar	0.74	Eşyalı Evde Kiracılar	0.69
Kamu Kesiminde Çalışanlar	0.32	Eşyasız Evde Kiracılar	0.78
Emekli	0.26	Ailesi İle Kalanlar	0.68
Esnaf	0.44	Sabit	-1.44

Tablo 9'da bulunan katsayılar kullanırsa doğru sınıflama oranı gösteren tablo 10 elde edilecektir. Lojistik regresyon modelinin doğru sınıflama oranı %80'lik veri seti üzerinde uygulandığında %75.5 olarak gerçekleşmiştir. Bu doğru sınıflama oranı %80'lik veri setinde bulunduğundan yüksek çıkmış olması doğaldır.

Tablo 10. Lojistik Fonksiyonu Doğru Sınıflama Oranı (Veri Setinin %80).

		Tahmin		
				Doğru Tahmin
		Krediyi Ödeyenler	Krediyi Ödemeyenler	
Gerçek	Krediyi Ödeyenler	731	19	97.5
	Krediyi Ödemeyenler	226	24	9.6
Genel Doğruluk Oranı				75.5

Kurulan modelin testi için ayrılan %20'lik veri seti kullanıldığında tablo 11 bulunmuştur. Doğru sınıflama oranına %20'lik veri seti için %68 olmaktadır.

Tablo 11. Lojistik Fonksiyonu Doğru Sınıflama Oranı (Veri Setinin %20).

		Tahmin		
				Doğru Tahmin
		Krediyi Ödeyenler	Krediyi Ödemeyenler	
Gerçek	Krediyi Ödeyenler	146	6	0.97
	Krediyi Ödemeyenler	65	8	0.11
Genel Doğruluk Oranı				0.68

Lojistik ve Diskriminant analizinin sonuçlarının karşılaştırıldığında diskriminant modeli gerçekte kredisini ödeyenlerin 519'unu doğru olarak tahmin ederken bu sayı lojistik regresyonda 731 e çıkmıştır. Fakat gerçekte kredisini ödemeyenler değerlendirildiğinde diskriminant fonksiyonu 138 kişiyi doğru tahmin ederken lojistik regresyon gerçekte kredisini ödemeyecek olan kişilerin sadece 24'ünü doğru olarak tahmin etmiştir.

5.2.3 Bootstrap Uygulaması Sonuçları

Diskriminant Analizi ve Lojistik Regresyon teknikleri eldeki veriye yeniden örnekleme tekniği kullanılarak yaratılacak olan veri setleri üzerinde uygulanacaktır. Üç tip yeniden örnekleme yapılacaktır: 250 adet, 500 adet ve 1000 adet veri seti yaratılarak sonuçlar incelenecektir. Üç tipin kullanılma sebebi yeniden seçilen örneklem sayısını artırınca doğru sınıflama oranında sonucun iyiye gidip gitmediğini görmektir. Her bir yeniden örnek tipi için ortalama değer hesaplanarak veri setinin ayrılmış olan %20'lik kısmında doğru sınıflama oranına bakılacaktır. Tablo 12'de yeniden örnekleme sonucu bulunan diskriminant analizi ve lojistik regresyon tahmin değerlerinin ortalama değerleri verilmiştir.

Bu tablodaki tahmin değerleri kullanılarak ana kütleden ayrılmış olan 225 veri için sınıflama işlemi yapılacaktır ve doğru sınıflama oranları diskriminant analizi ve lojistik regresyon sonuçları ile karşılaştırılacaktır.

Tablo 12. Bootstrap Ortalama Değerleri

	Diskriminant Analizi Sonuçları			**Lojistik Regresyon Sonuçları**		
	N=250	N=500	N=1,000	N=250	N=500	N=1,000
	Ortalama	Ortalama	Ortalama	Ortalama	Ortalama	Ortalama
Sabit	-5.129	-5.149	-5.146	-2.918	-2.588	-2.417
Yaş	0.021	0.021	0.02	0.013	0.012	0.012
Çocuk Sayısı	-0.072	-0.079	-0.076	-0.030	-0.037	-0.036
Telefonu Olanlar	0.62	0.588	0.589	0.332	0.316	0.299
Devlette Çalışanlar	1.978	2.1	2.061	2.177	1.990	1.787
Ev Kadını	1.411	1.494	1.419	1.928	1.665	1.490
Asker	1.338	1.369	1.427	1.785	1.632	1.545
Özel Sektörde Çalışanlar	1.849	1.971	1.947	2.089	1.877	1.688
Kamu Kesiminde Çalışanlar	1.117	1.235	1.244	1.735	1.516	1.312
Emekli	0.734	0.794	0.794	1.653	1.405	1.237
Esnaf	1.277	1.384	1.406	1.790	1.560	1.383
Öğrenci Olanlar	1.255	1.39	1.383	1.807	1.584	1.402
İşsiz Olanlar	0.112	0.185	0.199	1.222	1.146	0.988
Mesleği Farklı Olanlar	2.56	2.566	2.551	5.509	5.432	5.270
Gelir	0.00004	0.00004	0.00004	0.000026	0.000025	0.000025
Ev sahibi Olanlar	1.268	1.277	1.271	0.621	0.597	0.611
Eşyalı Evde Kiracılar	1.405	1.41	1.437	0.725	0.692	0.697
Eşyasız Evde Kiracılar	1.546	1.596	1.577	0.817	0.757	0.774
Ailesi İle Kalanlar	1.401	1.367	1.368	0.691	0.657	0.666

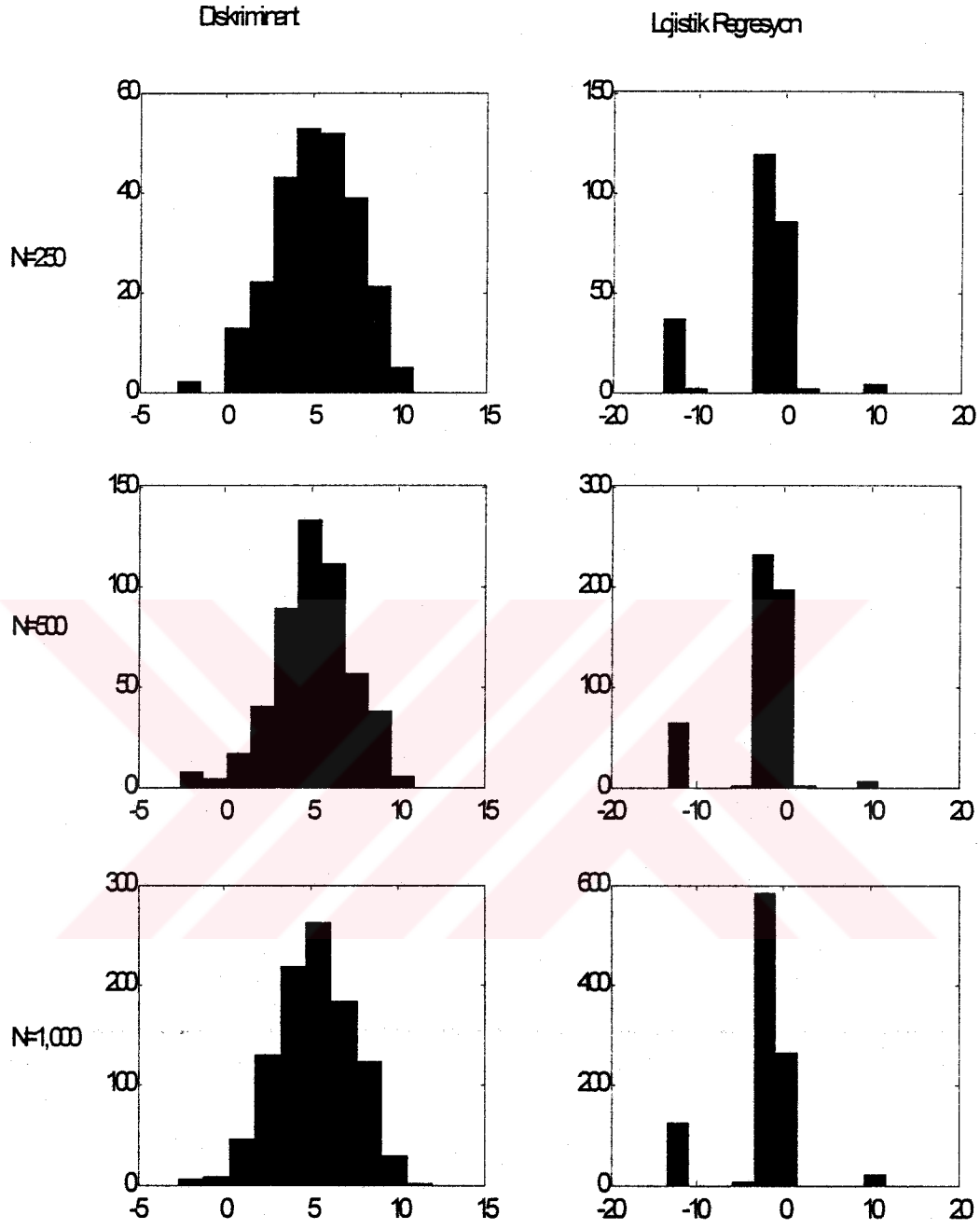
Tablo 12'de üç farklı bootstrap örneklemi sonrasında hesaplanan Diskriminant ve Lojistik Regresyon katsayılarının ortalamaları görülmektedir. Telefonu olanların katsayısı diskriminant analizinde 250 adet örnek için ortalama değeri 0.62, 500 adet örnek için 0.588 ve 1000 adet örnek için 0.589 hesaplanmıştır. Telefonu olanların katsayısı lojistik regresyon tekniğinde 250 adet örnek için 0.332, 500 adet örnek için 0.316 ve 1000 adet örnek için 0.299 olarak hesaplanmıştır.

Tablo 13'den görüleceği üzere %20'lik veri setine diskriminant analizi sonucu bulunan katsayılar kullanılarak sınıflama yapıldığında doğru sınıflama oranı sadece %42 olmuştur. Yeniden örnekleme tekniğini kullanarak elde edilen tahmin değerlerinin ortalamaları kullanıldığında doğru sınıflama oranı %58 çıkmaktadır. Lojistik regresyon sonucu bulunan doğru sınıflama oranlarında değişiklik olmadığı görülmektedir. Tahmin edilen değerlerin dağılımları bu durumu açıklamak için bir cevap olabilir. Diskriminant analizinde yeniden örnekleme tekniği kullanıldığında dağılımların normale yakın olduğunu, fakat lojistik regresyon ile elde edilen tahmin değerlerinin belli bir değer üzerinde yoğunlaşmış olduğunu ve bu değerin de lojistik regresyon sonucu elde edilmiş olan değere yakın olduğu görülmektedir. Diskriminant analizi ve lojistik regresyon ile elde edilen tahmin değerlerinin dağılımı aşağıda her bir değişken için tek tek verilmiştir.

Tablo 13. Diskriminant ve Lojistik Regresyon Doğru Sınıflama Oranları.

	Diskriminant Analizi	Lojistik Regresyon
Fonksiyon Katsayıları Kullanılırsa.	0.42	0.68
N=250	0.58	0.68
N=500	0.57	0.68
N=1000	0.58	0.68

Bootsrap ile elde edilen katsayıların dağılım şekil 3'de verilmektedir.



Şekil 3. Yeniden Örnekleme Sonucu Bulunan Sabit Terimin Dağılımı

Sol tarafta Diskriminant analizi, sağ tarafta ise lojistik regresyon sonucu elde edilmiş olan sabit teriminin N=250, N=500 ve N=1,000 adet yeniden örnek için dağılımı verilmektedir.

5.3 Finans Kurumunun Veri Seti

Uygulamanın bu bölümünde ise kredi kartı veren bir kuruluştan sağlanmış olan 4037 tane kredi kartı sahibinin durumları incelenecektir. Diskriminant ve Lojistik regresyon teknikleri kullanılarak kredinin geri ödenip ödenmeyeceği tahmin eden fonksiyonlar elde edilecektir. Veri setindeki 3037 (%75) kredi kartı sahibinin verileri kullanılarak fonksiyonlar tahmin edilecek 1000 (%25) adet veri ise bulunan fonksiyon katsayılarının test edilmesi için kullanılacaktır. Yeniden örnekleme sonrasında elde edilen fonksiyon katsayılarının ortalaması bulunacak ve 1000 adet veri için doğru sınıflama oranına bakılacaktır. Veri setini oluşturan bağımsız değişkenler şunlardır:

- Yaş
- Cinsiyet
- Öğrenim durumu
- Medeni hal
- Ev ve iş telefonunu bilgisinin verilmesi
- Cep telefonunun verilmesi
- İkamet durumu
- Çalışma süresi
- Mesleği
- Ekstre adresi
- Ek kart isteği
- Otomatik ödeme emri

Her bir sürekli değişkenin kategorileri veri setinin alındığı kuruluş tarafından daha önceden gruplanmıştır. Bu sebepten dolayı bu uygulamada bir öncekinden farklı olarak bütün bağımsız değişkenler kukla değişken olarak kullanılacaktır.

Yaş değişkeni altı kategoriye ayrılmıştır. 19 yaşından küçük olanlar, 20-22 yaş arasında olanlar, 23-26 yaş arasında olanlar, 27-31 yaşları arasında olanlar, 32-39 yaşları arasında olanlar, 40-45 yaşları arasında olanlar, 46 yaşından büyük olanlar ve yaş bilgisini boş bırakanlar.

Cinsiyet bağımsız değişkeni üç kategoriden oluşmaktadır. Erkek, kadın ve bilgi vermeyenler.

Öğrenim durumu yedi kategoriden oluşmaktadır. İlkokul mezunları, ortaokul mezunları, lise mezunları, iki yıllık ön lisans mezunları, dört yıllık lisans mezunları, yüksek lisans veya doktorası olanlar ve bu alanı boş bırakanlar.

Medeni hal bağımsız değişkeni ise dört kategoriden oluşmaktadır. Evli olanlar, bekar olanlar, dul veya boşanmış olanlar, diğer veya boş bırakanlar.

Ev ve iş telefonunun belirtilmesi bağımsız değişkeni dört kategoriye ayrılmıştır. Hem iş hem de ev telefonun verenler, sadece iş telefonunu verenler, sadece ev telefonunu verenler ve hem iş hem de ev telefonunu vermeyenler.

Cep telefonu bağımsız değişkeni ise iki kategoriden oluşmaktadır. Cep telefonunu verenler ve cep telefonunu vermeyenler.

İkamet durumu bağımsız değişkeni altı kategoriden oluşmaktadır. Ev sahibi olanlar, kiracılar, şirket evinde oturanlar, ailesinin evinde oturanlar, diğer ve son olarak bu alanı işaretlemeyenler.

Çalışma süresi yedi kategoriye ayrılmıştır. Bir yıldan az çalışanlar, 13-18 ay arasında çalışmış olanlar, 19-23 aya arasında çalışmış olanlar, 24-35 ay arasında çalışmış olanlar, 36-47 ay arasında çalışmış olanlar, 48 aydan fazla süredir çalışıyor olanlar ve emekli veya ev kadını veya öğrenci olanlar.

Meslek bağımsız değişkeni on kategoriye ayrılmıştır. Kamu kesiminde çalışanlar, özel sektörde çalışanlar, şirket ortağı olanlar, küçük işletme sahipleri, tüccar olanlar, emekli olanlar, ev kadını olanlar, öğrenci olanlar, işsizler ve bu alanı boş bırakanlardır.

Ekstre adresi değişkeni üç kategoriye ayrılmıştır. Ev adresine isteyenler, iş adresine isteyenler ve bu alanı boş bırakanlar.

Ek kart bağımsız değişkeni iki kategoriden oluşmaktadır. Ek kart talep edenler ve ek kart talep etmeyenler.

Otomatik ödeme talimatı bağımsız değişkeni dört kategoriden oluşmaktadır. Hesabından otomatik olarak en az miktarın çekilmesini kabul edenler, bütün borcun çekilmesini kabul edenler, otomatik ödeme talimatı verip de yukarıdaki ödeme şekillerinden hangisi olacağını belirtmemiş olanlar ve otomatik ödeme talimatı vermeyenler.

Eldeki veri seti, her bir bağımsız değişkenin kategorik başlıkları kukla değişken olarak diskriminant ve lojistik fonksiyonlarında yerine koyarak sonuçlar analiz edilecektir. Bazı kategorik değişkenler örnek veri seti içerisinde hiç gözlenmediğinden bu kategoriler bağımsız değişkenlerin arasından çıkarılarak kırk adet bağımsız değişken ile model tahmin edilecektir.

Tablo 14 ödenen ve ödenmeyen kredi borçlarının her bir değişken için frekans tablosunu vermektedir. Her bir değişken için kredi borcunu ödeyenler ve kredi borcunu ödemeyenlerin sayıları ve yüzde oranı verilmiştir. Örneğin; 20-22 yaş arasında olanların

kredi kartı borcunu ödeyenler 60 kişi (%53) ve kredi kartı borcunu ödemeyenler 54 kişidir (%47).

Tablo 14. Değişkenlerin Kategorilerine Göre Kredi Kartı Borcunu Ödeyen ve Ödenmeyenlerin Frekans Tablosu.

	Kredi Borcunu Ödeyenler	Yüzde	Kredi Borcunu Ödemeyenler	Yüzde
20-22 Yaş Arası	60	53%	54	47%
23-26 Yaş Arası	321	58%	235	42%
26-31 Yaş Arası	498	63%	294	37%
32-39Yaş Arası	644	68%	297	32%
40-45 Yaş Arası	313	68%	150	32%
46 Yaş ve Üstü	22	13%	144	87%
Erkek	1325	58%	942	42%
İlkokul Mezunu	145	51%	137	49%
Ortaokul Mezunu	738	56%	590	44%
Lise Mezunu	102	63%	59	37%
İki Yıllık Üniversite Mezunu	581	77%	176	23%
Dört Yıllık Üniversite Mezunu	37	93%	3	8%
Mastır veya Doktora	157	68%	75	32%
Evli	1185	62%	740	38%
Bekar	617	61%	400	39%
İş Telefonunu Verenler	125	54%	105	46%
Hem iş Hem de Ev Telefonunu Verenler	1633	63%	979	37%
Cep Telefonunu Verenler	1585	64%	897	36%
Ev Sahibi Olanlar	915	67%	458	33%
Kiracı Olanlar	397	56%	316	44%
Şirketi Evi Olanlar	54	59%	38	41%
Ailesinin Evi Olanlar	388	56%	304	44%
13 –18 Ay Arası Çalışanlar	168	55%	136	45%
19 –23 Ay Arası Çalışanlar	229	57%	175	43%
24 –35 Ay Arası Çalışanlar	227	52%	212	48%
36 –47 Ay Arası Çalışanlar	151	52%	139	48%
48 Aydan Fazla Çalışanlar	1027	69%	472	31%
Kamu Kesiminde Çalışanlar	362	64%	206	36%
Özel Sektörde Çalışanlar	890	56%	706	44%

Tablo 14.(Devam)

Şirket Ortakları	214	80%	52	20%
Küçük İşletme Sahipleri	71	63%	42	37%
Esnaf	170	77%	52	23%
Emekli	14	18%	62	82%
Ev kadını	31	67%	15	33%
Öğrenci	105	72%	41	28%
Ekstresini Ev Adresi Yazanlar	1099	61%	717	39%
Ek Kart İsteyenler	93	66%	48	34%
Borcun Tamamının Otomatik Ödeme Emri Olanlar	1808	75%	595	25%
Otomatik Ödeme Emri Olup Seçim Yapmamış Olanlar	0	0%	351	100%
Otomatik Ödeme Emri Vermeyenler	0	0%	186	100%

5.3.1 Diskriminant Analizi Uygulaması ve Sonuçları

Diskriminant analizi sonucu bulunan fonksiyonun katsayıları Tablo 15'de verilmiştir. Tablo 15'den 20-22 yaşları arasındaki bir kişinin kredisini geri ödeme olasılığının 46 yaş ve üstüne göre daha yüksek olduğunu görmekteyiz.

Tablo 15. Diskriminant Analizi Sonucu Bulunan Fonksiyon Katsayıları.

	Fonksiyon Katsayısı		Fonksiyon Katsayısı
Sabit Terim	-0.672	Şirketi Evi Olanlar	-0.017
20-22 Yaş Arası	0.5	Ailesinin Evi Olanlar	0.007
23-26 Yaş Arası	0.594	13 –18 Ay Arası Çalışanlar	0.291
26-31 Yaş Arası	0.789	19 –23 Ay Arası Çalışanlar	0.059
32-39Yaş Arası	0.884	24 –35 Ay Arası Çalışanlar	-0.034
40-45 Yaş Arası	0.816	36 –47 Ay Arası Çalışanlar	0.088
46 Yaş ve Üstü	-0.517	48 Aydan Fazla Çalışanlar	0.443
Erkek	-0.328	Kamu Kesiminde Çalışanlar	-0.76
İlkokul Mezunu	0.151	Özel Sektörde Çalışanlar	-0.917
Ortaokul Mezunu	0.345	Şirket Ortakları	-0.306
Lise Mezunu	0.624	Küçük İşletme Sahipleri	-0.61

Tablo 15 (Devam)

İki Yıllık Üniversite Mezunu	0.946	Esnaf	-0.373
Dört Yıllık Üniversite Mezunu	1.379	Emekli	-1.105
Mastır veya Doktora	0.603	Ev kadını	-1.449
Evli	0.158	Öğrenci	-0.594
Bekar	0.128	Ekstresini Ev Adresi Yazanlar	-0.113
İş Telefonunu Verenler	-0.323	Ek Kart İsteyenler	-0.082
Hem iş Hem de Ev Telefonunu Verenler	-0.23	Borcun Tamamının Otomatik Ödeme Emri Olanlar	0.647
Cep Telefonunu Verenler	0.166	Otomatik Ödeme Emri Olup Seçim Yapmamış Olanlar	-2.061
Ev Sahibi Olanlar	0.282	Otomatik Ödeme Emri Vermeyenler	-2.276
Kiracı Olanlar	-0.056		

Yukarıdaki tabloda verilen Diskriminant katsayıları kullanılarak 3037 kişi kredisini ödeyenler ve kredisini ödemeyenler olarak ayrıştırıldığında, tablo 16 elde edilmiştir.

Tablo 16. Diskriminant Analizinin Doğru Sınıfla Matrisi (Veri Setinin %75)

			Tahmin Edilen		
			Kredi Borcunu Ödeyenler	Kredi Borcunu Ödemeyenler	Toplam
Gözlenen	Adet	Kredi Borcunu Ödeyenler	1771	89	1860
		Kredi Borcunu Ödemeyenler	470	707	1177
	%	Kredi Borcunu Ödeyenler	95.2	4.8	100
	%	Kredi Borcunu Ödemeyenler	39.9	60.1	100

Tablo 16'ya göre doğru sınıflama oranı [(1771+707)/(3037)] %81.6'dır. Bu oran yüksek bir oran olmasına rağmen, katsayılar bulunmasında da aynı veri seti kullanıldığı için yanlıdır.

Yanlılık sorununu aşmak için daha önceden ayrılmış olan 1000 adet veri kullanılacak ve bu kişilerin içinde doğru sınıflama oranı bulunmuştur. 1000 kişiye ait sınıflama sonuçları tablo 17'de verilmiştir. 1000 kişilik veri seti kullanıldığında doğru sınıflama oranı %66'dır [(403+257)/1000].

Tablo 17. Diskriminant Analizi Doğru Sınıflama Matrisi (Veri Setinin %75).

			Tahmin Edilen		
			Kredi Borcunu Ödeyenler	Kredi Borcunu Ödemeyenler	Toplam
Gözlenen	Adet	Kredi Borcunu Ödeyenler	403	97	500
		Kredi Borcunu Ödemeyenler	243	257	500
	%	Kredi Borcunu Ödeyenler	81	19	100
		Kredi Borcunu Ödemeyenler	49	51	100

5.3.2 Lojistik Regresyon Uygulaması ve Sonuçları

Veri setinin %75'lik kısmı kullanıldığında Lojistik regresyon sonucu bulunan fonksiyonun katsayıları Tablo 18'de verilmiştir. 20-22 yaş grubundaki kredi müşterisinin katsayısının –0.112 olduğu bulunmuştur.

