



LOJİSTİK REGRESYON ANALİZİNDE KATSAYILARIN GENETİK ALGORİTMA İLE BELİRLENMESİ

Hazırlayanlar:
Şebnem ER, Timur KESKİNTÜRK

*İstanbul Üniversitesi İşletme Fakültesi
Sayısal Yöntemler Anabilim Dalı*



Lojistik Regresyon Analizi

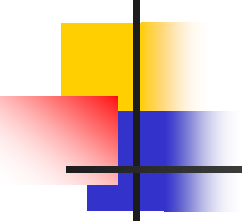
Lojistik regresyon analizi bağımlı değişkenin kategorik olması durumunda kullanılan bir regresyon analizidir.

$$Y = \begin{cases} 1 \\ 0 \end{cases}$$

Örneğin; $Y = \{\text{Başarılı } (Y=1) \vee \text{Başarısız } (Y=0)\}$

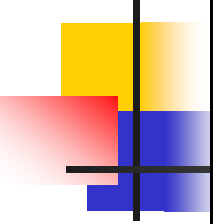
$Y = \{\text{Hasta } (Y=1) \vee \text{Hasta Değil } (Y=0)\}$

$Y = \{\text{Müşteri } (Y=1) \vee \text{Müşt.Değil } (Y=0)\}$


$$Y_i = b_0 + b_1X_{1i} + b_2X_{2i} + \dots + b_pX_{pi} + e_i$$

şeklinde Y bağımlı deęişkeninin 0, 1 gibi deęerler aldığı bir modele, klasik regresyon analizi uygulanarak katsayılar belirlendięinde;

- hataların normal daęılmaması,
- hataların deęişen varyanslı olması,
- tahmin edilen bağımlı deęişkenin 0-1 aralığı dıőında deęerler alması,
- düşük belirlilik katsayısı gibi sorunlarla karşılaşılmaktadır.



Y _i	Gerçekleşme Olasılığı	
1	$p = P(Y_i=1)$	Başarı olasılığı
0	$q = P(Y_i=0) = 1 - P(Y_i=1)$	Başarısızlık olasılığı
Toplam	1	

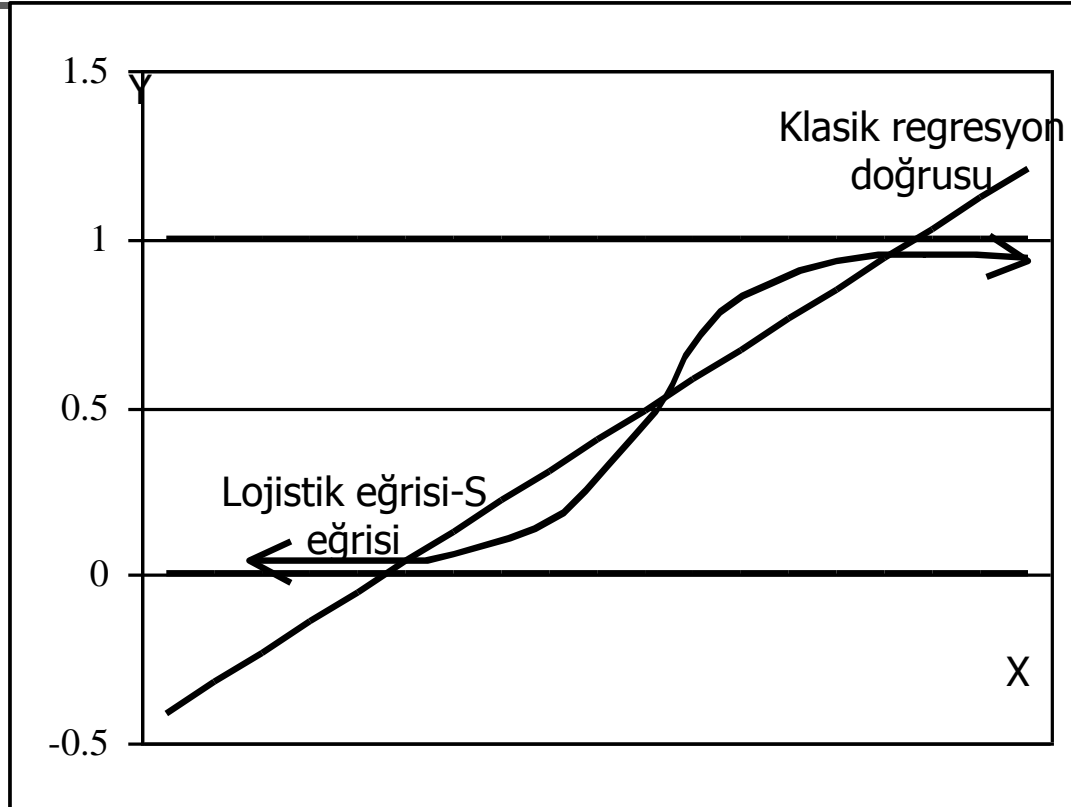
Lojistik Fonksiyon:

$$\text{Logit}(Y) = \ln \left[\frac{p}{1-p} \right] = \ln \left[\frac{p}{q} \right] = b_0 + b_1 X_1 + b_2 X_2 + \dots + b_p X_p + e_i$$

Tahmini Olasılıklar:

$$p = \frac{e^{b_0 + \sum_{k=1}^p b_k X_k}}{1 + e^{b_0 + \sum_{k=1}^p b_k X_k}}$$

S-Eğrisi



Böylelikle tahmini olasılıkların 0 ile 1 arasında olması sağlanmış olmaktadır.