Tablo 18. Lojistik Regresyon Sonucu Bulunan Fonksiyon Katsayıları.

	Fonksiyon Katsayısı		Fonksiyon Katsayısı
Sabit Terim	-4.276	Şirketi Evi Olanlar	-0.05
20-22 Yaş Arası	-0.112	Ailesinin Evi Olanlar	-0.083
23-26 Yaş Arası	-0.266	13 –18 Ay Arası Çalışanlar	-0.608
26-31 Yaş Arası	-0.617	19 –23 Ay Arası Çalışanlar	-0.195
32-39Yaş Arası	-0.786	24 –35 Ay Arası Çalışanlar	-0.101
40-45 Yaş Arası	-0.687	36 –47 Ay Arası Çalışanlar	-0.328
46 Yaş ve Üstü	1.899	48 Aydan Fazla Çalışanlar	-0.949
Erkek	0.662	Kamu Kesiminde Çalışanlar	5.245
İlkokul Mezunu	-0.404	Özel Sektörde Çalışanlar	5.48
Ortaokul Mezunu	-0.698	Şirket Ortakları	4.129

Tablo 18. (Devam)

Lise Mezunu	-1.291	Küçük İşletme Sahipleri	4.961
İki Yıllık Üniversite Mezunu	-1.884	Esnaf	4.438
Dört Yıllık Üniversite Mezunu	-3.62	Emekli	6.618
Mastır veya Doktora	-1.119	Ev kadını	6.955
Evli	-0.235	Öğrenci	4.9
Bekar	-0.145	Ekstresini Ev Adresi Yazanlar	0.25
İş Telefonunu Verenler	0.951	Ek Kart İsteyenler	0.088
Hem iş Hem de Ev Telefonunu Verenler	0.809	Borcun Tamamının Otomatik Ödeme Emri Olanlar	-0.788
Cep Telefonunu Verenler	-0.378	Otomatik Ödeme Emri Olup Seçim Yapmamış Olanlar	10.525
Ev Sahibi Olanlar	-0.585	Otomatik Ödeme Emri Vermeyenler	11.614
Kiracı Olanlar	0.057		

Tablo 19'da 3037 (veri setinin %75) kredi müşterisinden elde edilmiş olan lojistik regresyon fonksiyonunun katsayıları kullanılarak elde edilen doğru sınıflama oranı verilmiştir. Bu veri seti kullanıldığında doğru sınıflama oranının %81.6 [(1761+719)/3037] olduğu bulunmuştur. Bu oran diskriminant analizinde olduğu gibi yanlıdır.

Tablo 19. Lojistik Regresyon Doğru Sınıflama Tablosu (Veri Setinin %75).

			Tahmin Edilen		
			Kredi Borcunu Ödeyenler	Kredi Borcunu Ödemeyenler	Toplam
Gözlenen	Adet	Kredi Borcunu Ödeyenler	1761	99	1860
		Kredi Borcunu Ödemeyenler	458	719	1177
	%	Kredi Borcunu Ödeyenler	95	5	100
		Kredi Borcunu Ödemeyenler	39	61	100

Bulunan katsayıların ne kadar güvenilir olduğunu anlamak için ayrılmış olan 1000 (veri setinin %25) kişilik veri seti kullanılacaktır. Tablo 20 de görüleceği gibi doğru sınıflama oranı sadece %32 [(65+256)/1000] olarak bulunmuştur. Bu oran oldukça düşüktür.

Tablo 20. Lojistik Regresyon Doğru Sınıflama Tablosu (Veri Setinin %25)

			Tahmin Edilen		
			Kredi Borcunu Ödeyenler	Kredi Borcunu Ödemeyenler	Toplam
Gözlenen	Adet	Kredi Borcunu Ödeyenler	65	435	500
		Kredi Borcunu Ödemeyenler	244	256	500
	%	Kredi Borcunu Ödeyenler	13	87	100
		Kredi Borcunu Ödemeyenler	49	51	100

5.3.3 Bootstrap Uygulaması Sonuçları

Çalışmanın bundan sonraki kısmında amaç yeniden örnekleme sonucu bulunan 250, 500 ve 1000 adet örnek için elde edilmiş olan Diskriminant ve Lojistik regresyon fonksiyonunun katsayılarının ortalamalarını alarak 1000 kişilik kredi kartı sahiplerinin hangi oranda doğru tahmin edildiğini belirlemektir. Tablo 21 bootstrap sonucu elde edilen örnekler için katsayıların ortalamalarını göstermektedir. 20-22 yaş grubundaki bir kişinin 250 örnek için hesaplanan katsayı ortalaması 0.4713, 500 örnek için hesaplanan katsayı ortalaması 0.5247 ve 1000 örnek için hesaplanan katsayı ortalaması 0.5145'dir.

Tablo 21. Diskriminant Fonksiyonu Yeniden Örnekleme Ortalamaları

	N=250	N=500	N=1000
Sabit	-0.4101	-0.5025	-0.4778
20-22 Yaş Arası	0.4713	0.5247	0.5145
23-26 Yaş Arası	0.5667	0.6161	0.6092
26-31 Yaş Arası	0.7584	0.8109	0.8037
32-39Yaş Arası	0.8574	0.9099	0.9016
40-45 Yaş Arası	0.7845	0.8409	0.8314
46 Yaş ve Üstü	-0.5702	-0.5190	-0.5198
Erkek	-0.3390	-0.3333	-0.3349
İlkokul Mezunu	0.1526	0.1496	0.1496
Ortaokul Mezunu	0.3417	0.3454	0.3440
Lise Mezunu	0.6245	0.6351	0.6306
İki Yıllık Üniversite Mezunu	0.9462	0.9447	0.9457
Dört Yıllık Üniversite Mezunu	1.3795	1.3815	1.3801
Mastır veya Doktora	0.5986	0.5980	0.5973
Evli	0.1476	0.1451	0.1453
Bekar	0.1093	0.1088	0.1109
İş Telefonunu Verenler	-0.3160	-0.3175	-0.3147
Hem iş Hem de Ev Telefonunu Verenler	-0.2299	-0.2241	-0.2259
Cep Telefonunu Verenler	0.1726	0.1738	0.1723
Ev Sahibi Olanlar	0.3030	0.2940	0.2980
Kiracı Olanlar	-0.0438	-0.0514	-0.0474
Şirketi Evi Olanlar	-0.0135	-0.0167	-0.0116
Ailesinin Evi Olanlar	0.0256	0.0201	0.0233
13 -18 Ay Arası Çalışanlar	0.2924	0.2830	0.2893
19 -23 Ay Arası Çalışanlar	0.0513	0.0445	0.0518
24 -35 Ay Arası Çalışanlar	-0.0376	-0.0458	-0.0401
36 -47 Ay Arası Çalışanlar	0.0881	0.0838	0.0841
48 Aydan Fazla Çalışanlar	0.4419	0.4359	0.4415
Kamu Kesiminde Çalışanlar	-0.7571	-0.7291	-0.7450
Özel Sektörde Çalışanlar	-0.9151	-0.8881	-0.9045
Şirket Ortakları	-0.2961	-0.2724	-0.2894
Küçük İşletme Sahipleri	-0.5920	-0.5754	-0.5902
Esnaf	-0.3614	-0.3400	-0.3537
Emekli	-1.1051	-1.0728	-1.0903
Ev kadını	-1.4895	-1.4542	-1.4639
Öğrenci	-0.5939	-0.5594	-0.5772
Ekstresini Ev Adresi Yazanlar	-0.1231	-0.1179	-0.1183

Tablo 21. (Devam)

Ek Kart İsteyenler	-0.0868	-0.0931	-0.0884
Borcun Tamamının Otomatik Ödeme Emri Olanlar	0.6352	0.6513	0.6434
Otomatik Ödeme Emri Olup Seçim Yapmamış Olanlar	-2.0986	-2.0873	-2.0920
Otomatik Ödeme Emri Vermeyenler	-2.3142	-2.3004	-2.30522

Tablo 22'de bootstrap örneklerinin lojistik regresyon analizi sonrası hesaplanan ortalama değerlerini göstermektedir. 250 adet örnek kullanılarak hesaplanan 20-22 yaş grubundaki bir kişinin ortalama katsayı değeri 0.0471 olarak hesaplanmıştır. Örnek hacmini 500 alırsak 20-22 yaş grubundaki bir kişinin ortalama katsayısı –0.425 olarak hesaplanmıştır. Örnek hacmini 1000 alırsak 20-22 yaş grubundaki bir kişinin ortalama katsayısı –0.44 olarak hesaplanmıştır.

Tablo 22. Lojistik Regresyon Yeniden Örnekleme Ortalamaları.

	N=250	N=500	N=1000
Sabit	-6.4192	-6.0376	-6.1891
20-22 Yaş Arası	0.0471	-0.4250	-0.4400
23-26 Yaş Arası	-0.1316	-0.5874	-0.5841
26-31 Yaş Arası	-0.4846	-0.9427	-0.9493
32-39Yaş Arası	-0.6523	-1.1022	-1.1059
40-45 Yaş Arası	-0.5710	-1.0165	-1.0114
46 Yaş ve Üstü	2.1481	1.6670	1.6714
Erkek	0.6882	0.6811	0.6699
İlkokul Mezunu	-0.4365	-0.4198	-0.4127
Ortaokul Mezunu	-0.7357	-0.7086	-0.7193
Lise Mezunu	-1.3514	-1.3251	-1.3273
İki Yıllık Üniversite Mezunu	-1.9437	-1.9245	-1.9259
Dört Yıllık Üniversite Mezunu	-5.3182	-5.4831	-5.4194
Mastır veya Doktora	-1.1791	-1.1414	-1.1532
Evli	-0.2248	-0.2281	-0.2333
Bekar	-0.1225	-0.1346	-0.1388
İş Telefonunu Verenler	1.0110	1.0089	0.9932
Hem iş Hem de Ev Telefonunu Verenler	0.8792	0.8772	0.8535
Cep Telefonunu Verenler	-0.3676	-0.3744	-0.3867

Tablo 22. (Devam)

Ev Sahibi Olanlar	-0.5803	-0.5837	-0.6031
Kiracı Olanlar	0.0718	0.0784	0.0485
Şirketi Evi Olanlar	-0.0509	-0.0499	-0.0753
Ailesinin Evi Olanlar	-0.0776	-0.0785	-0.0868
13 -18 Ay Arası Çalışanlar	-0.5586	-0.6129	-0.6185
19 -23 Ay Arası Çalışanlar	-0.1517	-0.2012	-0.1820
24 -35 Ay Arası Çalışanlar	-0.0565	-0.0830	-0.0839
36 -47 Ay Arası Çalışanlar	-0.2919	-0.3210	-0.3156
48 Aydan Fazla Çalışanlar	-0.9233	-0.9507	-0.9418
Kamu Kesiminde Çalışanlar	7.1495	7.2399	7.4490
Özel Sektörde Çalışanlar	7.3862	7.4741	7.6882
Şirket Ortakları	6.0138	6.0746	6.2944
Küçük İşletme Sahipleri	6.8575	6.9409	7.1333
Esnaf	6.2972	6.4089	6.6121
Emekli	8.5614	8.6441	8.8819
Ev kadını	8.8871	9.0284	9.2185
Öğrenci	6.7709	6.8502	7.0790
Ekstresini Ev Adresi Yazanlar	0.2493	0.2622	0.2574
Ek Kart İsteyenler	0.0674	0.0925	0.0841
Borcun Tamamının Otomatik Ödeme Emri Olanlar	-0.8076	-0.8075	-0.7988
Otomatik Ödeme Emri Olup Seçim Yapmamış Olanlar	12.4233	12.3594	12.3746
Otomatik Ödeme Emri Vermeyenler	13.5942	13.7561	13.7454

Tablo 23'de elde edilen ortalama değerleri kullanılarak veri setinin %25 üzerinde tahminde bulunulacaktır. Diskriminant analizinde elde edilen katsayılar kullanıldığında doğru sınıflama oranı %66 olarak hesaplanmıştır. Bootstrap sonrası elde edilen ortalama katsayıların kullanılması durumunda ise %70'lik doğru sınıflama oranı hesaplanmıştır. Lojistik regresyondan elde edilen katsayılar kullanıldığında %32'lik doğru sınıflama oranı hesaplanmıştır. Bootstrap sonrası elde edilen katsayı ortalamaları kullanıldığında doğru sınıflama oranının değişmediği görülmektedir.

Tablo 23. Doğru Sınıflama Oranları (Veri Setinin %25).

	Diskriminant Analizi Doğru Sınıflama Oranı	Lojistik Regresyon Doğru Sınıflama Oranı
Fonksiyon Katsayıları Kullanılırsa	%66	%32
Yeniden Örnekleme N=250	%70	%32
Yeniden Örnekleme N=500	%70	%32
Yeniden Örnekleme N=1000	%70	%32

5.4 İki Uygulamanın Karşılaştırılması

Çalışmanın uygulama bölümünde iki ayrı veri seti kullanılmıştır. İlk veri seti Thomas tarafından sağlanan veri setidir. Veri seti öncelikle %80 ve %20 olmak üzere iki bölüme ayrılmış ve %80'lik kısmı ile model parametreleri tahmin edilmiş, yeniden örnekleme tekniği kullanılarak elde edilen 250, 500, 1000, adet alt örneklemden tahmin edilen parametrelerin ortalamaları bulunmuştur. Karşılaştırma yapabilmek için %20'lik veri seti üzerinde bulunan ağırlıklar kullanılarak doğru sınıflama oranları bulunmuştur. Sonuçlar tablo 24'de gösterilmiştir. Tablo 24'den de görüldüğü gibi doğru sınıflama oranı %68 çıkan Lojistik regresyon yönteminin doğru sınıflama oranı %42 çıkan Diskriminant analizine göre üstün olduğu anlaşılmaktadır. Yeniden örnekleme tekniği kullanıldığında ise Lojistik regresyonda bulunan doğru sınıflama oranı değişmezken, Diskriminant analizinde bulunan doğru sınıflama oranının artmış olduğu görülmektedir. Bunun olası sebebi olarak şekil 3'de yeniden örnekleme tekniği ile elde edilmiş olan katsayıların dağılımlarına bakmak bir cevap olacaktır. Dağılımların şekilleri incelendiğinde Lojistik regresyon katsayılarında fazla bir değişiklik gözlenmezken, Diskriminant analizi ile elde edilmiş katsayıların normal dağılıma benzer bir şekil alması ile açıklanabilir. Sonuç olarak yeniden örnekleme tekniğinin Diskriminant

olarak yeniden örnekleme tekniğinin Diskriminant analizinde doğru sınıflama oranında bir iyileştirme yaparken Lojistik regresyon tekniğinde etkili olmadığı görülmektedir.

Tablo 24. Doğru sınıflama Oranı (Birinci Veri Seti)

	Diskriminant Analizi	Lojistik Regresyon
Fonksiyon Katsayıları Kullanılırsa	0.42	0.68
Yeniden Örnekleme N=250	0.58	0.68
Yeniden Örnekleme N=500	0.57	0.68
Yeniden Örnekleme N=1000	0.58	0.68

İkinci çalışmanın verisi ise bir finans kurumundan alınmıştır. Veri seti %75 ve %25'lik bölümlere ayrılmış ve ilk kısmı ile Diskriminant analizi ve Lojistik regresyon katsayıları tahmin edilmiş. Yeniden örnekleme tekniği %75'lik veri seti üzerinde kullanılarak ana kütle parametreleri tahmin edilmiştir. Parametre tahminleri %25'lik veri seti üzerinde kullanılarak doğru sınıflama oranları bulunmuştur. İkinci veri setinin doğru sınıflama oranları tablo 25'de verilmiştir. Tablo 25'de Diskriminant analizinin Lojistik regresyon doğru sınıflama oranından büyük olduğu görülmektedir. Bir önceki uygulamada olduğu gibi Diskriminant analizi sonuçları Bootstrap ile iyileşirken Lojistik regresyon sonuçları değişmemeltedir.

Tablo 25. Doğru sınıflama oranı (İkinci Veri Seti)

	Diskriminant Analizi Doğru Sınıflama Oranı	Lojistik Regresyon Doğru Sınıflama Oranı
Fonksiyon Katsayıları Kullanılırsa	%66	%32
Yeniden Örnekleme N=250	%70	%32
Yeniden Örnekleme N=500	%70	%32
Yeniden Örnekleme N=1000	%70	%32

Sonuç

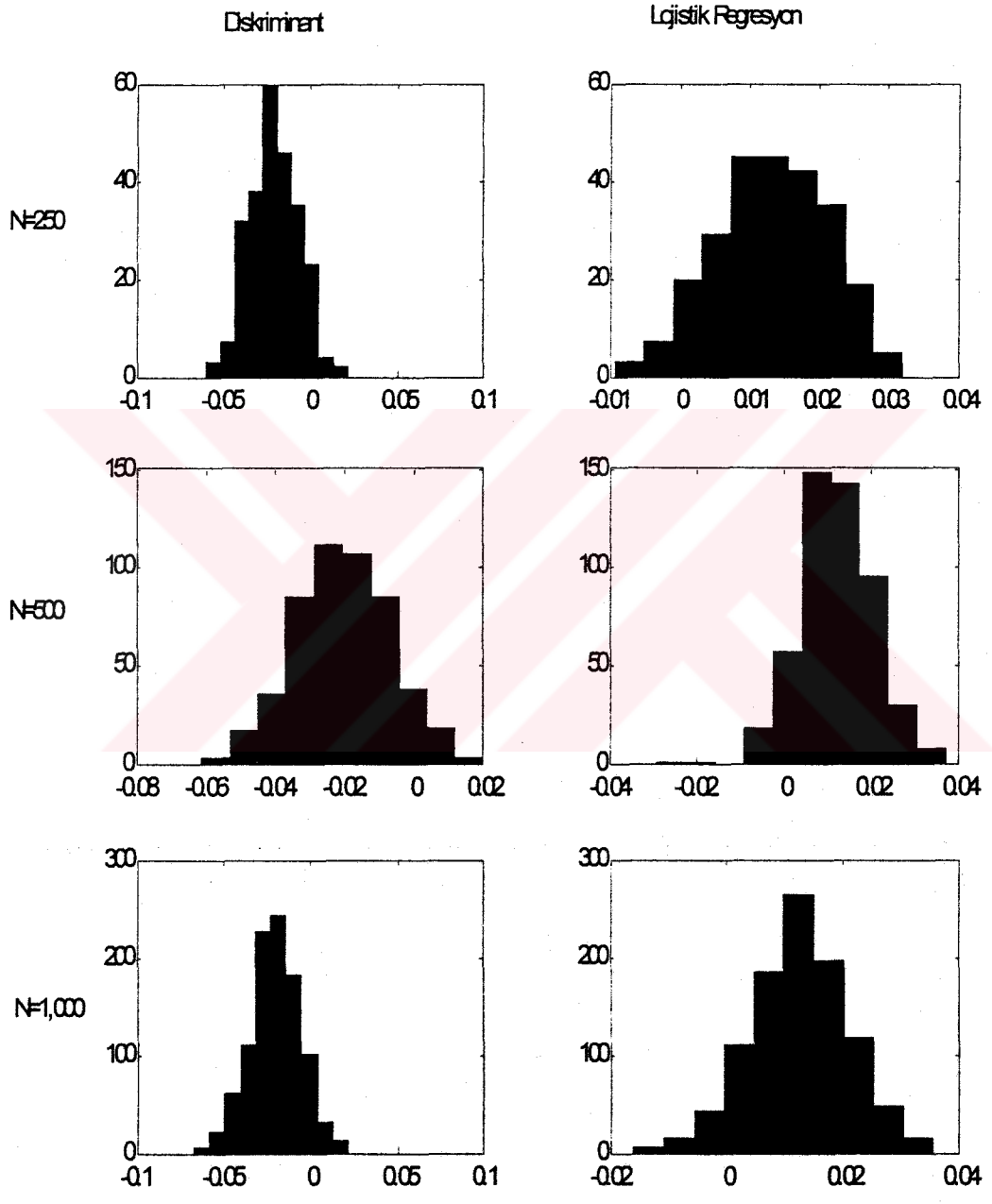
Bu çalışmada bir yeniden örnekleme tekniği olan Bootstrap tekniğinin Diskriminant Analizi ve Lojistik Regresyon tekniklerinde kullanımı amaçlanmıştır. Diskriminant analizi ve Lojistik regresyon tekniklerinde bulunan katsayılar bir tek örnek kullanılarak karar vermek yerine, yaratılan örneklerin oluşturduğu bir örneklem dağılımından elde edilen ortalamalar ana kütle parametresinin tahmin edicisi olarak kullanılmakta ve böylece daha etkili tahmin yapmak mümkün olmaktadır. Kredi kartı kullanımının bu kadar yaygın olduğu günümüzde kredi kartı borcunun geri ödenip ödenmemesi ekonomik anlamda oldukça önemlidir. Geçmişte kredi kartı sahibi olup kredi borcunu ödeyenler ve ödemeyenler açısından ayrıştırıcı bir fonksiyonun karar vericinin elinde olması önemlidir. Kredi kartının verilmesi kararında kullanılan skor kartı tekniği yerine istatistik bazı yöntemlerin kullanılması gerekliliği karar verici açısından doğruluğun ve güvenilirliğin sağlanmasında önemlidir.

Tablo 25'da ikinci çalışmaya ait doğru sınıflama oranları verilmiştir. Bu veri setinde bir önceki veri setinden farklı olarak Diskriminant analizi Lojistik regresyondan daha iyi sonuç verdiği görülmüştür.

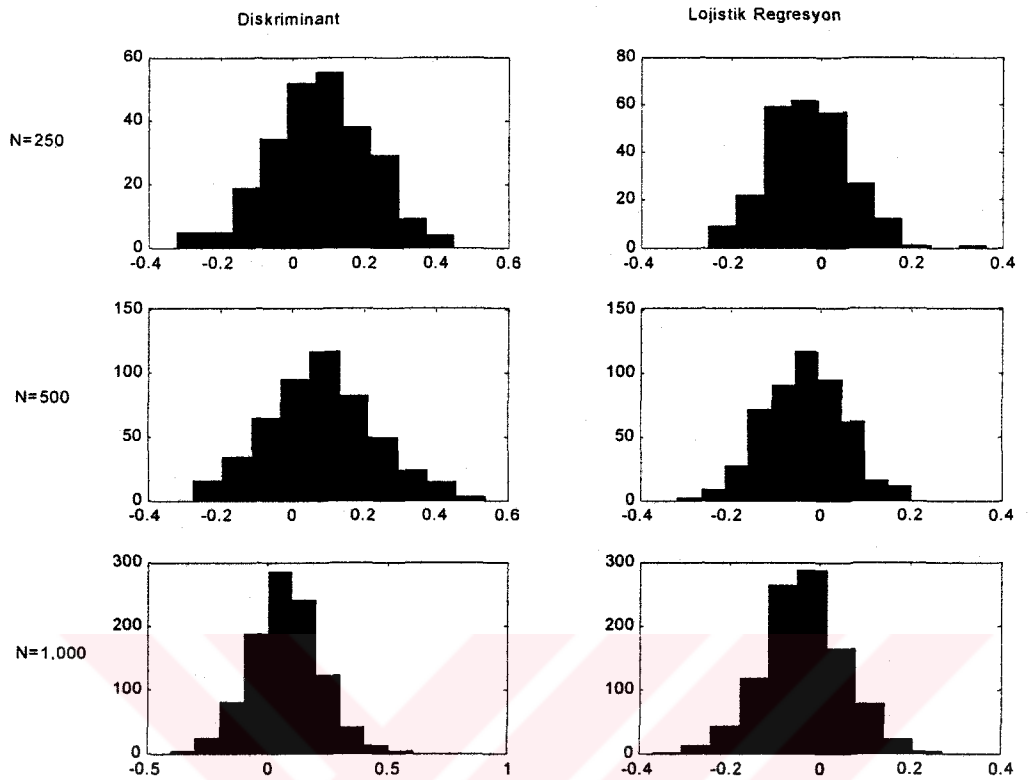
Sonuç olarak her iki uygulamadan elde edilen bulgular göstermektedirki yeniden örnekleme tekniğinin kullanılması Diskriminant Analizinde doğru sınıflama oranı açısından bir iyileştirme yapmış olmasına rağmen Lojistik Regresyonda bir iyileştirme yapmamıştır. Yeniden örnekleme tekniğinin kullanılması sonucunda Diskriminant Analizinde elde edilen tahmin edicilerin dağılımlarının normal dağılıma benzediğini Lojistik Regresyonda elde edilen katsayıların ise bir sayı etrafında toplandığı görülmektedir. Bu durumda cok değişkenli tekniklerden Diskriminant Analizinde bootstrap tekniğinin uygulanması Lojistik regresyon ile karşılaştırıldığında daha uygun sonuçlar elde edilmesini sağlayacaktır.