Lojistik Regresyon Analizinde Katsayıların Belirlenmesi

- Lojistik regresyon modelinde katsayılar en yüksek benzerlik yöntemi (Maximum Likelihood Estimation) kullanılarak belirlenmektedir.
- En yüksek benzerlik yönteminde, anakütle ile anakütleden çekilen örnek arasındaki benzerlik ilişkisinden yararlanarak incelenen örneğin elde edilmesi olasılığını maksimum yapan parametre değeri tahmin edilmektedir.



En Yüksek Benzerlik Yöntemi

- Başarı olasılığı $P_i = E(Y_i = 1 | X_i)$
 - Başarısızlık olasılığı $(1 - P_i) = E(Y_i = 0 | X_i)$
- olarak tanımlandığından i 'nci gözlem için olasılık,

$$P(Y_i | X_i) = P_i^{y_i} (1 - P_i)^{1 - y_i} ; i = 1, \dots, n$$

biçiminde yazılmaktadır. Bağımsız oldukları varsayılan n gözlem için ortak olasılık yoğunluk fonksiyonu n tane tekil fonksiyonun çarpımı olarak

$$L = \prod_{i=1}^n P_i^{y_i} (1 - P_i)^{1 - y_i} \quad \text{şeklinde yazılmaktadır.}$$

- 
- Birden fazla bağımsız değişkenli model için benzerlik fonksiyonu ise

$$L = \prod_{i=1}^n \left(\frac{e^{b_0 + \sum_{k=1}^p b_k X_{ik}}}{1 + e^{b_0 + \sum_{k=1}^p b_k X_{ik}}} \right)^{y_i} \left(\frac{1}{1 + e^{b_0 + \sum_{k=1}^p b_k X_{ik}}} \right)^{(1-y_i)}$$

şeklindedir.



En Yksek Benzerlik Yntemi

- Lojistik regresyon analizinde benzerlik fonksiyonunu maksimize eden katsayılar seilir.
- Fonksiyonun trevinin alınması yerine belirli bir bařlangı deęeri seilerek iterasyon sonucu uygun parametreler bulunur.
- Burada en nemli nokta bařlangı deęerinin doęru belirlenmesidir.
- Bu alıřmada sezgisel bir yntem olan genetik algoritma kullanılarak benzerlik fonksiyonunu maksimize eden deęiřkenlerin neler olduęu ve bu deęiřkenlere ait katsayıların tahmini yapılmıřtır.



GENETİK ALGORİTMAYA GENEL BAKIŞ

- 1970' li yıllarda Amerika' da geliştirilmiştir,
- J. Holland, D. Goldberg,
- Sezgisel optimizasyon tekniğidir,
- Çözümlerin kodlanması, Amaç fonksiyonu,
- Genetik operatörler: Seçim, Çaprazlama, Mutasyon



Genetik Algoritma

- Başlangıç popülasyonu
- Uygunluk fonksiyonu
 - Seçim
 - Çaprazlama
 - Mutasyon
- Bitir ya da adım 2'ye dön

Lojistik Regresyon Analizinde Katsayıların Belirlenmesinde GA Uygulaması

- Çözümler katsayıların gerçek değerlerle ifade edildiği kromozomlarla temsil edilmektedir.

0,698	-1,236	0	1,305	-4,025	0
-------	--------	---	-------	--------	---

- Genin aldığı değer "0" ise ilgili değişken modele dahil edilmez.
- Çaprazlama: Orta Seviye Üretim (Intermediate);
- Mutasyon: Tek nokta mutasyonun özel bir biçimi kullanılmıştır; Seçilen genin değeri "0" ise -1 ile 1 arası rasgele bir sayı atanır, yani değişken modele dahil edilir; genin değeri "0" dan farklı ise "0" yapılır, yani modelden çıkarılır.



Amaç Fonksiyonu

- Genetik algorithmanda katsayıların belirlenmesi aşamasında uygunluk fonksiyonu aşağıdaki gibi kodlanmaktadır.

```
z=0;
aa=[0 x(2) x(3) x(4) x(5) x(6) x(7) x(8) x(9) x(10) x(11) x(12) x(13) x(14) x(15) x(16) x(17) ...];
indices=find(aa);
bb=aa(indices);
for j=1:1:size(dizix,1)
    d=exp(x(1)+sum(bb.*dizix(j,indices)));
    z=z-dizix(j,1)*(log((d/(1+d))))-(1-dizix(j,1))*log((1/(1+d)));
end
```



GA Parametreleri

- Populasyon büyüklüğü: 20-150
- Çaprazlama Oranı: 0.9
- Mutasyon Oranı: 0.4
- Matlab 7.0 GA Toolbox

Örnek Problemlere İlişkin Sonuçlar

Problem	Birim Sayısı	Bağımsız Değişken Sayısı	Uygunluk Fonksiyonu Değeri Sonuçları (InL)	
			SPSS	Genetik Algoritma
1	n=315	k=2	-183.137	-182.444
2	n=977	k=2	-443.241	-443.241
3	n=172	k=3	-74.601	-74.601
4	n=561	k=3	-301.507	-301.507
5	n=690	k=3	-446.510	-442.783
6	n=145	k=7	-10.548	-10.553

Best: 442.7831 Mean: 442.8199