Ek 1

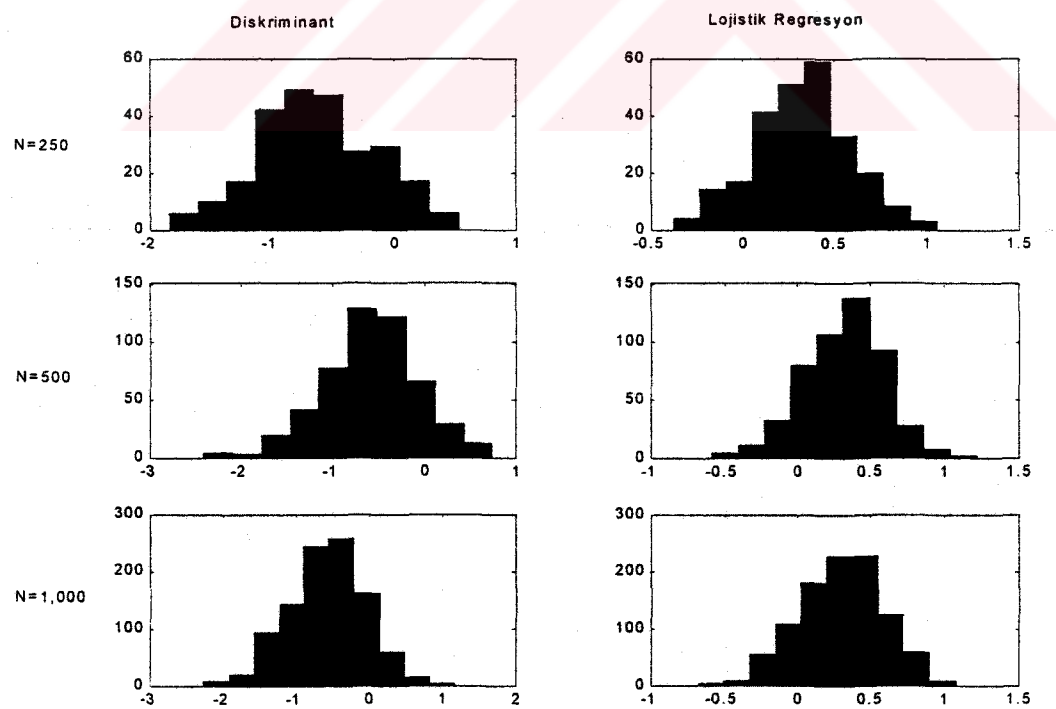
Birinci çalışmada yeniden örnekleme sonucu tahmin edilen parametrelerin dağılımı.



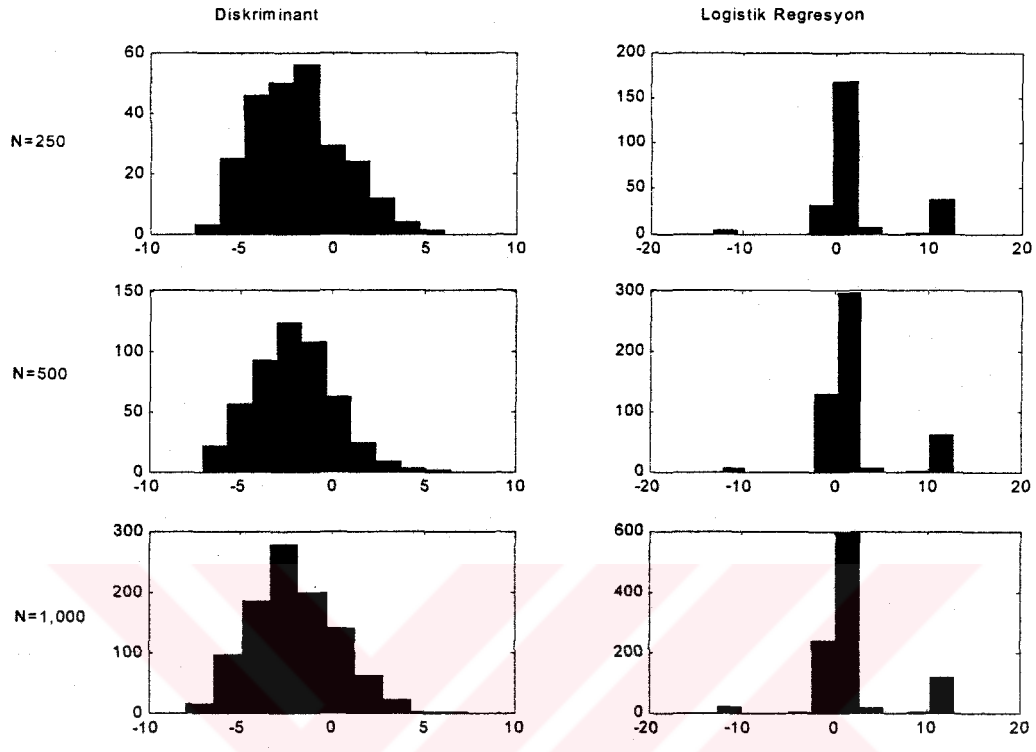
Şekil 4. Yaş Değişkeninin Yeniden Örnekleme Sonrası Bulunan Değerlerinin Dağılım



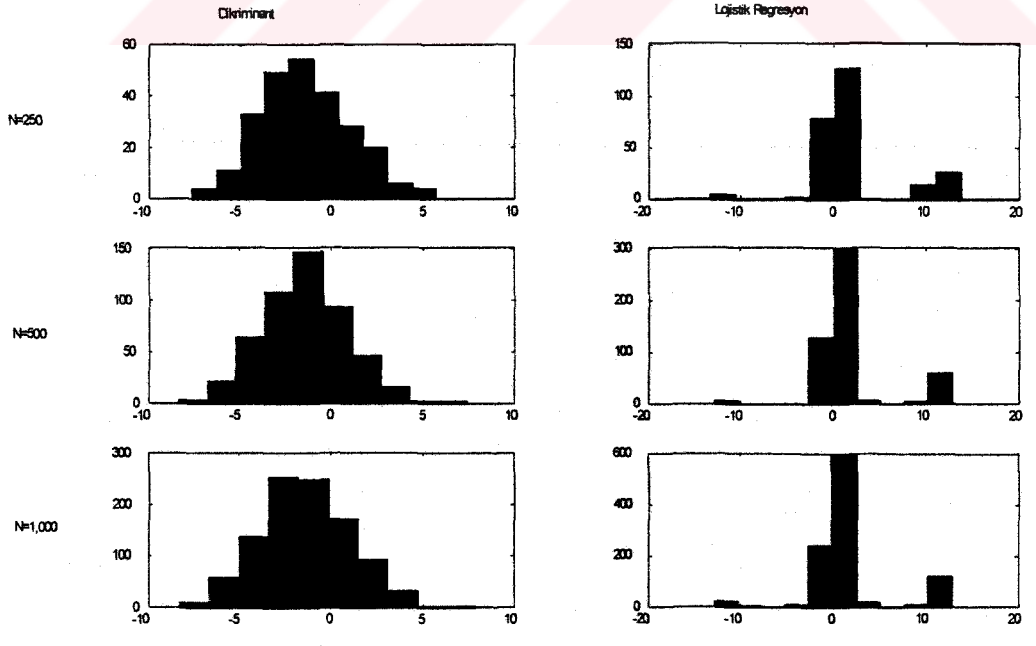
Şekil 5. Cocuk Sayısı Değişkeninin Yeniden Örnekleme Tahminlerinin Dağılımı.



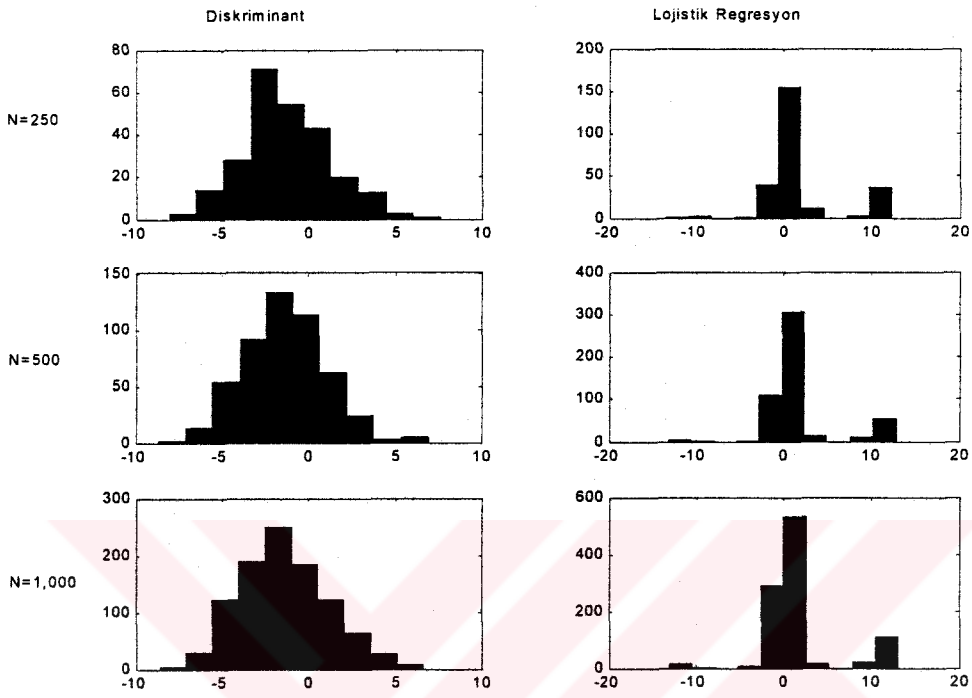
Şekil 6. Telefon Değişkeninin Yeniden Örnekleme Tahmin Değerlerinin Dağılımı.



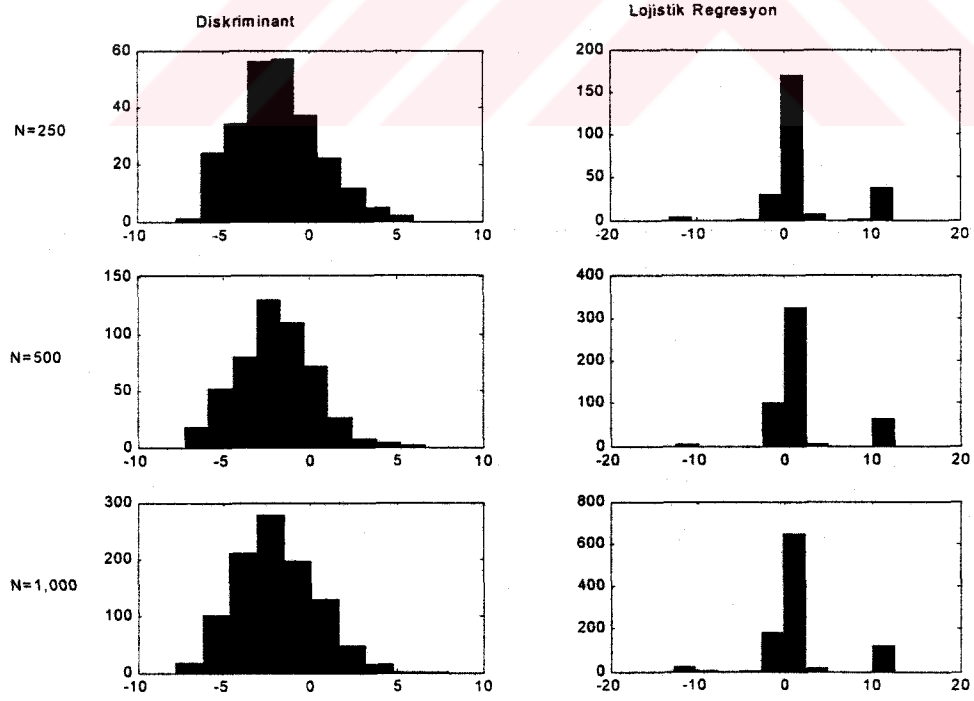
Şekil 7. Devlette Çalışanların Yeniden Örnekleme ile Tahmin Değerlerinin Dağılımı.



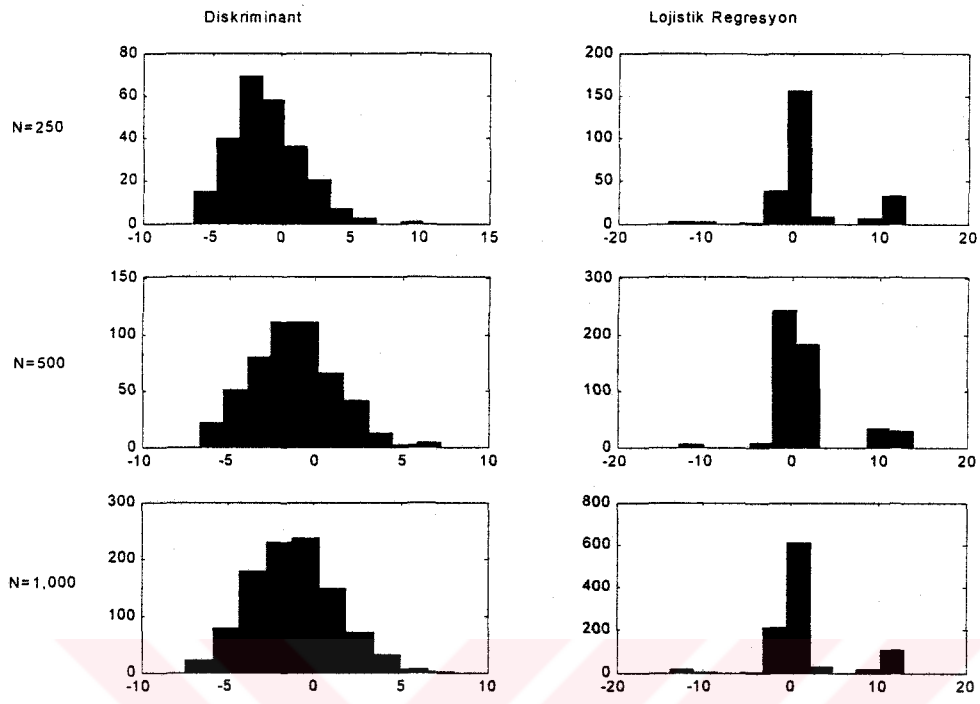
Şekil 8. Evkadınları Değişkeninin Yeniden Örnekleme Tahmin Değerlerinin Dağılımı.



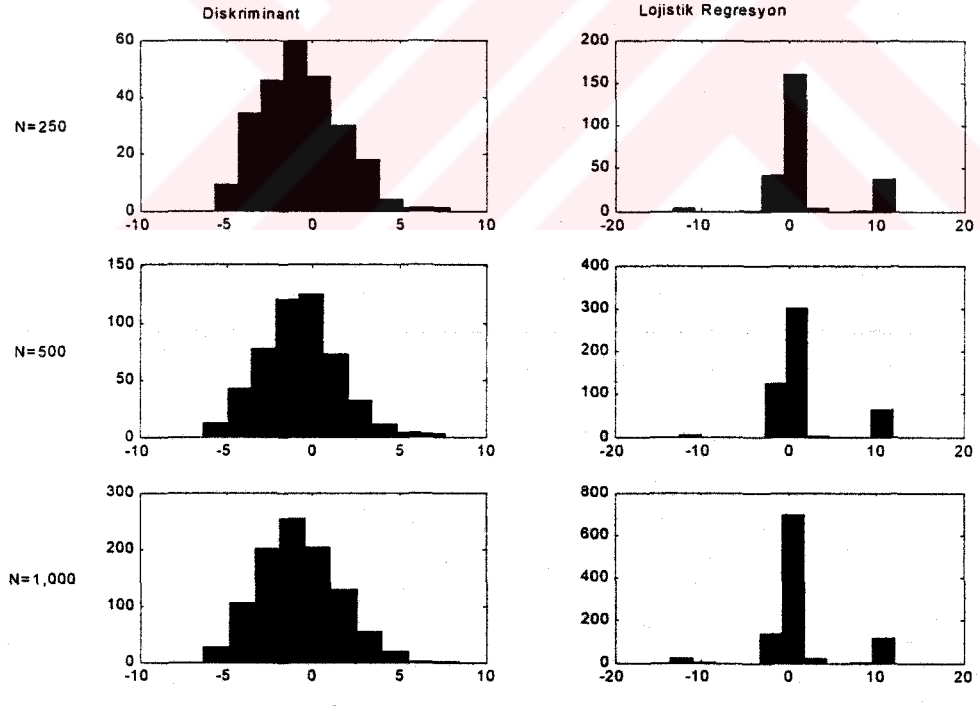
Şekil 9. Mesleği Asker Olanların Yeniden Örnekleme Tahminlerinin Dağılımı



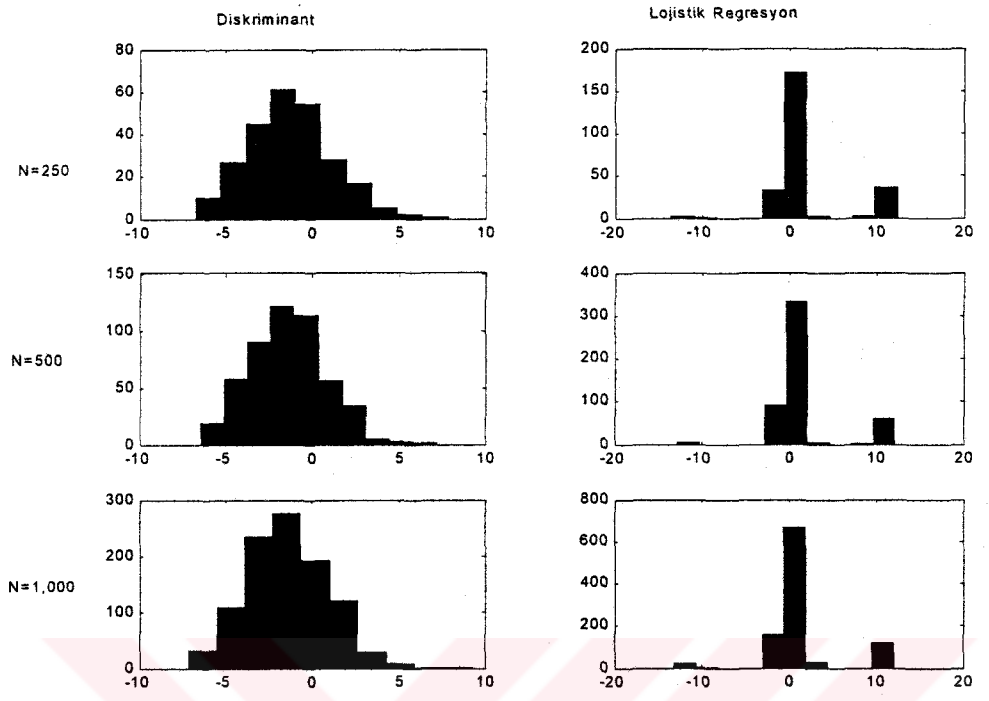
Şekil 10. Özel Sektörde Çalışanların Yeniden Örnekleme Tahmin Değerlerinin Dağılımı



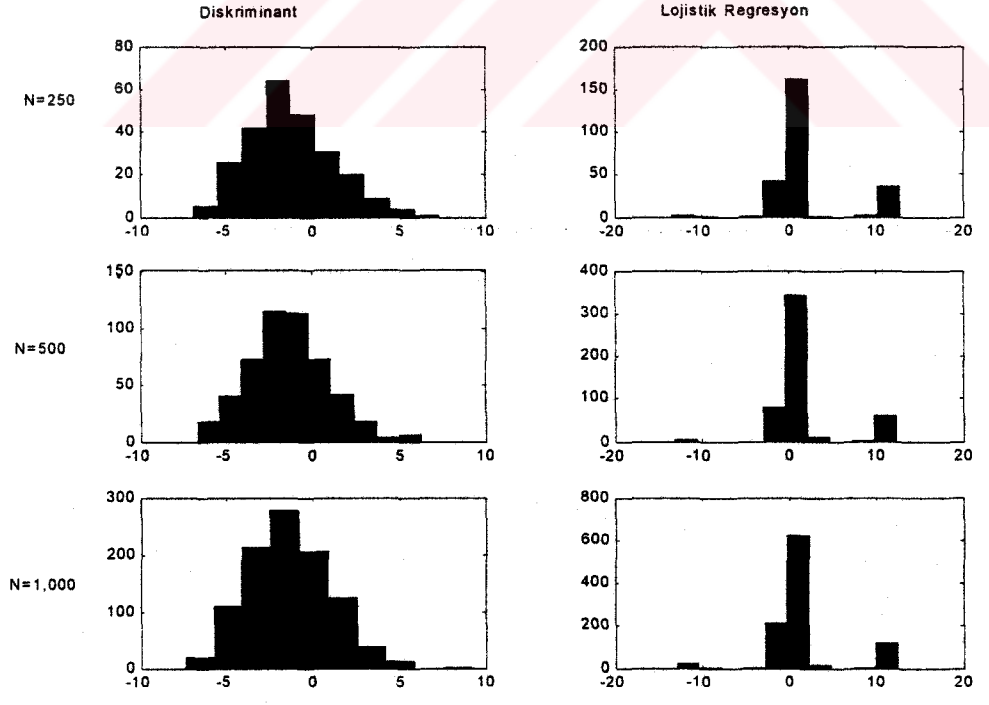
Şekil 11. Kamu Kesiminde Çalışanların Yeniden Örnekleme Tahminlerinin Dağılımı.



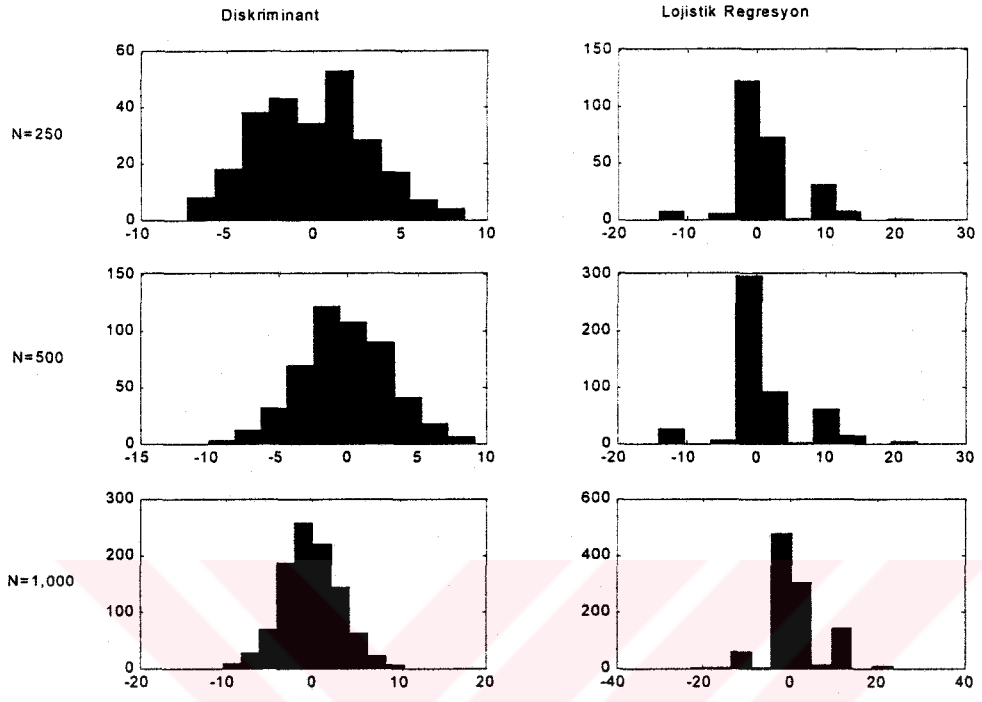
Şekil 12. Emeklilerin Yeniden Örnekleme Tahminlerinin Dağılımı.



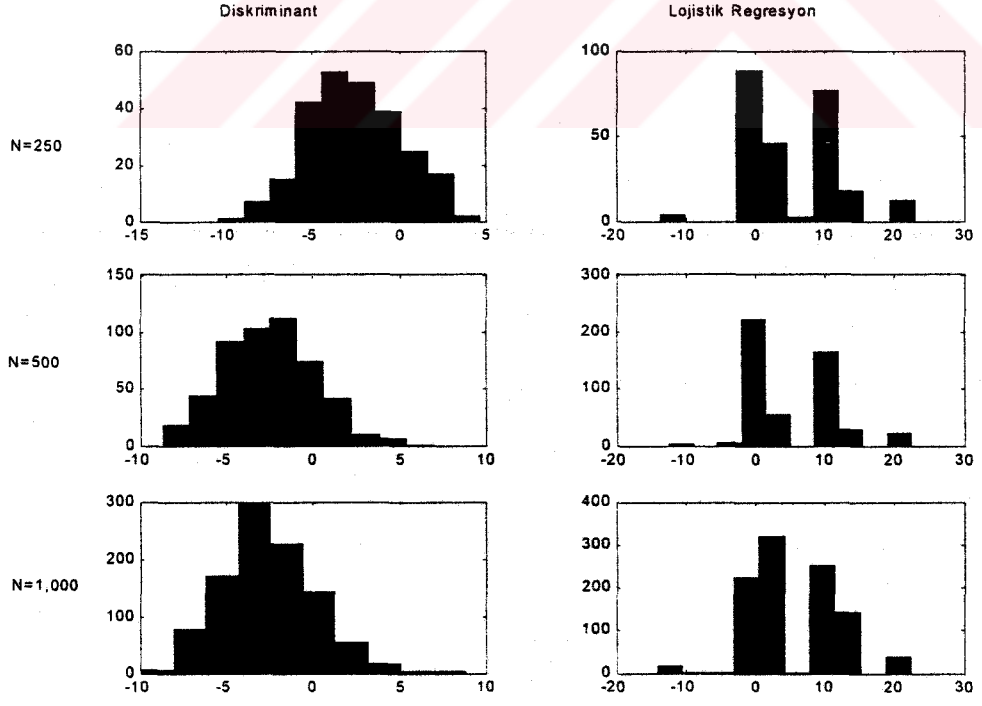
Şekil 13. Esnafın Yeniden Örnekleme Tahmin Değerlerinin Dağılımı.



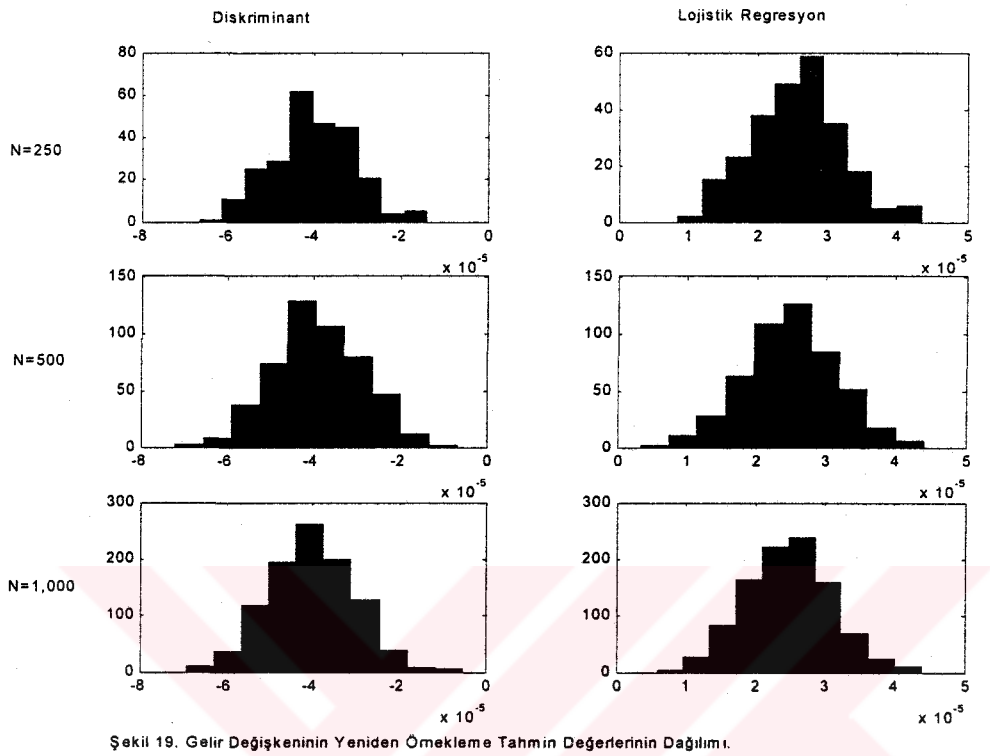
Şekil 14. Ögrencilerin Yeniden Örnekleme Tahminlerinin Dağılımı



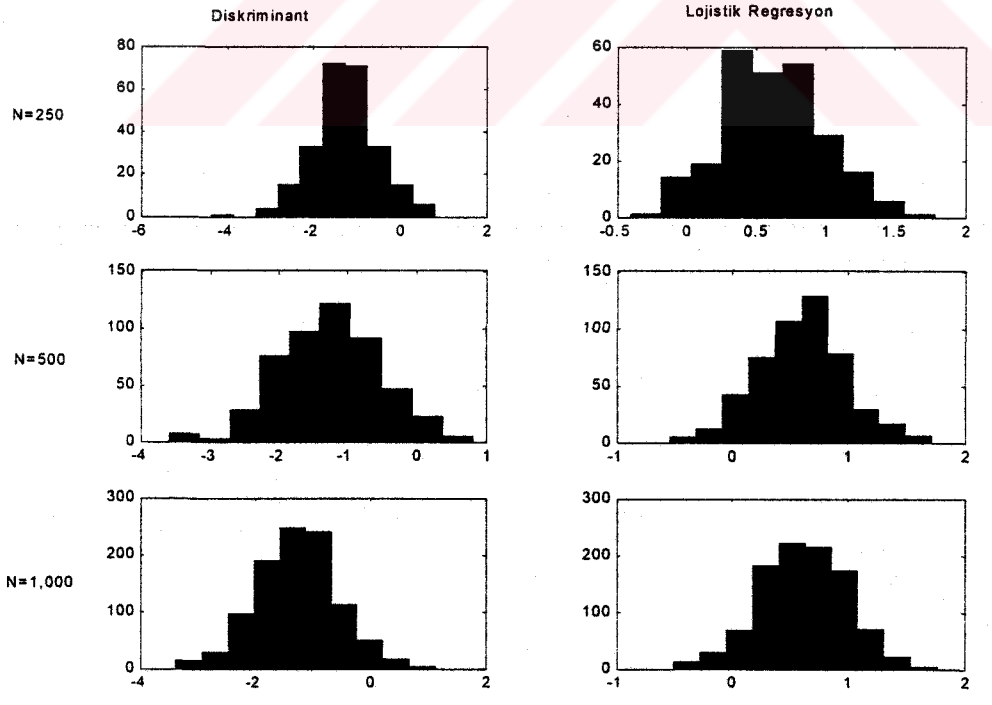
Şekil 15. İşsizlerin Yeniden Örnekleme Tahmin Değerlerinin Dağılımı.



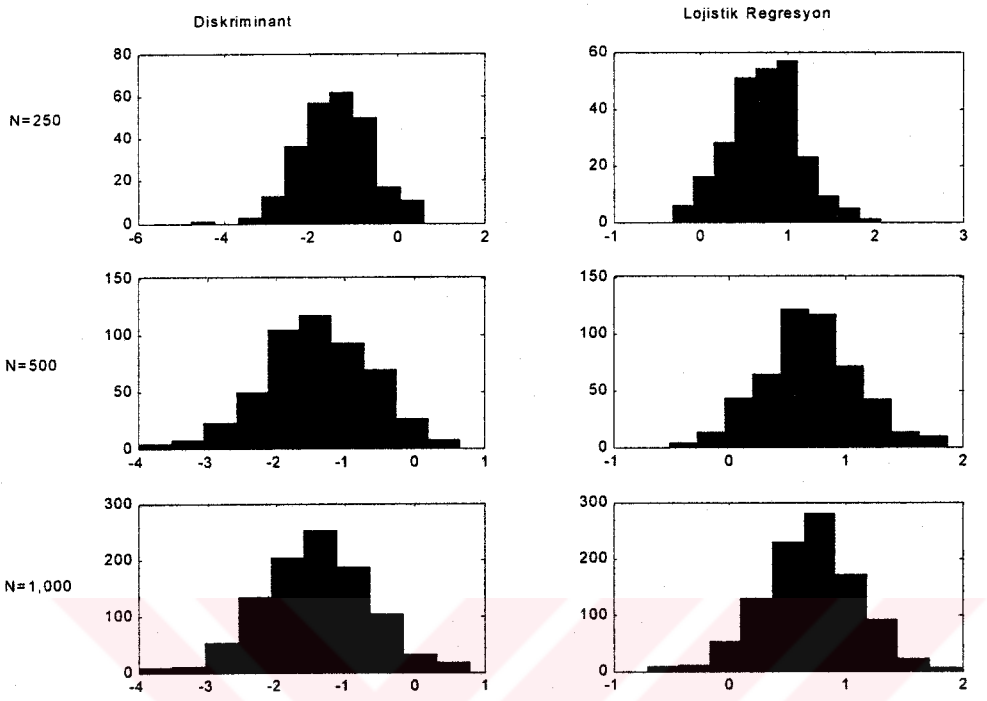
Şekil 16. Diğer Meslek Grubundakilerin Yeniden Örnekleme Tahmin Değerlerinin Dağılımı.



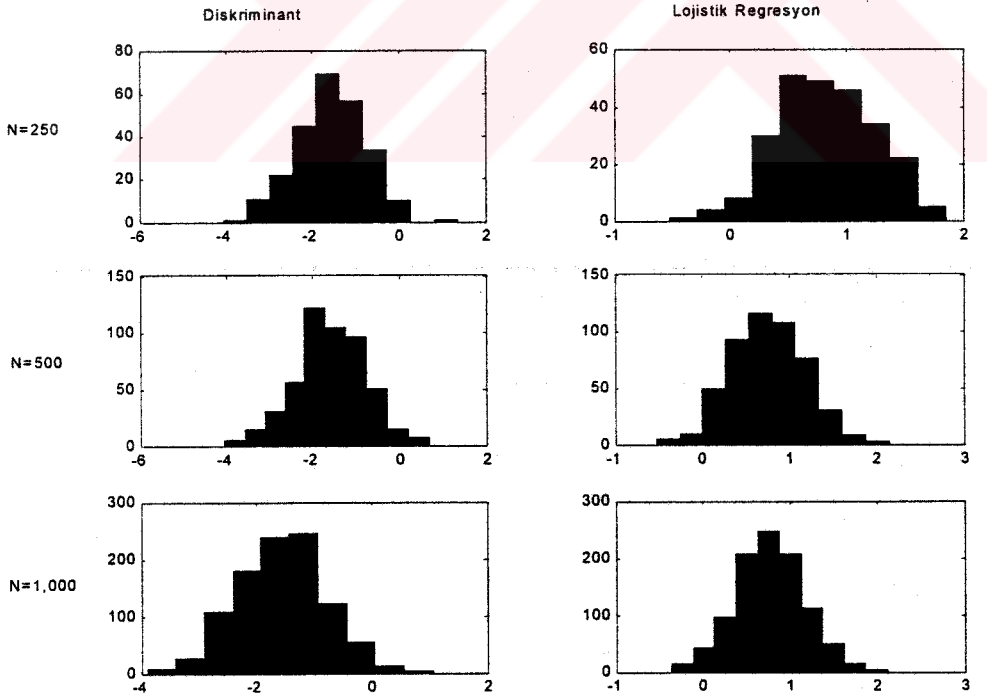
Şekil 19. Gelir Değişkeninin Yeniden Örnekleme Tahmin Değerlerinin Dağılımı.



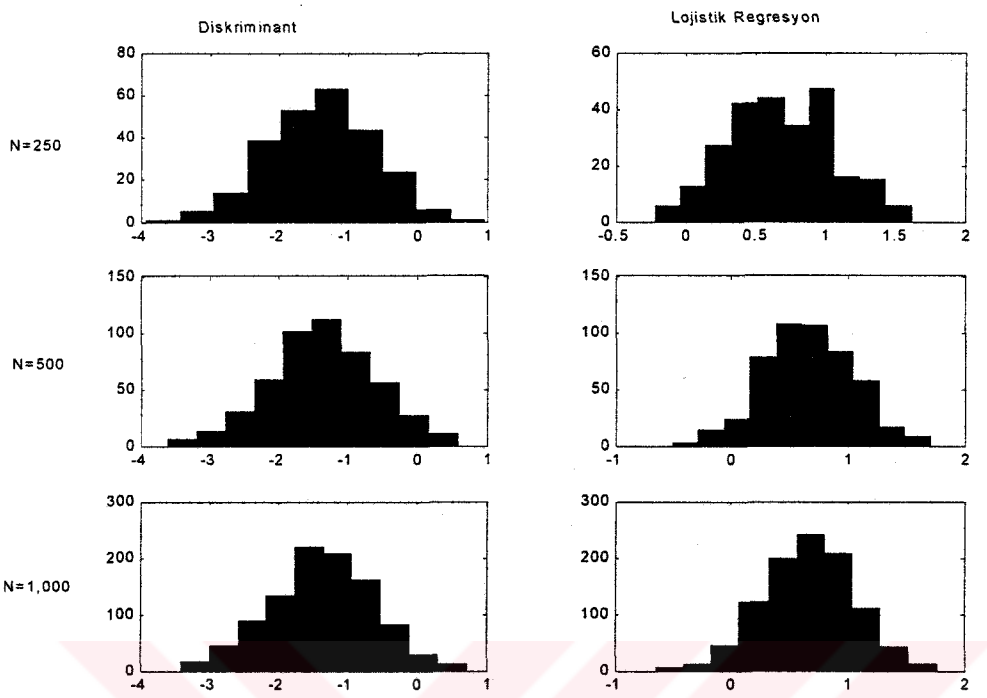
Şekil 20.Ev Sahibi Olanların Yeniden Örnekleme Tahmin Değerlerinin Dağılımı.



Şekil 21. Eşyalı Evde Kiracı Olanların Yeniden Örnekleme Tahminlerinin Dağılımı.

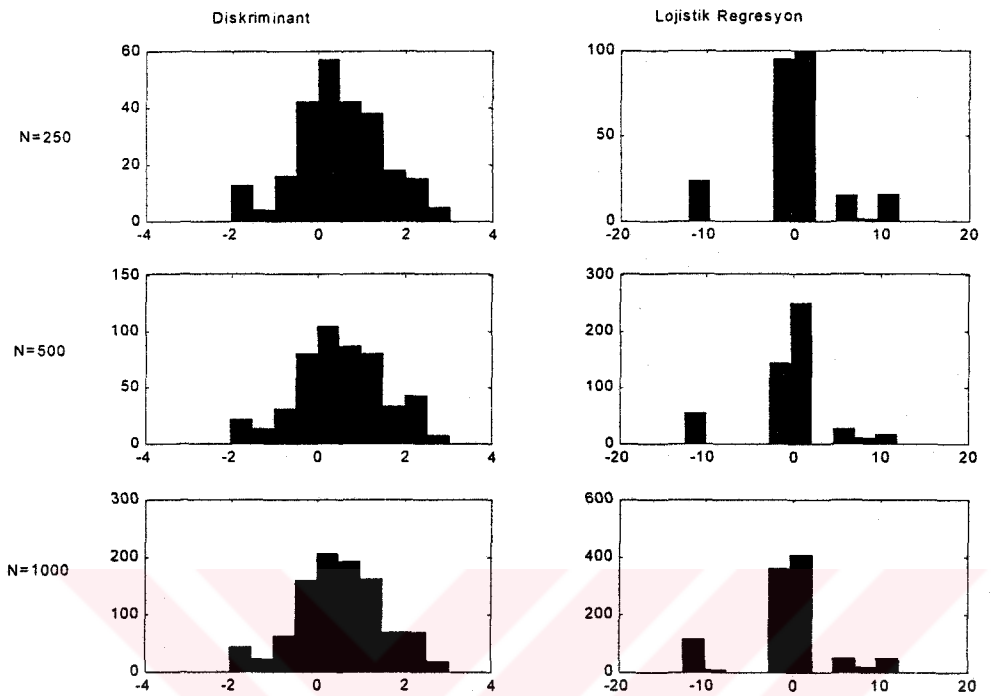


Şekil 22. Eşyasız Evde Kiracıların Yeniden Örnekleme Tahmin Değerlerinin Dağılımı.

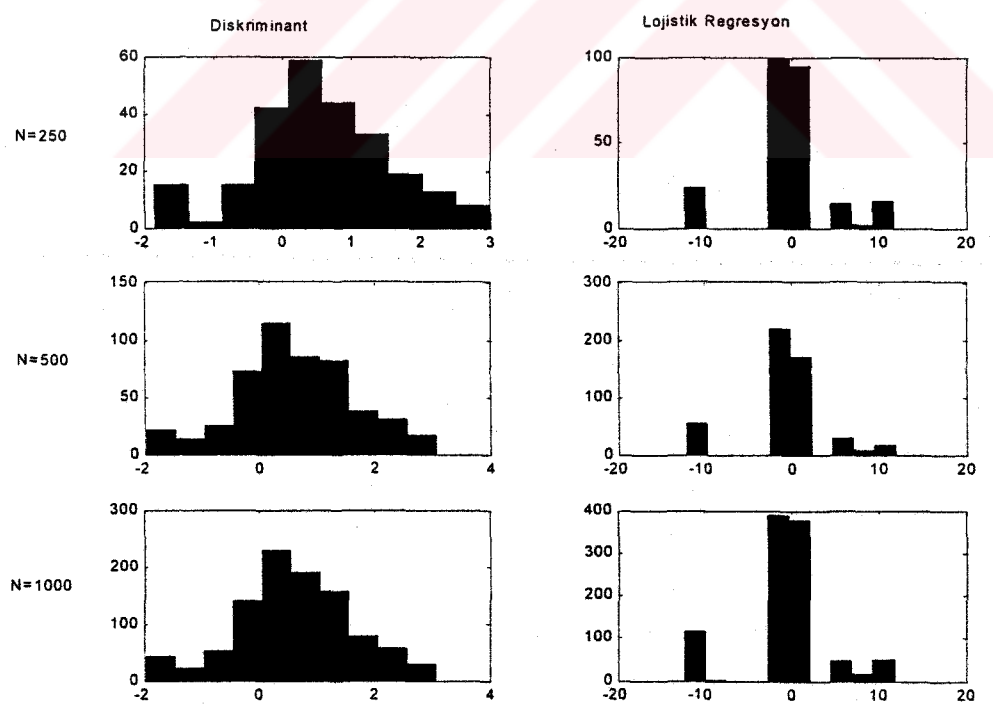


Şekil 23. Ailesi ile Kalanların Yeniden Örnekleme Tahminlerinin Dağılımı.

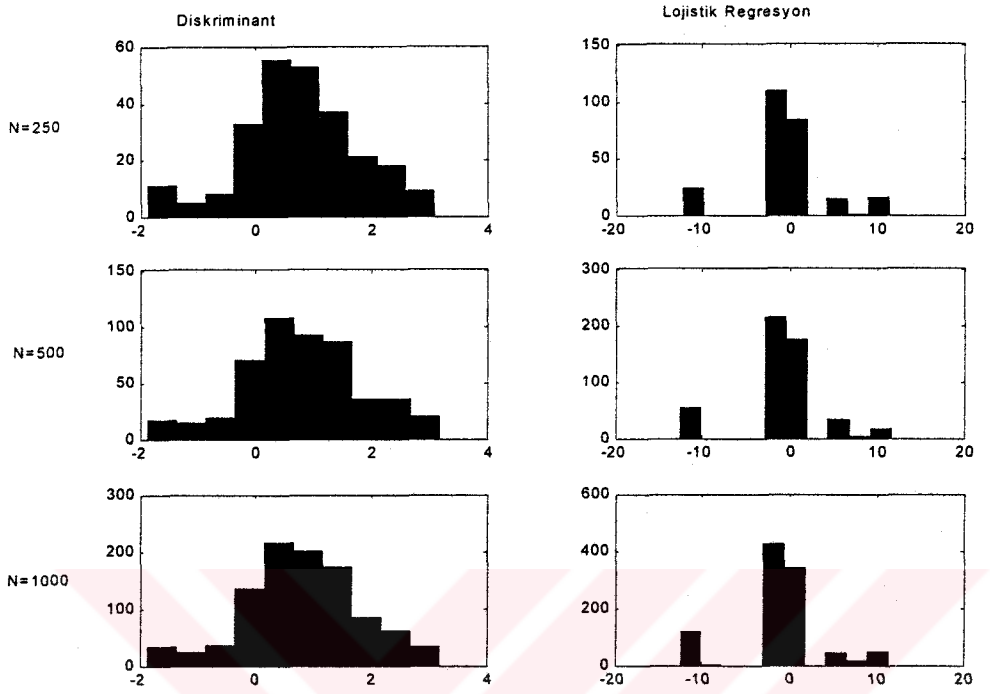
Ek 2
İkinci Uygulamadaki Değişkenlerin Dağılımı.



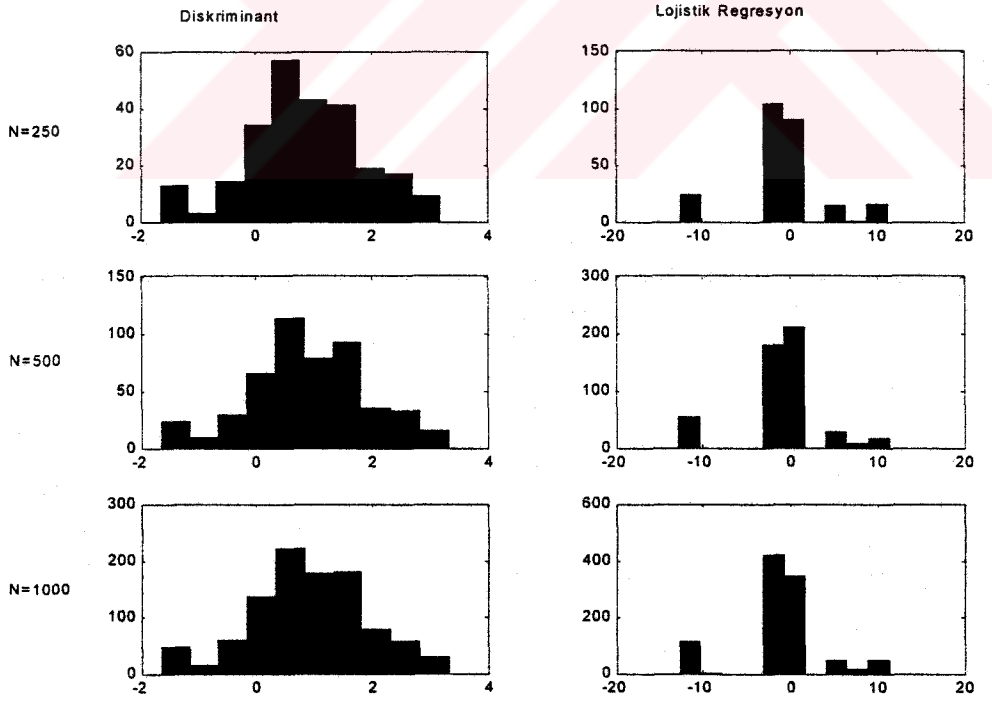
Şekil 24. 20-22 Yaş Grubu Yeniden Örnekleme Tahmin Sonuçlarının Dağılımı



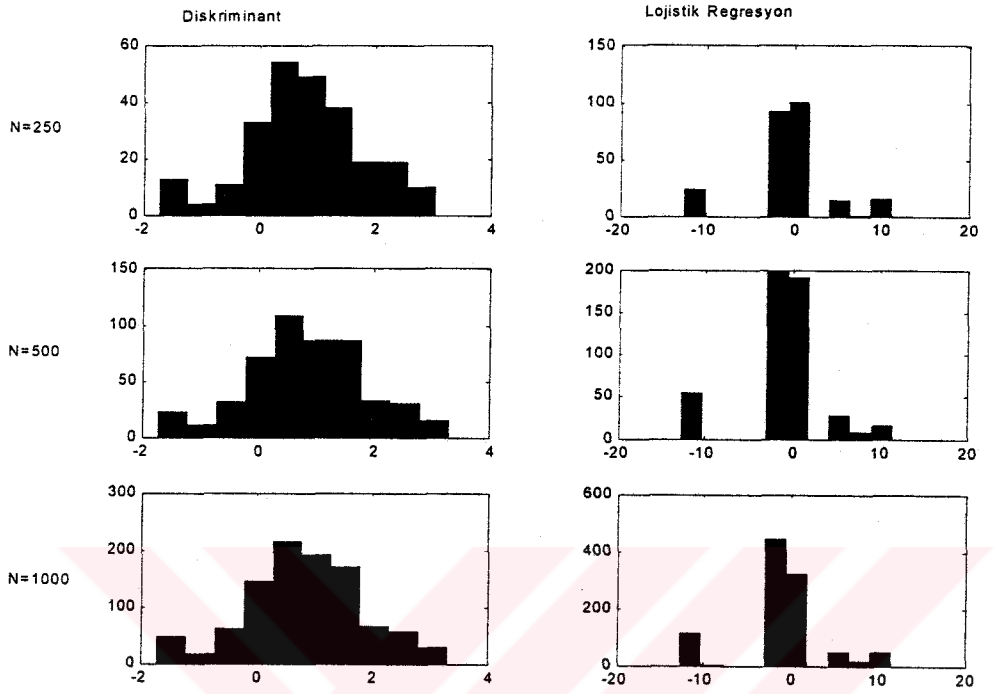
Şekil 25. 23-26 Yaş Grubu Yeniden Örnekleme Tahmin Sonuçları Dağılımı



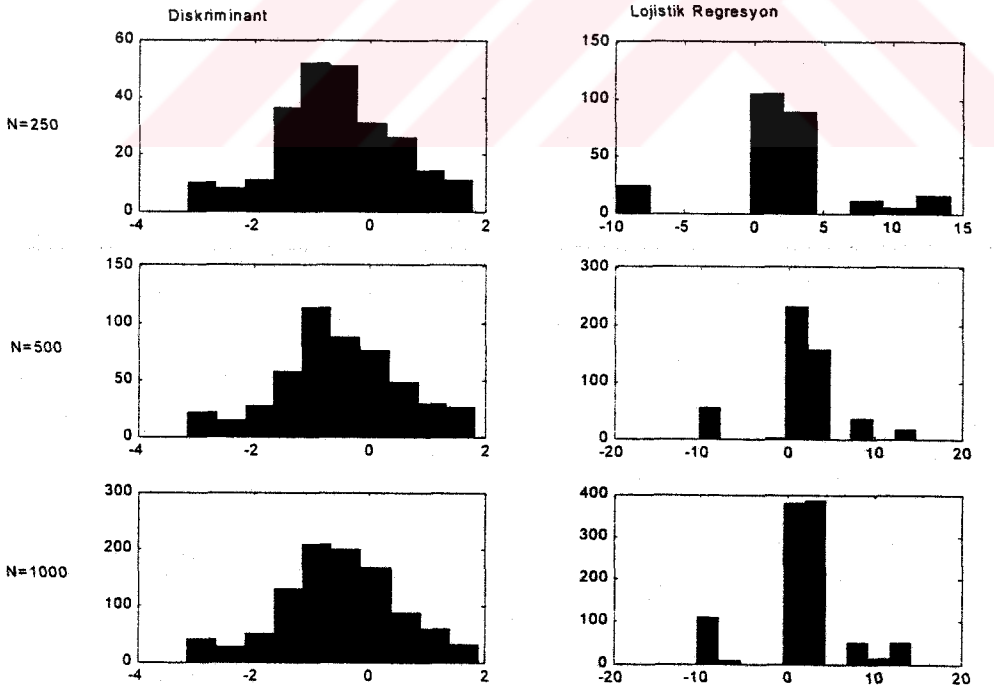
Şekil 26. 26-31 Yaş Grubu Yeniden Örnekleme Tahmin Sonuçları Dağılımı



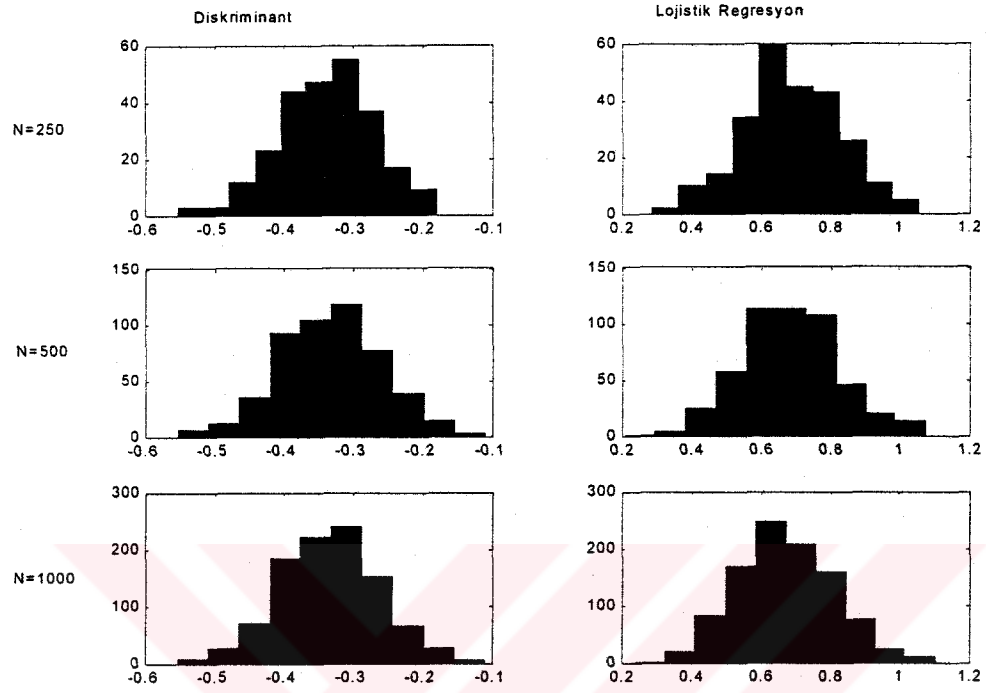
Şekil 27. 32-39 Yaş Grubu Yeniden Örnekleme Tahmin Sonuçlarının Dağılımı



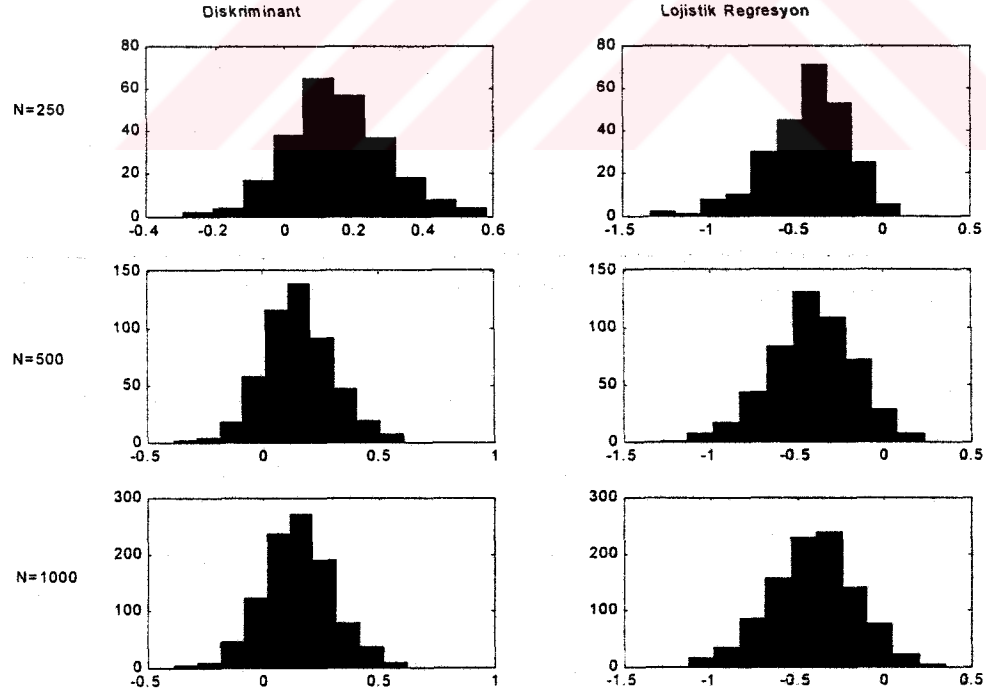
Şekil 28. 40-45 Yaş Grubu Yeniden Örnekleme Tahmin Sonuçları Dağılımı



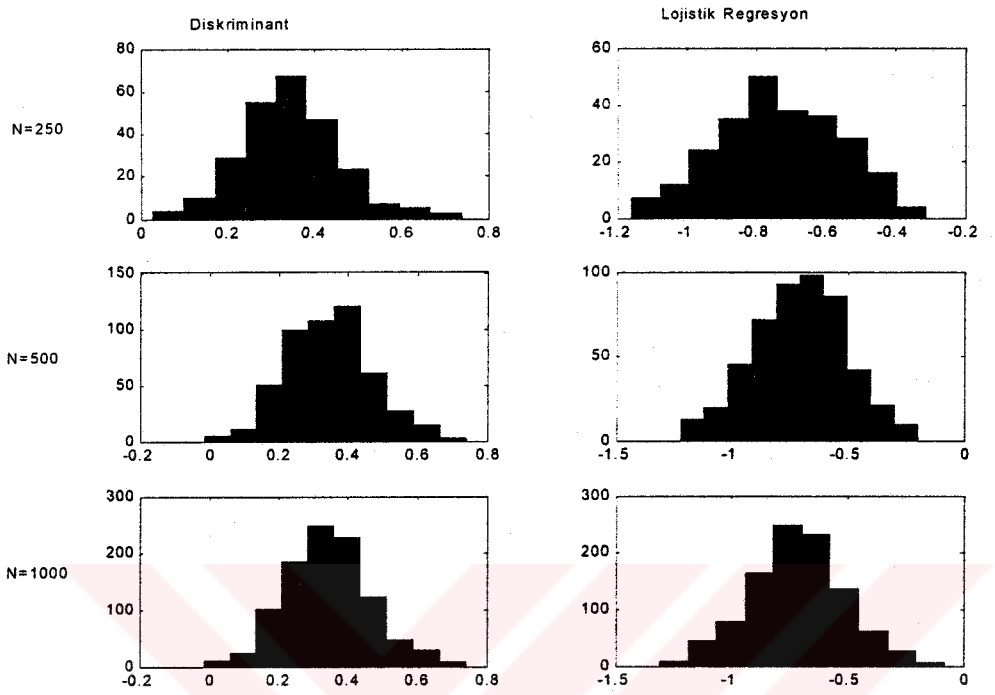
Şekil 29. 46 Yaş Üstü Yeniden Örnekleme Tahmin Sonuçlarının Dağılımı



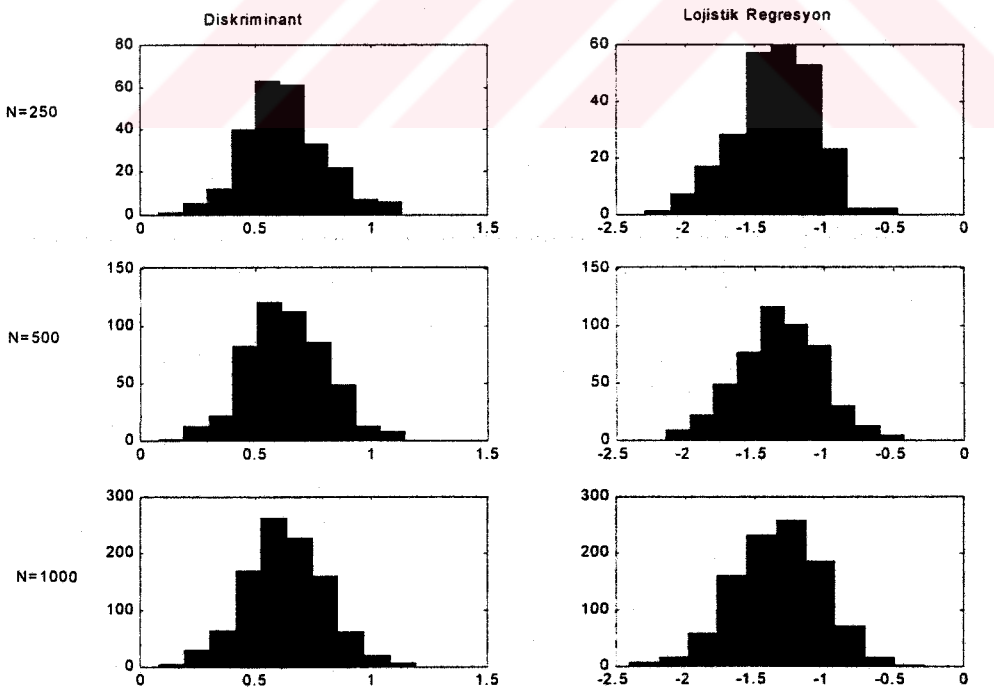
Şekil 30. Erkeklerin Yeniden Örnekleme Tahmin Sonuçlarının Dağılımı



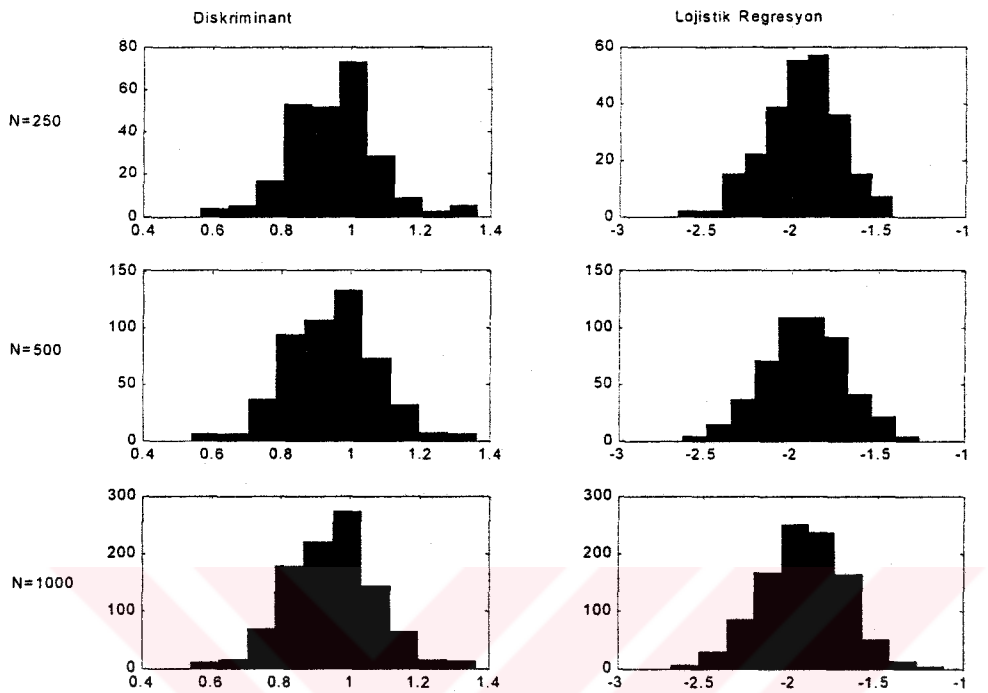
Şekil 31. İlkokul Mezunu Olanların Yeniden Örnekleme Tahmin Sonuçlarının Dağılımı



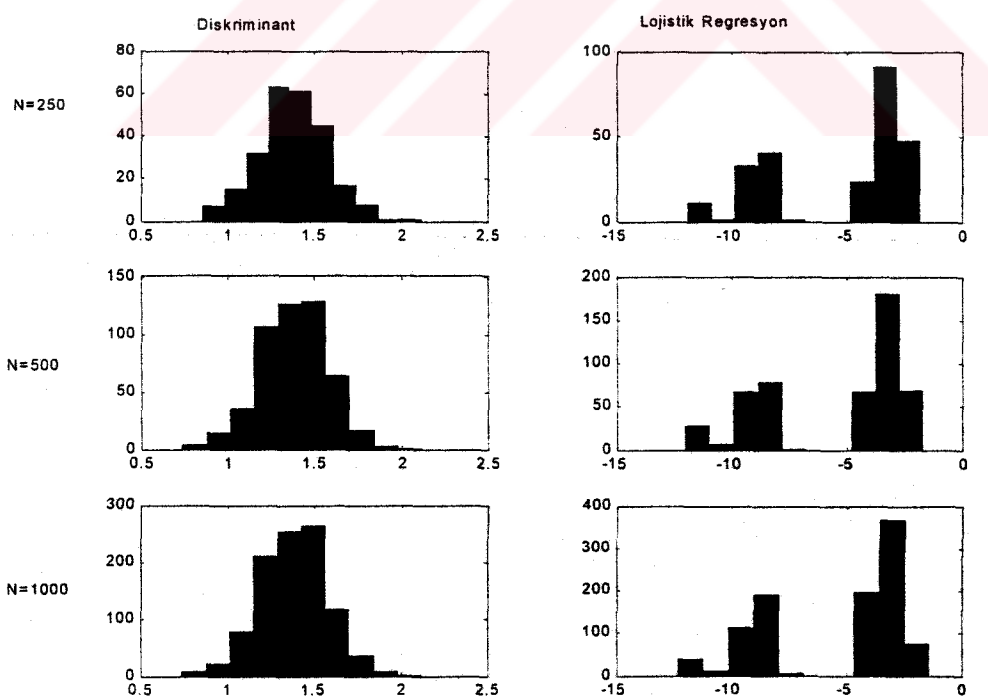
Şekil 32. Ortaokul Mezunlarının Yeniden Örnekleme Tahmin Sonuçlarının Dağılımı



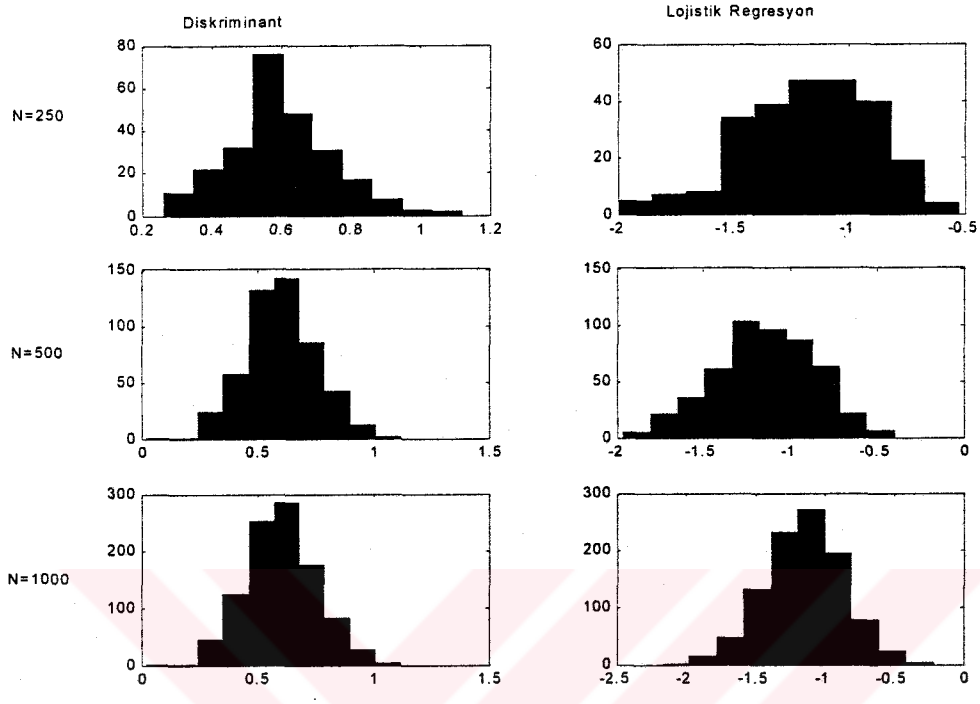
Şekil 33. Lise Mezunlarının Yeniden Örnekleme Tahmin Sonuçlarının Dağılımı



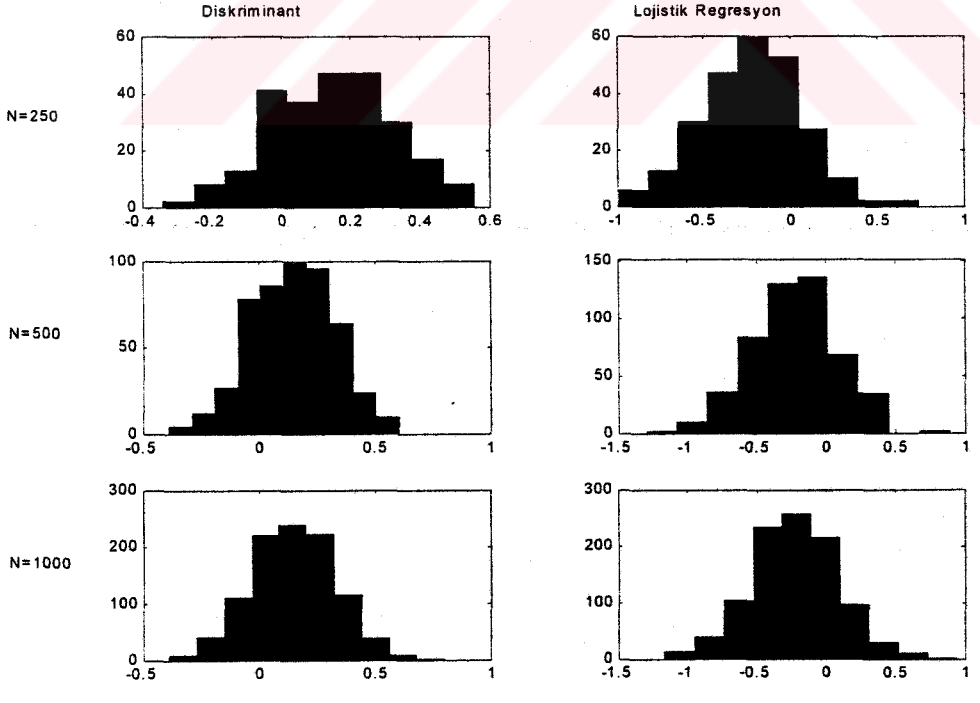
Şekil 34. 2 Yıllık Üni. Mezununlarının Yeniden Örnekleme Tahmin Sonuçları Dağılımı



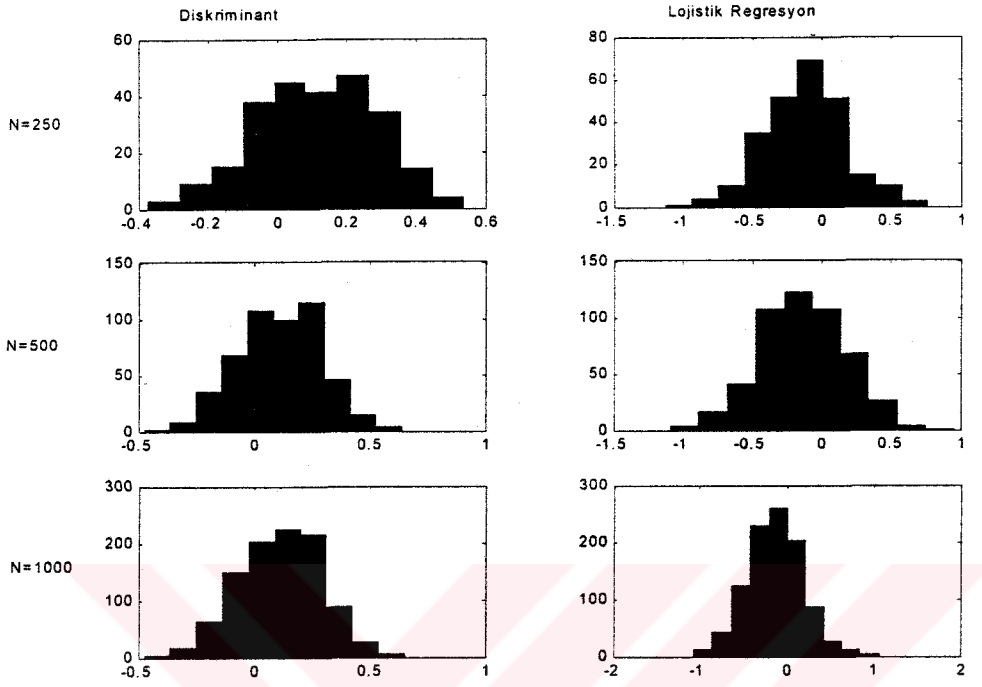
Şekil 35. 4 Yıllık Üniversite Mezunlarının Yeniden Örnekleme Tahmin Sonuçlarının Dağılımı



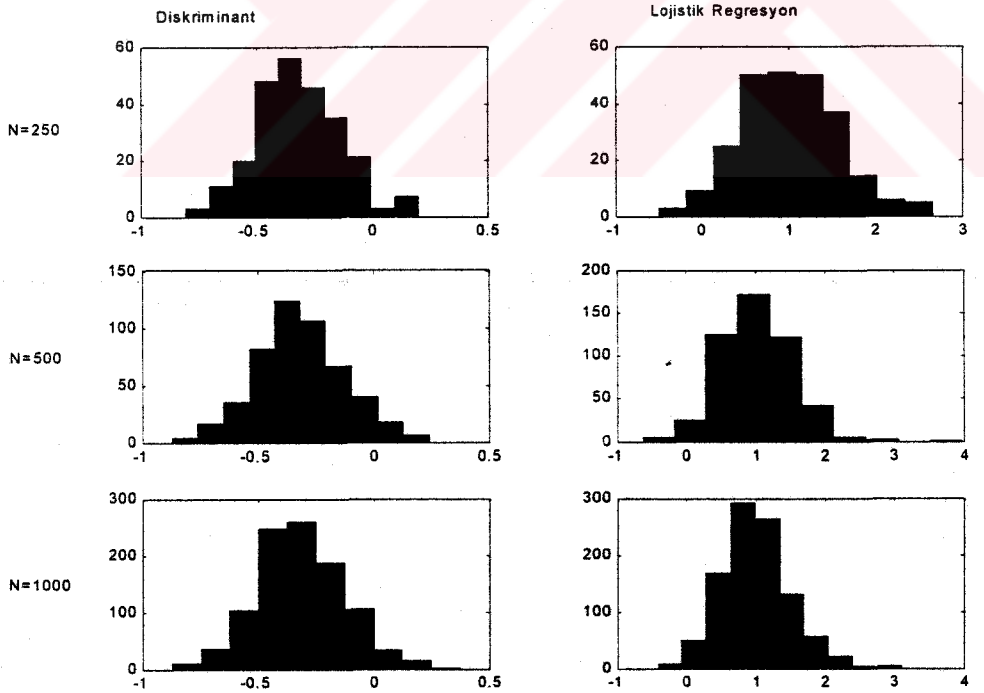
Şekil 36. Mastır/Doktorası Olanların Yeniden Örnekleme Tahmin Sonuçlarının Dağılımı



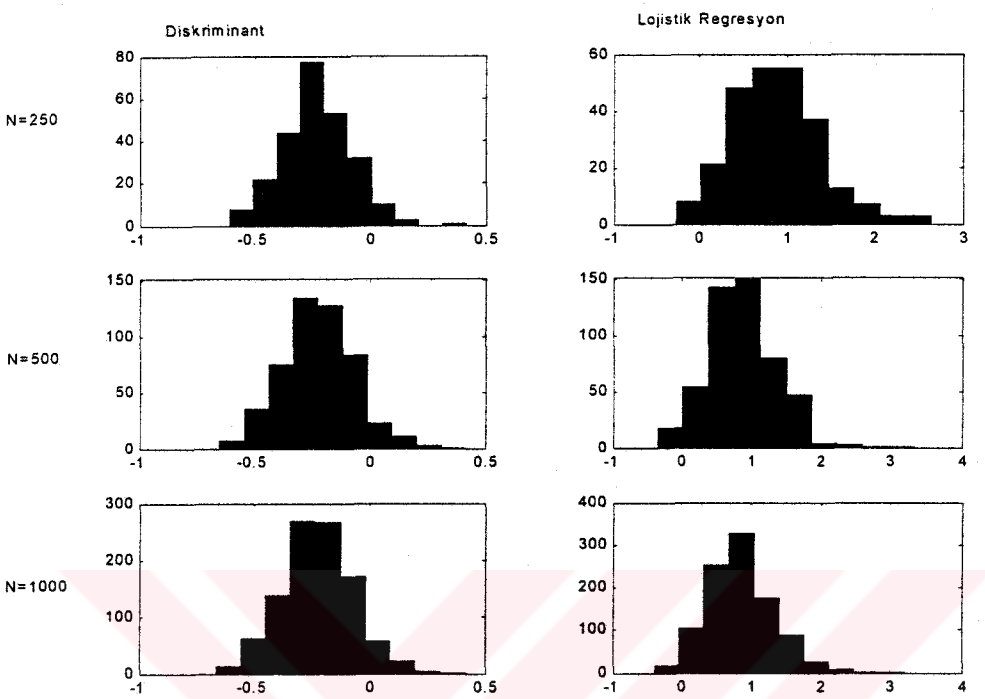
Şekil 37. Evlilerin Yeniden Örnekleme Tahmin Sonuçlarının Dağılımı



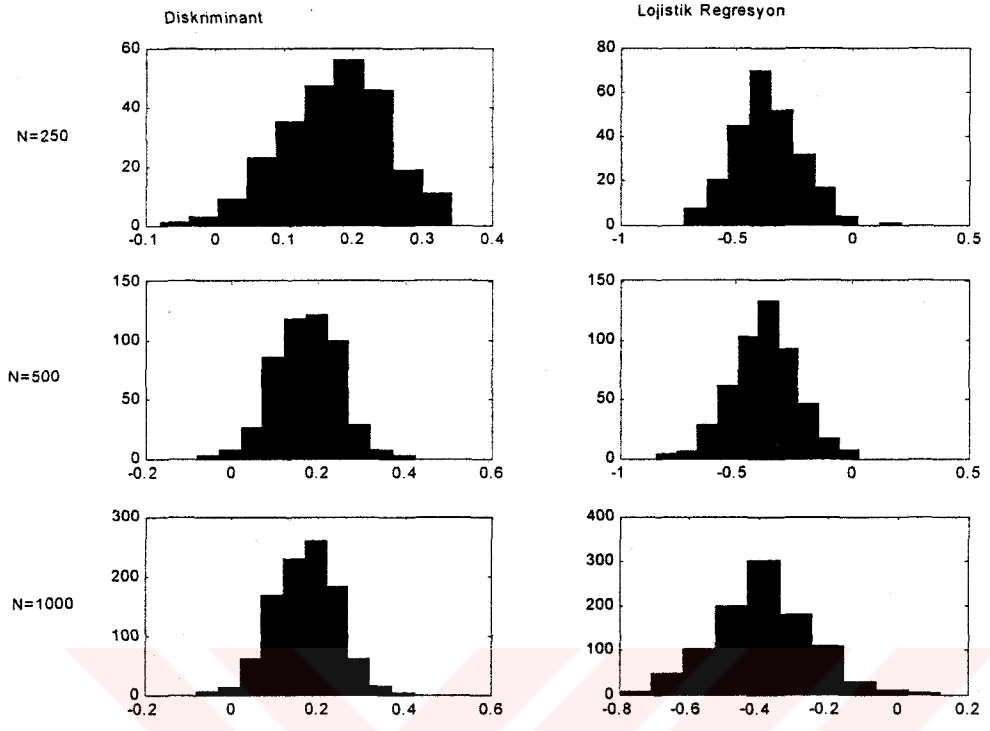
Şekil 38. Bekarların Yeniden Örnekleme Tahmin Sonuçlarının Dağılımı



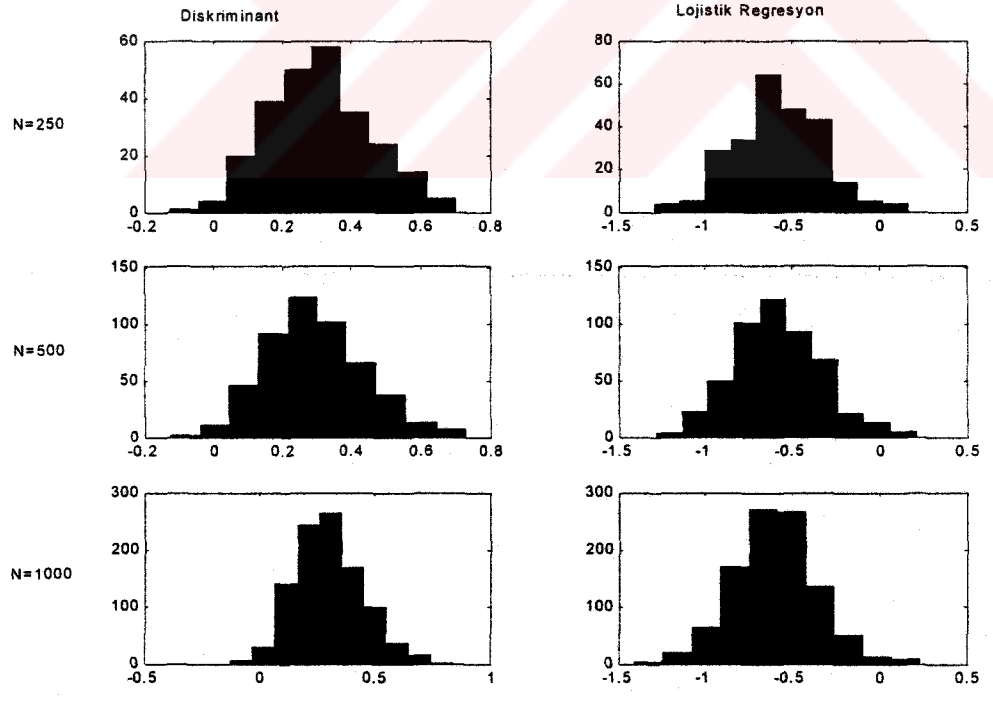
Şekil 39. İş Tel. Verenlerin Yeniden Örnekleme Tahmin Sonuçlarının Dağılımı.



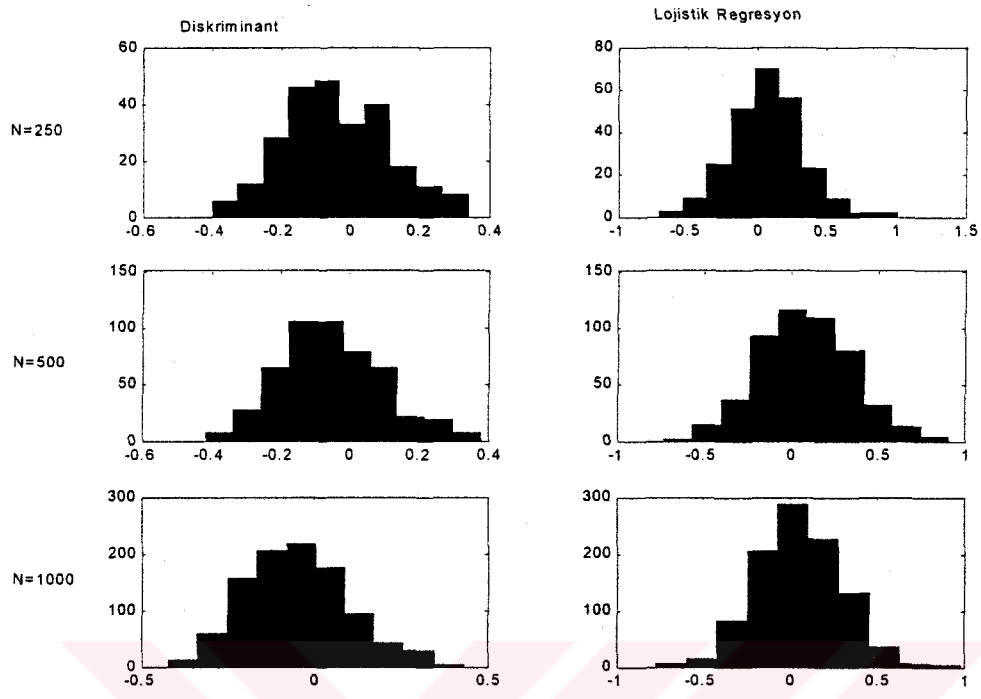
Şekil 40. İş ve Ev Tel. Verenlerin Yeniden Örnekleme Tahmin Sonuçlarının Dağılımı



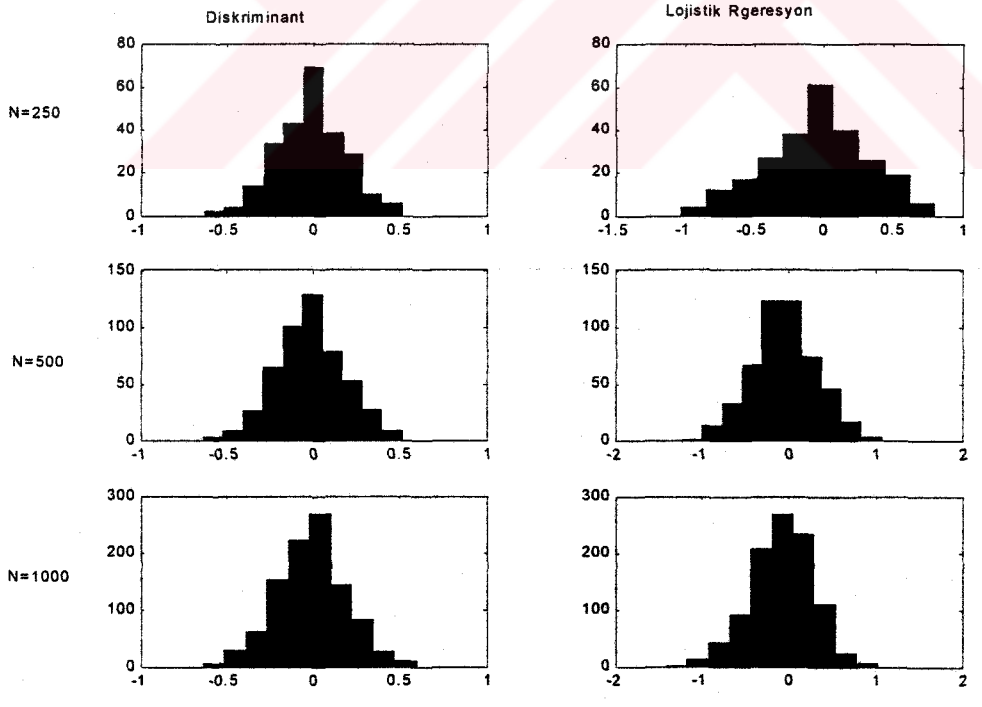
Şekil 41. Cep Tel. Verenlerin Yeniden Örnekleme tahmin Sonuçlarının Dağılımı



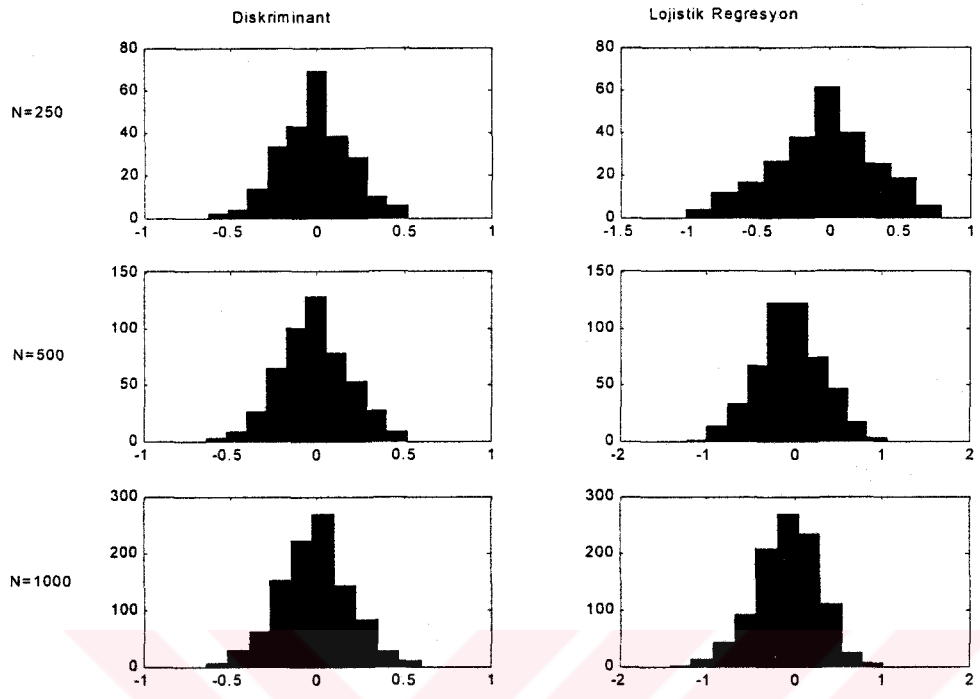
Şekil 42. Ev Sahibi Olanların Yeniden Örnekleme Tahmin Sonuçlarının Dağılımı



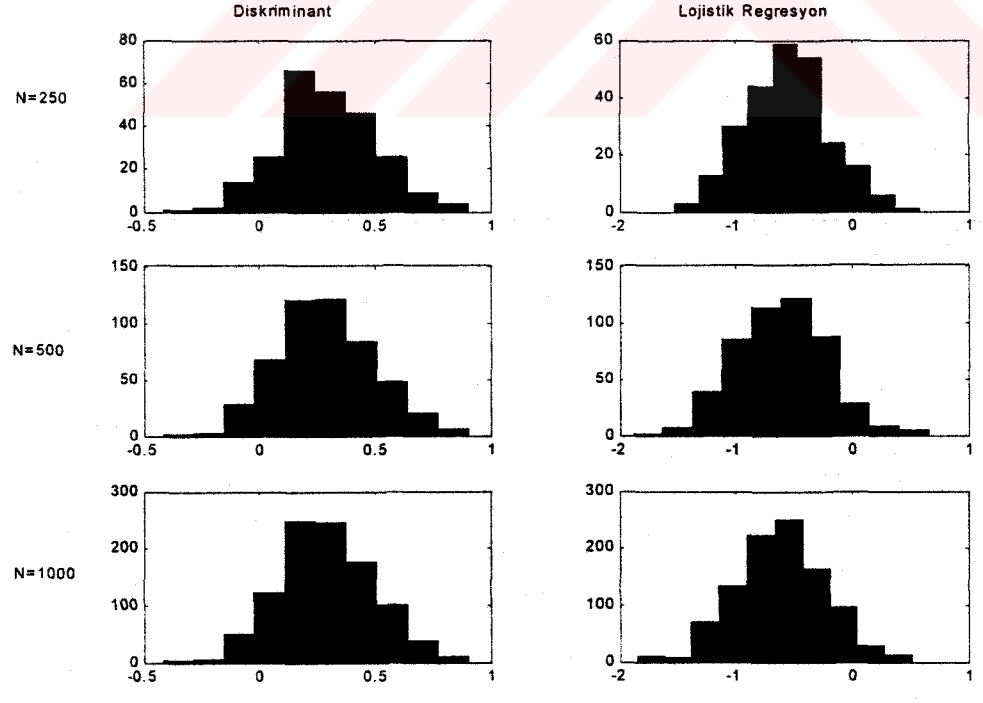
Şekil 43. Kiracı Olanlar Yeniden Örnekleme Tahmin Sonuçlarının Dağılımı



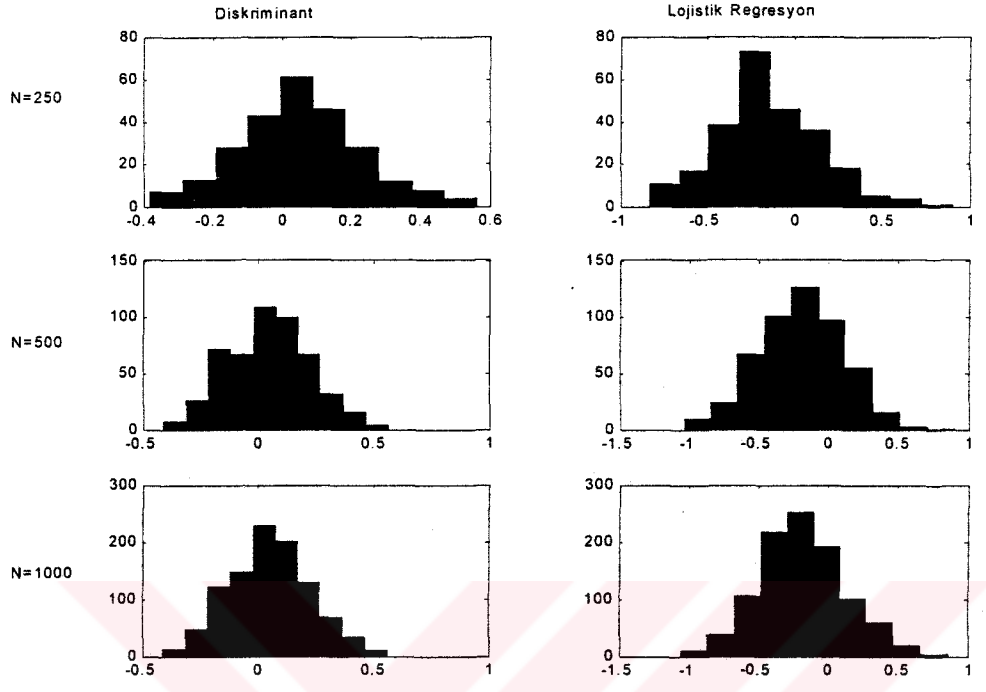
Şekil 44. Şirket Evinde Oturanlar Yeniden Örnekleme Tahmin Sonuçlarının Dağılımı



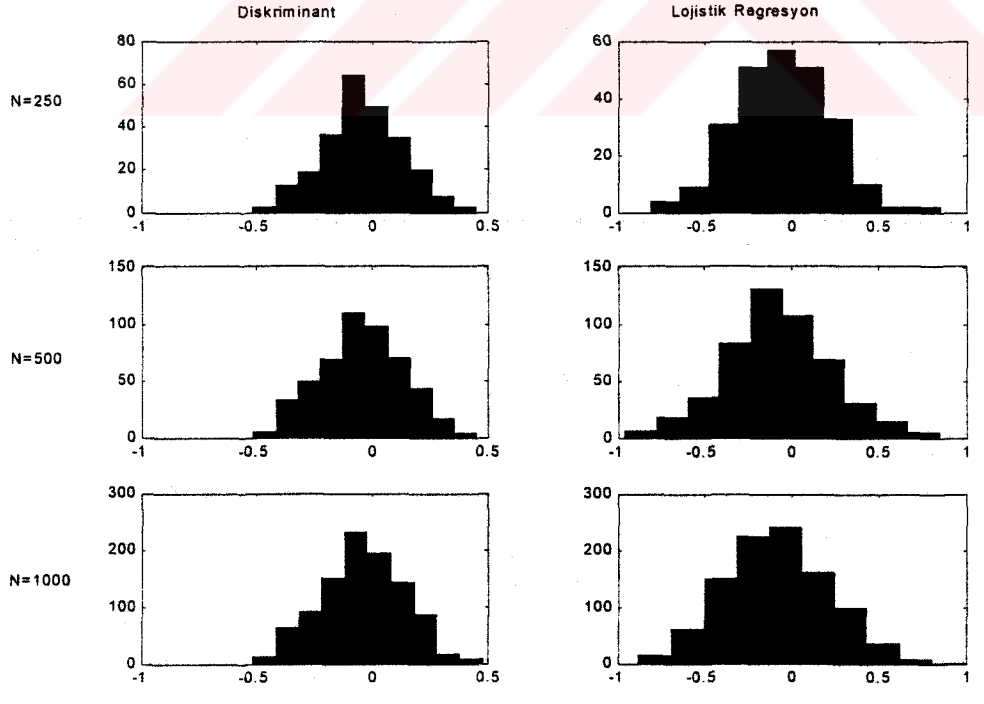
Şekil 45. Ailesinin Evi Olanlar Yeniden Örnekleme Tahmin Sonuçları Dağılımı



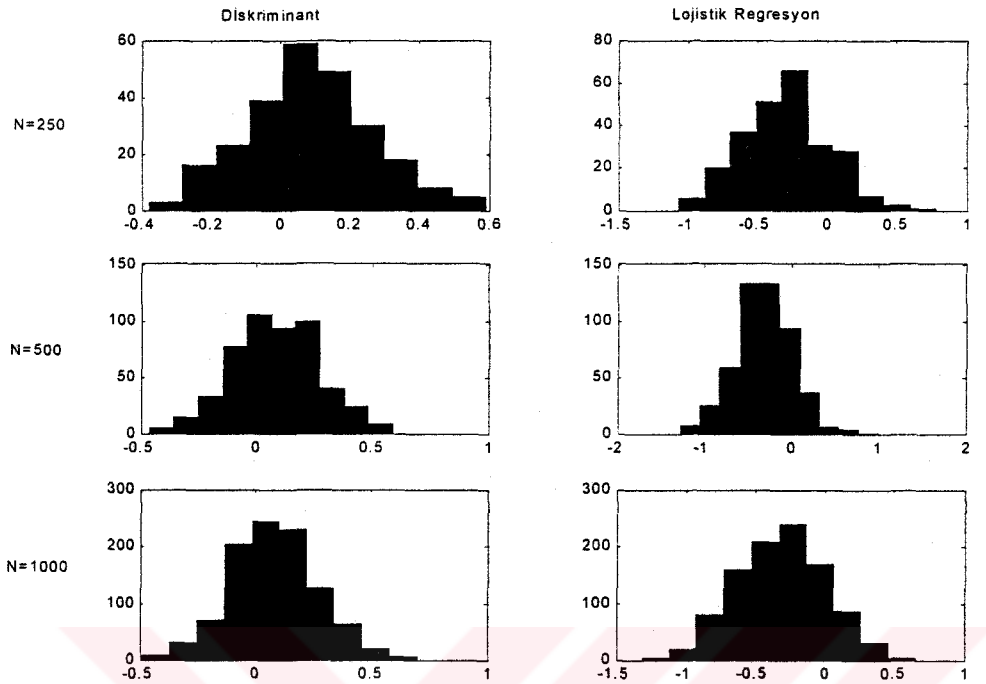
Şekil 46. 13-18 Ay Çalışanlar Yeniden Örnekleme Tahmin Sonuçları Dağılımı



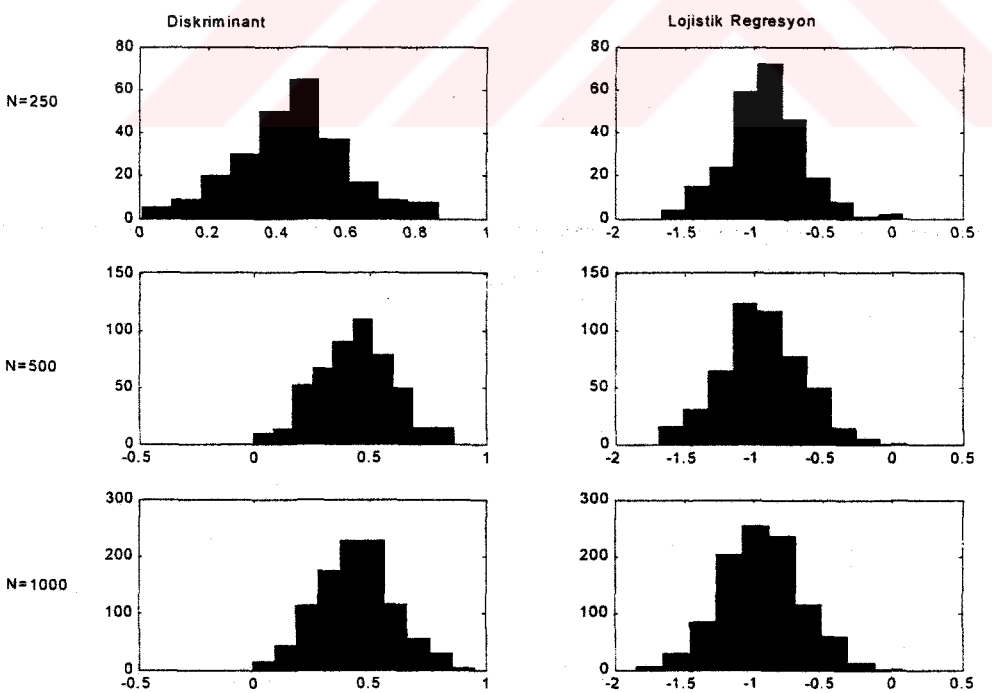
Şekil 47. 19-23 Ay Çalışsanlar Yeniden Örnekleme Tahmin Sonuçları Dağılımı



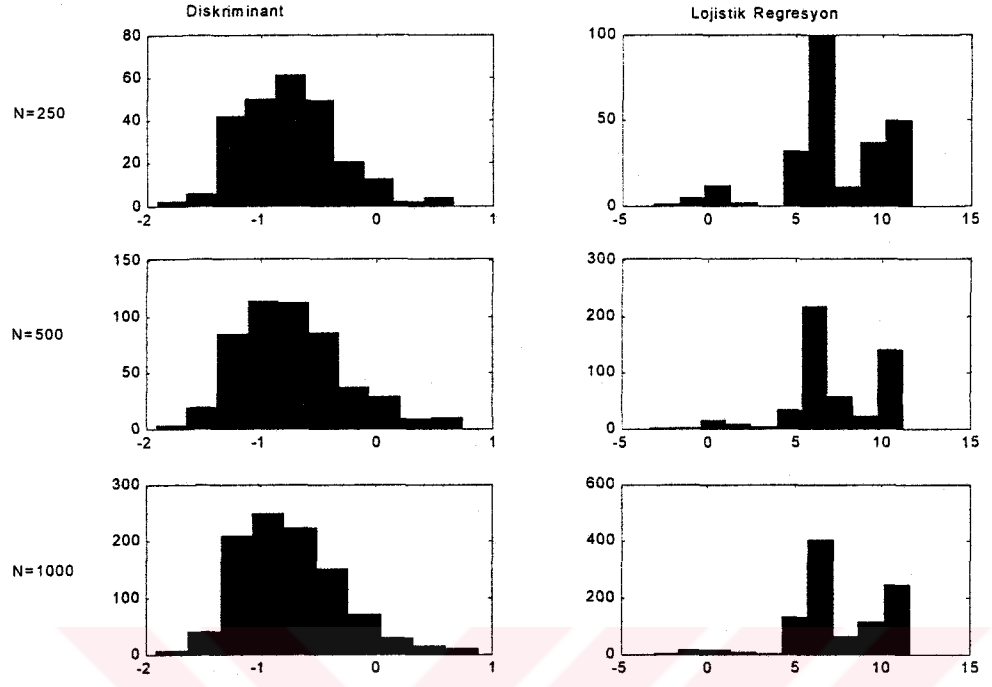
Şekil 48. 24-35 Ay Çalışanlar Yeniden Örnekleme Tahmin Sonuçları Dağılımı



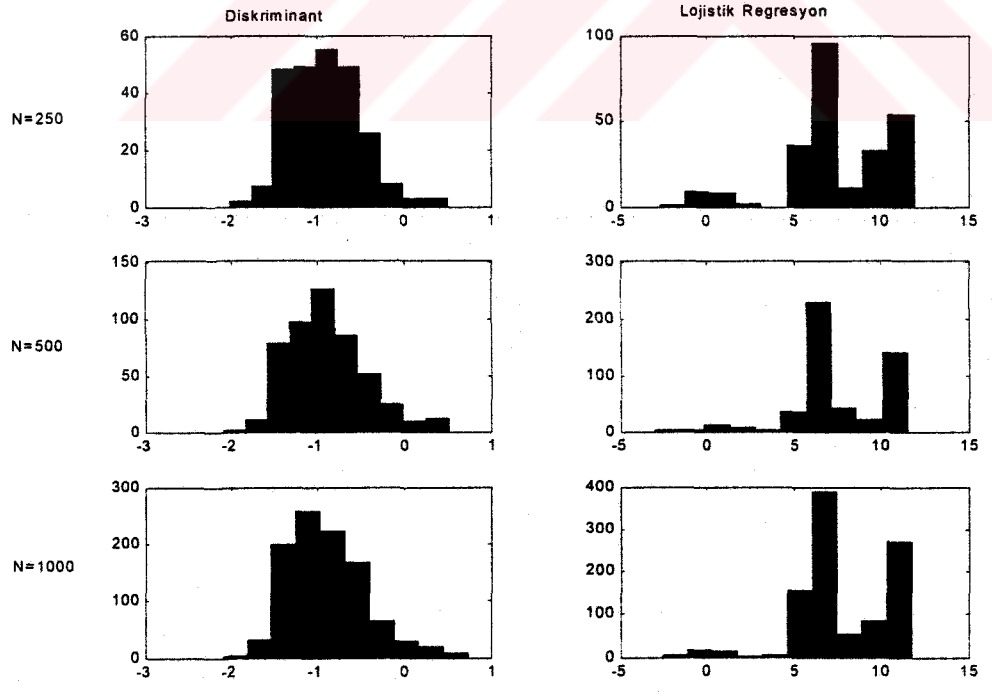
Şekil 49. 36-47 Ay Çalışnalar Yeniden Örnekleme Tahmin Sonuçalrı Dağılımı



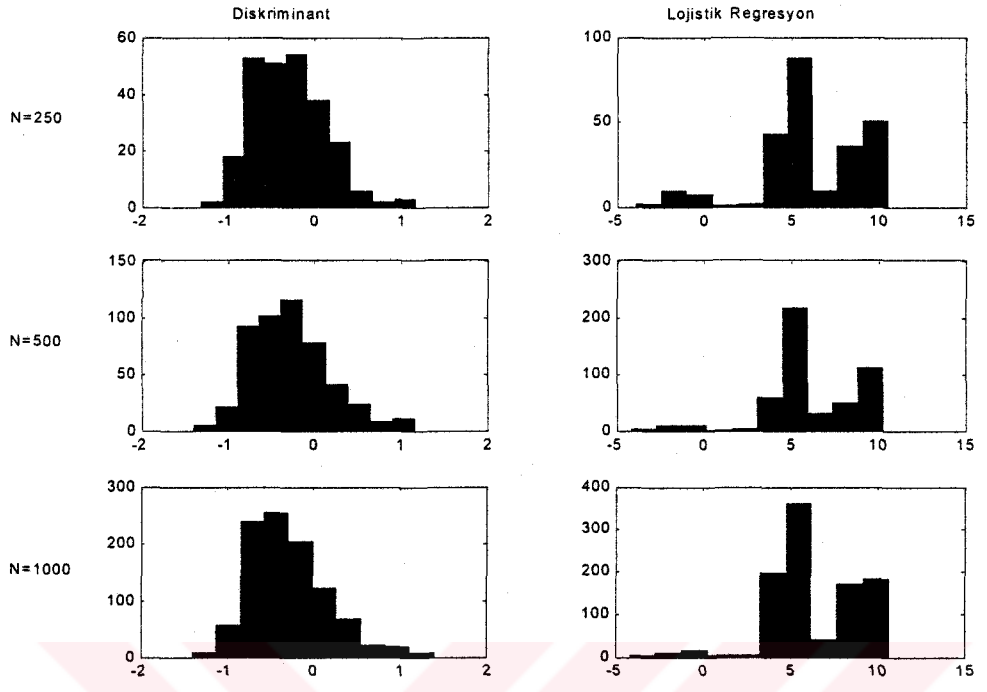
Şekil 50. 48 Aydan Fazla Çalışanlar Yeniden Örnekleme Tahmin Sonuçlarının Dağılımı



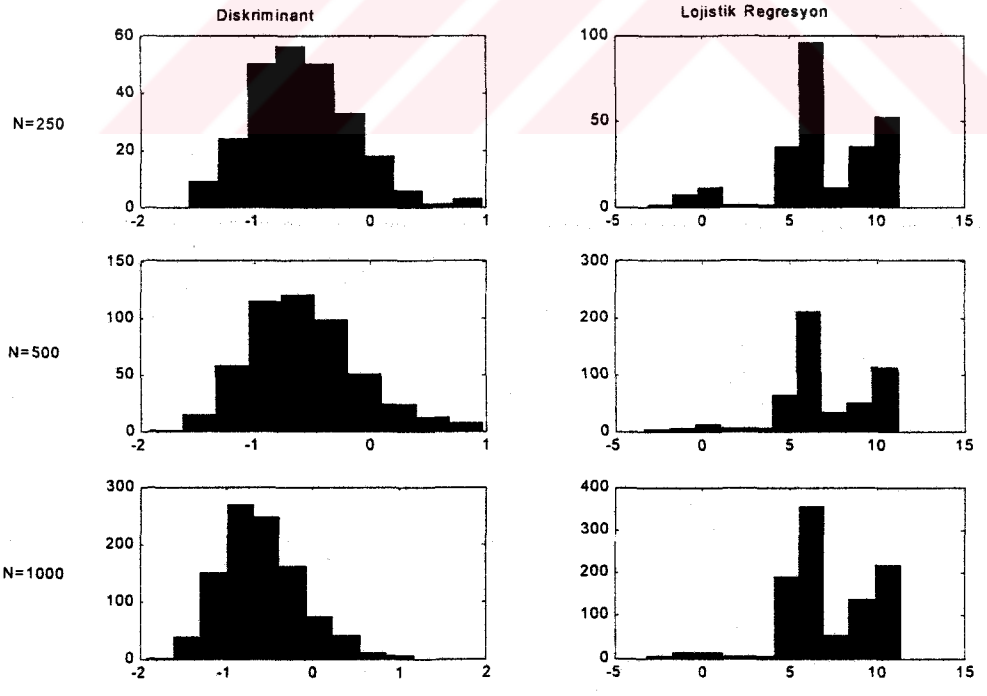
Şekil 51. Kamuda Çalışanlar Yeniden Örnekleme Tahmin Sonuçlarının Dağılımı



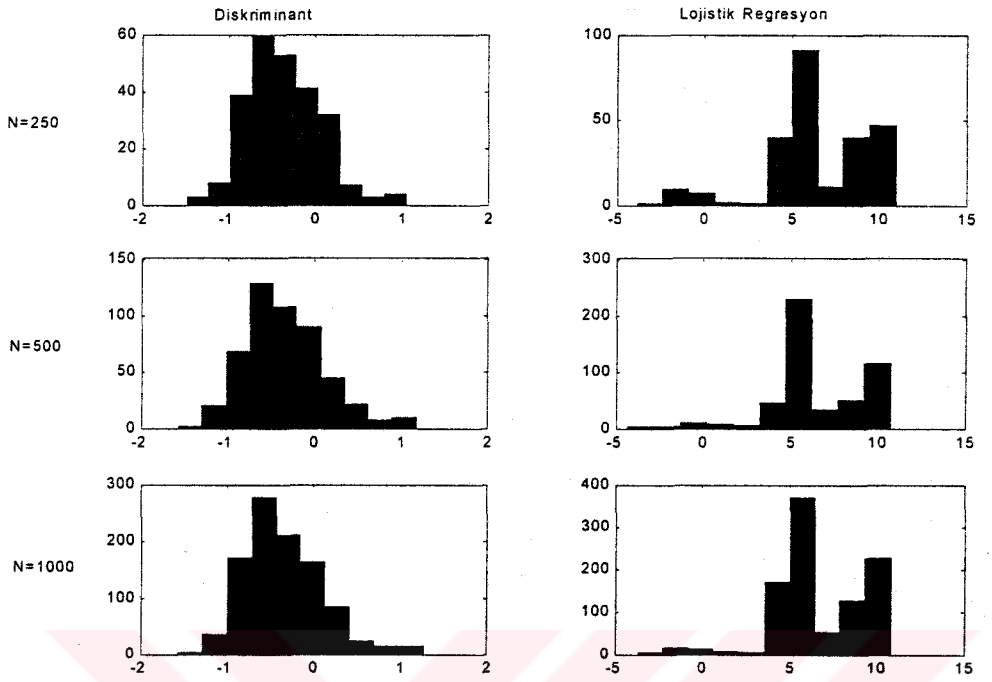
Şekil 52. Özel Sektörde Çalışanlar Yeniden Örnekleme Tahmin Sonuçlarının Dağılımı



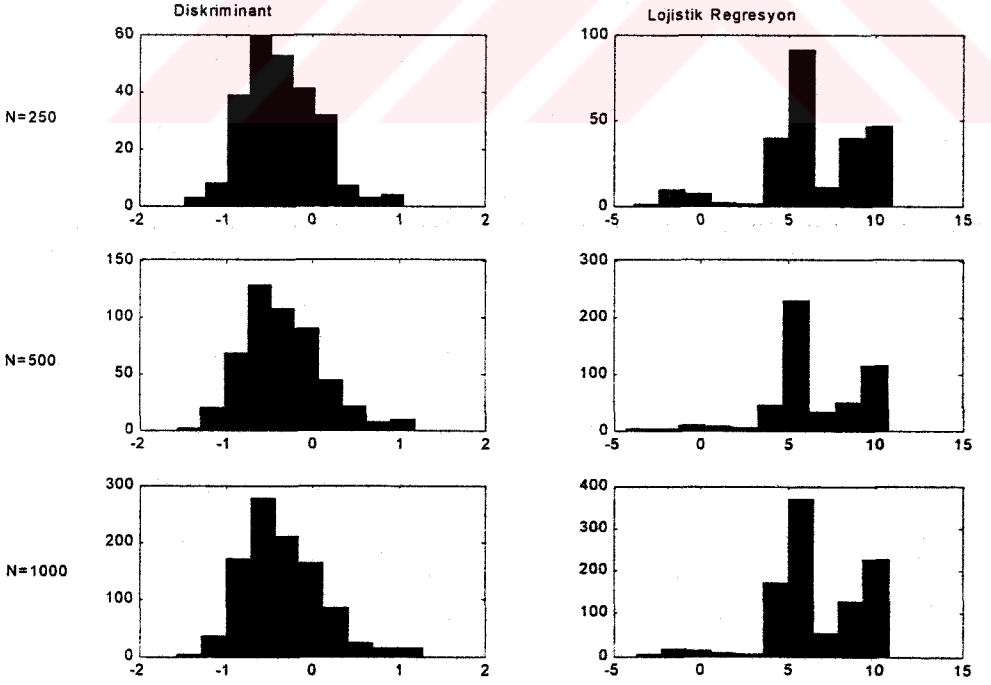
Şekil 53. Şirket Ortakları Yeniden Örnekleme Tahmin Sonuçları Dağılımı



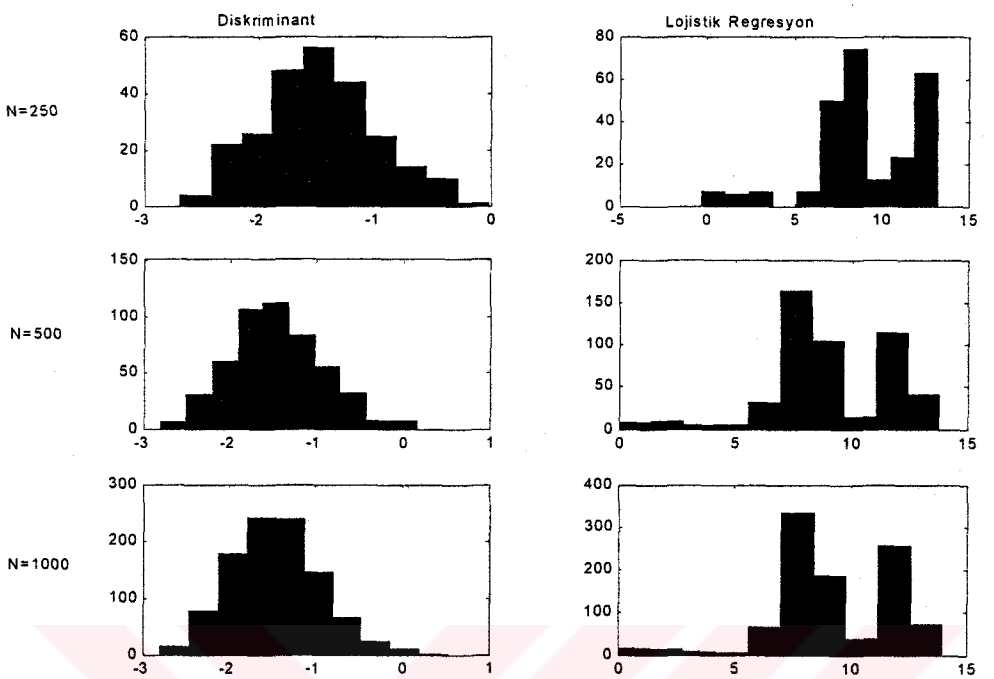
Şekil 54. Küçük İşletme Sahiplerinin Yeniden Örnekleme Tahmin Sonuçlarının Dağılımı



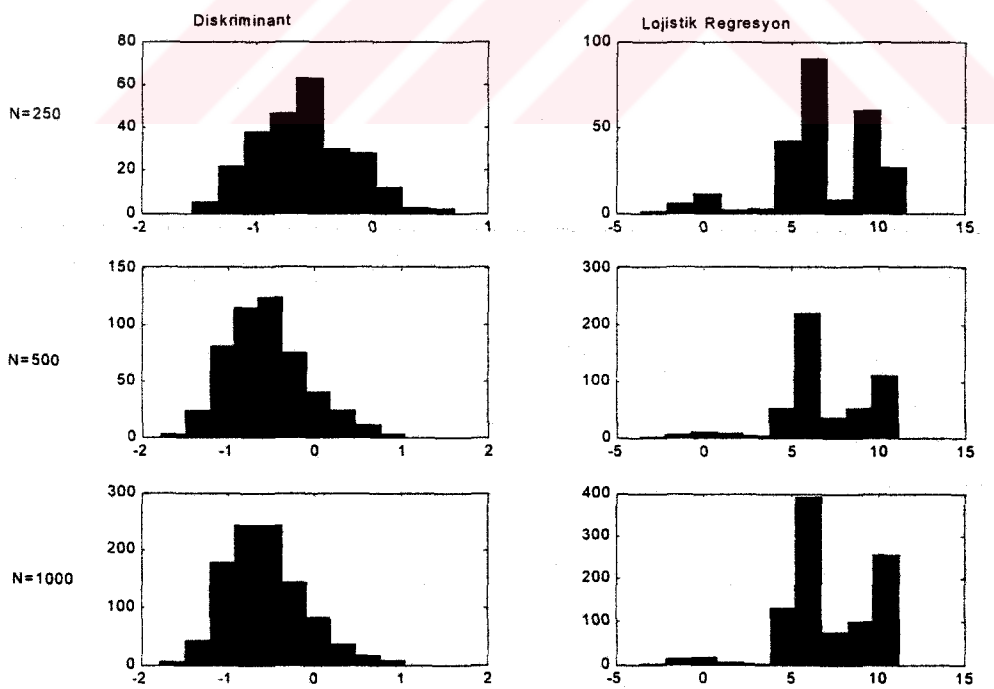
Şekil 55. Esnafların Yeniden Örnekleme Tahmin Sonuçlarının Dağılımı



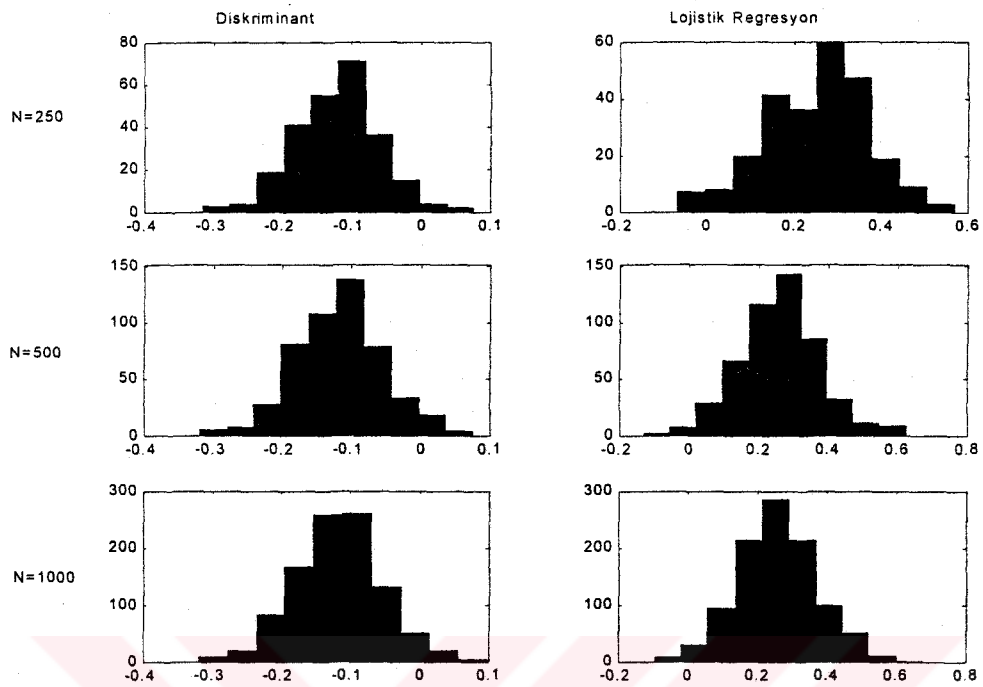
Şekil 56. Emeklilerin Yeniden Örnekleme Tahmin Sonuçlarının Dağılımı



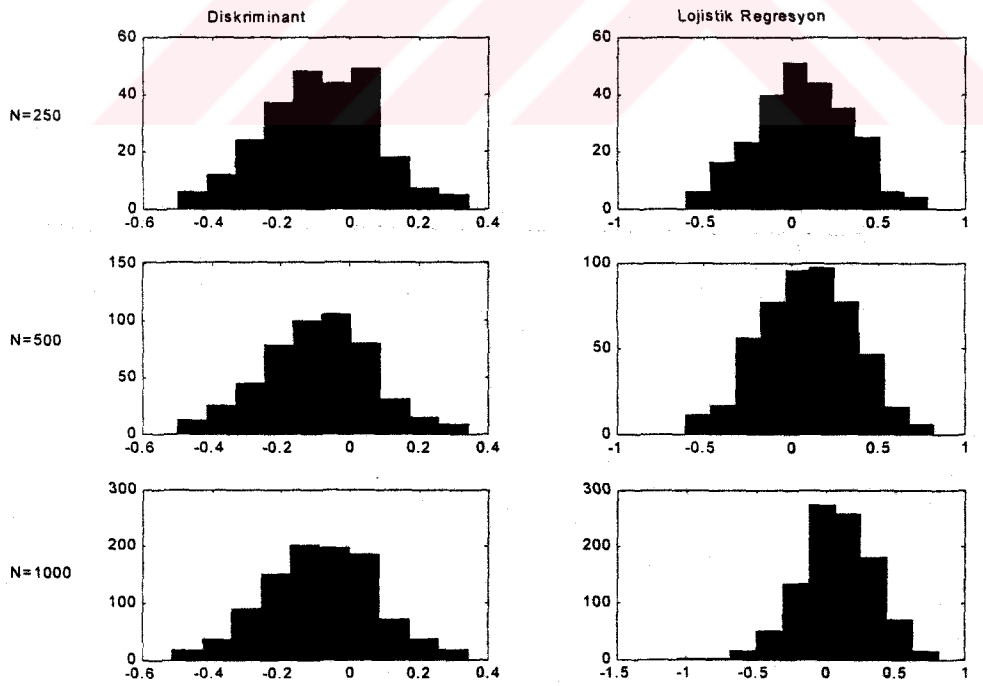
Şekil 57. Ev Kadınlarının Yeniden Örnekleme Tahmin Sonuçlarının Dağılımı



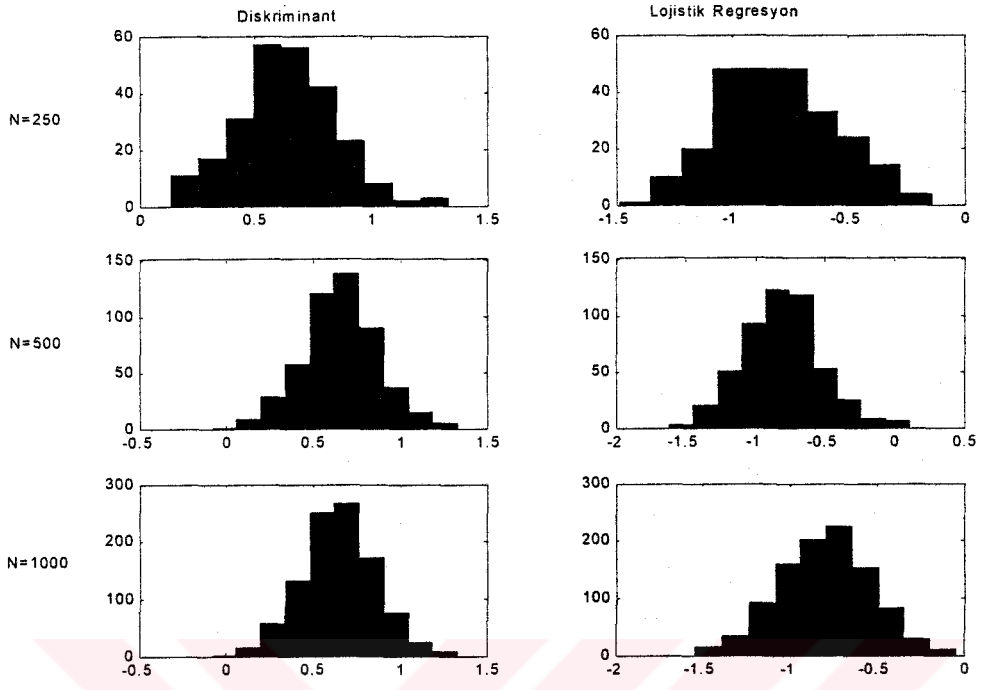
Şekil 58. Öğrencilerin Yeniden Örnekleme Tahmin Sonuçlarının Dağılımı



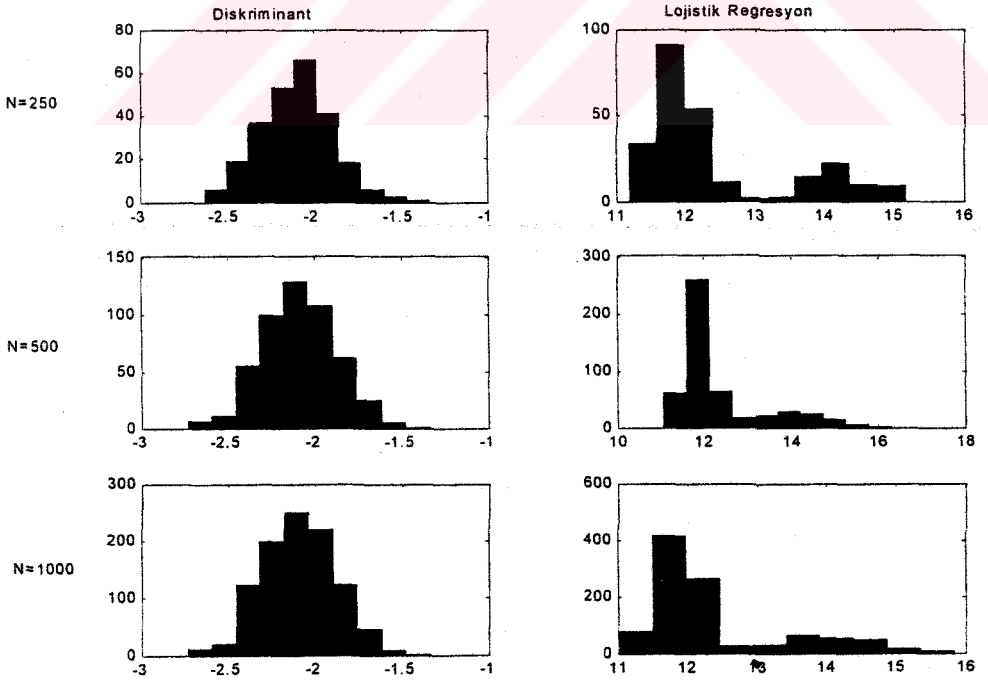
Şekil 59. Ekstresine Ev Adresine İsteyenlerin Yeniden Örnekleme Sonuçlarının Dağılımı



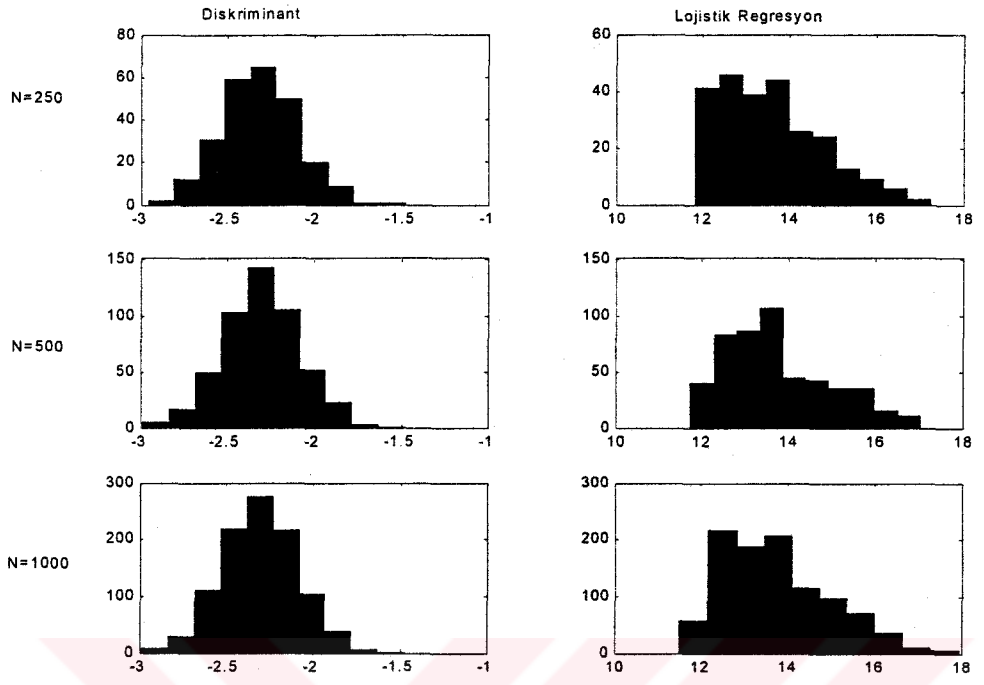
Şekil 60. Ek Kart İSteyenlerin Yeniden Örnekleme Tahmin Sonuçlarının Dağılımı



Şekil 61. Otomatik Ödemesi Olanların Yeniden Örnekleme Sonuçları Dağılımı



Şekil 62. Otomatik Emri Olup Seçim Yapmayanlar Yeniden Örnekleme Sonuçları Dağılımı



Şekil 63. Otomatik Ödeme Emri Olmayanlar Yeniden Örnekleme Sonuçları Dağılımı

KAYNAKLAR

Albright H.T., 1994, "Construction of a polynomial classifier for consumer loan applications using genetic algorithms" **Working Papers**, Department of Systems Engineering, University of Virginia.

Aldrich, J. H. , Nelson, D. F, **"Linear Probability, Logit, and Probit models",** Sage Pub. 1984

Altman E. I., "Financial ratios discriminant analysis and the prediction of corporate bankruptcy" **Journal of Finance**. 1968, 589-609

Altman E. I., Haldeman R., Narayanan P, "Zeta analysis: a new model to identify bankruptcy risk of corporations" **Journal of Banking and Finance** 1977, 29-54

Altman E. I., Saunders A. "Credit risk measurement: development over the last 20 years", **Journal of Banking and Finance**. 21, 1998, 1721-1742

Altman E.I., Marco G., Varetto F., "Corporate distress diagnosis: comparison using linear disriminant analysis and nueral networks (the Italian experience)" **Journal of Banking and Finance**, 18, 1994,505-529.

Apilado V.P., Warner D.C., Dauten J.J., "Evaluation techniques in consumer finance" **Journal of Financial and Quantitative Analysis**, March, 1974, 275-283

Bierman H., Hausman W.H., "The credit granting decision" **Management Science**,16, 8, 1970, 519-532

Boyes W.J., Hoffman D.L., Low S.A. "An econometric analysis of the bank credit scoring problem" **Journal of Econometrics** 40, 1989, 3-14

Boyle M., Crook J. N., Hamilton R., Thomas L.C., "Methods for credit scoring applied to slow payers in credit scoring and credit control" ed. L. C. Thomas, D.B. Edelman, Oxford University Press, Oxford, 1992, 75-90

Breiman L., Friedman J.H., Olshen R. A., Stone C.J., **"Classification and Regression Trees"** Wadsworth, Belmont, California. 1984,

Bridges S., Disney R., 2001, "Modelling consumer credit and default: The research agenda" Experian Centre for Economic Modelling, www.nottingham.ac.uk/economics/ExCEM.

Bugera V., Konno H., Uryasev S., "Credit cards scoring with quadratic utulity function" Research Report, University of Florida 2002-1

Capon N., "Credit scoring systems: a critical analysis" **Journal of Marketing**, 46, 1982, 82-91.

Carter C., Catlett J. "Assesing credit card applications using machine learning" **IEEE Expert** 2, 1987, 71-79

Chatterjee S. Barcun S., "A nonparametric approach to credit screening" **Journal of American Statistical Association**, March, 65, 1970,150-154

Cheng B., Titterington D.M., "Neural networks: a review from a statistical perspective" **Statistical Science** 9, 1994, 2-30

Cherney, M.R. **Bootstrap Methods: A Practitioner's Guide**, John Willey & Sons, Inc. 1999

Cizek, G.J, Fitzgerald, S. M.,"An introduction to logistic regression". **Measurement&Evaluation in Counseling&Development**, 31, 1999,223-245

Coffman J. Y., "The proper role of tree analysis in forecasting the risk behaviour of borrowers", MDS Reports, **Manangement Decision Systems**, Atlanta, 3,4,7 and 9. 1986,

Davis D.B., "Artificial intelligence goes to work" **High Technology** April, 1987, 16-17

Desai V.S., Convay D. G., Crock J. N., Overstreet G.A., "Credit scoring models in credit union environment using neural networks and genetic algorithms" **IMA Journal of Mathematics Applied in Business and Industry**, 8, 1997, 323-346.

Desai V.S., Crock J. N., Overstreet G.A., "A comparison of neural networks and linear scoring models in the credit environment" **European Journal of Operational Research**, 95, 1996, 24-37.

Durand D., "Risk elements in consumer instalment financing" **National Bureau of Economic Research**, New York. 1941

Efron and Tibshirani, **Introduction to the Bootstrap**, Chapman & Hall. 1993

Eisenbeis R.A., "Pitfalls in the application of discriminant analysis in business, finance and economics" **Journal of Finance**, 32, 1977,875-900.

Fisher R. A., "The use of multiple measurements in taxonomic problems" **Annals of Eugenics** 7, 1936, 179-188

Fitzpatrick D.B., "An analysis of bank credit card profit" **Journal of Bank Research**. 7, 1976, 199-205

Fogarty T.C., Ireson N.S., "Evolving bayesian classifiers for credit control-a comparison with other machine learning methods" **IMA Journal of Mathematics Applied in Business and Industry**, 5, 1993, 63-76.

Freed N., Glover F., "A linear programing approach to the discriminant problem" **Decision Science**, 12, 1981a, 68-74,

Freed N., Glover F., "Simple but powerful goal programming formulations for the discriminant problem" **European Journal of Operational Research**, 7, 1981b, 44-60

Glen J.J., "Integer programming models for normalisation and variable selection in mathematical programming models for discriminant analysis", **Procedings of Credit Scoring and Credit Control V**, Credit Research Center, University of Edinburgh. 1997

Goldberger, A. S., **Econometric Theory**. New York: John Wiley. 1964

Grablowsky B.J., Talley W.K., "Probit and discriminant function for classifications for classifying credit applicants a comparison" **Journal of Economic and Business**, 33, 1981, 254-261

Hair J. F., Anderson R. E., Totham R. L., Grablowisky B. J., **Multivariate Data Analysis With Readings**, Macmillan. 1984.

Hair J. F., Anderson R. E., Totham R. L., Grablowisky B. J., **Multivariate Data Analysis**, Printice Hall. 1998.

Hand D.J., "New instruments for identifying good and bad credit risk: a feasibly study" **Report Trustee Saving Bank**, London, 1986.

Hardy W.E., Adrian J. L., "A linear programming alternative to discriminant analysis in credit scoring" **Abribus** 1, 1985, 285-292.

Henley W.E., "Statistical aspects of credit scoring" PhD Thesis. The Open University, Milton Keynes, 1995.

Henley W.E., Hand D.J., "A k-nearest neighbour classifier for assessing consumer credit risk" **Statistician**, 45, 1996, 77-95

Hopper M. A., Lewis E.M., "Behavior Scoring and Adaptive Control Systems", **In Credit Scoring and Credit Control**, ed. By L.C. Thomas, J.N.Crook, D.B.Edelman. Oxford University Press, Oxford. 1992, pp. 257-276

JoachimsthalerE. W., Stam A., "Mathematical programming approaches for the classification problem in two-group discriminant analysis" **Multivariate Behavioural Research** 25, 1990, 427-454

Klecka W. R., **Discriminant Analysis.** Sage Publication. 1980

Kolesar P., Showers J.L., "A robust credit screening model using categorical data" **Management Science** 31(2), 1985, 123-133

Lachenbruch P.A., **Discriminant Analysis.** New York: Hafner, 1975.

Lane S., "Submarginal credit risk classification" **Journal of Financial and Quantitative Analysis**, January, 1972, 1379-1385

Leonard K.J., "Empirical bayes analysis of the commercial loan evaluation process" **Statistical Probability Letters**, 18, 1993a, 289-296

Leonard K.J., "Detecting credit fraud using expert systems" **Computer Industrial Engineer**, 25, 1993b, 103-106

Leonard K.J., "A fraud alert model for credit cards during the authorization process" **IMA Journal of Mathematical Applied Business Industry**, 5, 1993c, 57-62

Lucas A., "Updating scorecards: removing the mystique" **In Credit Scoring and Credit Control** (eds. L.C. Thomas, J.N. Crook, D.B. Edelman), Oxford: Claredon. 1992, pp 180-197,

Mahalanobis, P. C., "On the generalized distance in statistics", **Proceedings of the National Institute of Social Science**, India, 12, 1963; 49-55

Makowski P. "Credit scoring branches out" **The Credit World**, 75, 1985, 30-37

Mangasarian O.L., "Linear and nonlinear separation separation of patterns by linear programing" **Operations Research** 13, 1965, 444-452

McNab H., Wynn A., **Principles and Practice of consumer credit Risk Management**, CIB Publishing, Canterbury. 2000,

Mehta D., "The formulation of credit policy models" **Management Science** 15, 1968, 30-50

Menard, S. , **Applied Logistic Regression Analysis**, Sage. 1995

Moses D., Liao S.S., "On developing models for failure prediction" **Journal of Commercial Bank Lend**, 69, 1987, 27-38

Myers J.H., Forgy E. W., "The development of numerical credit evaluation systems" **Journal of American Statistical Association** 58, 1963 799-806

Nath R., Jackson W.M., Jones T. W., "A comparison of the classical and linear programming approaches to the classification problem in discriminant analysis" **Journal of Statistical Computation and Simulation** 41, 1992, 73-93

Orgler Y.E., "A credit scoring models for commercial loans" **Journal of Money Credit Banking**, November, 1970, 31-37

Orgler Y.E., "Evaluating of bank consumer loans with credit scoring models" **Journal of Bank Research**, 1, Spring, 1971, 31-37

Reichert A. K., Cho C.C., Wagner G.M., "An examination of the conceptual issues involved in developing credit scoring models" **Journal of Business and Ecomic Statistics,** 1, 1983, 101-114

Safavain S. F., Landgrebe D., "A survey of decision tree classifier methodology" **IEEE Trans. On Systems, Man and Cybernetics**, 21, 1991, 660-674.

Salas V., Saurina J., "Credit risk in two institutional regimes:Spanish commercial and savig banks" **Journal of Financial Service Reserach**, 22:3, 2002, 203-224.

Sharma, S. **Applied Multivariate Techniques**, John Wiley & Sons. 1996

Showers J.L., Chakrin L.M., "Reducing uncollectable revenue from residental telephone customers" **Interfaces**, 11, 1981, 21-31

Srinivasan V. Kim Y.H., "Credit granting: a comparative analysis of classification procedures" **Journal of Finance**, 42, 1987a, 665-683.

Tam K.Y., Kiang M.Y., "Managerial applications of neural networks: the case of bank failure prediction" **Management Science**, 38, 1992, 926-947.

Tatsuoka, M. M., **Multivariate Analysis**. New York: John Wiley. 1971

Thomas L. C., " A survey of credit and behavioural scoring: forecasting financial risk of lending to consumerd." **International Journal of Forecasting**, 16, 2000, 149-172

Thomas, L. C, Edelman, D. B., Crook J. N., "Credit scoring and its applications", **Society for Industrial and Applied Mathematics**, Philadelphia 2002

Titterington D.M "Discriminant analysis and related topics" In: Thomas L.C., Croock J.C and Edelman D.B. (eds), "**Credit scoring and credit control**", Oxford University Press, Oxford, 1992 53-73

Wagner G.M., Reichert A.K., Cho C.C, "Conceptual issues in credit credit scoring models". **Credit World**, 71, (May/June), 1983, 22-25

Wiginton J.C., "A note on the comparison of logit and discriminant models of consumer credit behavior" **Journal of Financial and Quantitative Analysis** XV,(3), 1980, 757-770

Yobas M.B., Crook J.N., Ross P., "Credit scoring using neural and evolutionary techniques" **Working Paper** 97/2, Credit research Centre, University of Edinburg.

Ziari H.A., Leatham D.J., Ellinger P.N., "Development of statistical discriminant mathematical programming model via resampling estimation techniques" **American Journal of Agricultural Economics**, 79, 1997, 1352-1362.

Zocco D.P., "A framework for expert systems in bank loan management" **Journal of Commercial Bank Lending**, 67, 1985, 47-54