Gazi Mühendislik Bilimleri Dergisi
2017, 3(3): 14-26

gmbd.gazipublishing.com

Kısıtlandırılmamış Kısmi Oransal Odds Modelinin Doğru Sınıflandırma Performansı Üzerine Bir Çalışma

Hatice DAĞLIOĞLU[*,a], Semra ORAL ERBAŞ[b]

[a,*] *TÜBİTAK Savunma Sanayii Araştırma ve Geliştirme Enstitüsü, ANKARA 06261, TÜRKİYE*
[b] *Gazi Üniversitesi Fen Fakültesi İstatistik Bölümü, ANKARA 06500, TÜRKİYE*

MAKALE BİLGİSİ

Alınma: 15.08.2017
Kabul: 20.09.2017

Anahtar Kelimeler:
Oransal odds modeli, kısıtlandırılmamış kısmi oransal odds modeli doğru sınıflandırma oranı, simülasyon.

[*]*Sorumlu Yazar:*
e-posta:
hatice.daglioglu@
tubitak.gov.tr

ÖZET

Birimlerin iki ya da daha fazla düzeyli kategorik değişkenler bakımından sınıflandırılmasında birçok yöntem kullanılmaktadır. Bu yöntemlerden bazıları, bağımlı değişken düzeyinin ikiden fazla ve sıralı bir yapıda olması durumunda kullanılan oransal odds modeli ve bu modele ait temel varsayımın bazı değişkenler için sağlanıp bazı değişkenler için sağlanmaması durumunda kullanılan kısıtlandırılmamış kısmi oransal odds modelidir. Bu çalışmada kısıtlandırılmamış kısmi oransal odds modeli ele alınarak bağımlı değişken düzeyinin sayısı, bağımsız değişken sayısı ve örneklem büyüklüğü değiştirildiğinde doğru sınıflandırma oranları incelenmiştir. Ayrıca bağımsız değişkenlerin tümünün sürekli olması durumu için simülasyon çalışması yapılmış ve bu veriler için oransal odds modeli, kısıtlandırılmamış kısmi oransal odds modeli, doğrusal diskriminant analizi ve karesel diskriminant analizi yöntemlerinin birbirlerine göre üstünlükleri ortaya konulmaya çalışılmıştır.

DOI:

A Study on Correct Classification Performance of Unconstraint Partial Proportional Odds Model

ARTICLE INFO

Received: 15.08.2017
Accepted: 20.09.2017

Keywords:
Proportional odds model, unconstraint partial proportional odds model, correct classification rate, simulation.

[*]***Corresponding Authors***
e-mail:
hatice.daglioglu@
tubitak.gov.tr

ABSTRACT

Several methods are used to classify units in terms of categorical variables having two or more levels. Some of these methods are proportional odds model which is used when the level of the dependent variable is more than two and a ordinal pattern and the unconstraint partial proportional odds model which is used when the fundamental assumption of this model is established from the some of the variables and is not satisfied some of the variables. In this study, the correct classification rates are investigated with varying the number of classes of dependent variables, the number of independent variables and the sample size by taking unconstraint partial proportional odds model. A simulation study is performed in the case of all independent variables are continuous and also superiority of proportional odds model, unconstraint partial proportional odds model, linear discriminant analyze and quadratic discriminant analyze to each other are tried to revealed for these data.

1. Giriş (Introduction)

Regresyon analizinde bağımlı değişkenin kategorik olduğu durumlar göz önünde bulundurulduğunda çeşitli lojistik regresyon teknikleri kullanılmaktadır. Bu teknikler bağımlı değişkene ait düzey sayısının iki veya ikiden fazla olması durumunda farklılaşmaktadır [4].

Bağımlı değişkene ait düzey sayısının ikiden fazla ve sıralı olması durumunda sıralı lojistik regresyon veya diğer bir ifadeyle oransal odds modelinin kullanılması söz konusu olmaktadır. Bu teknikte önemli bir varsayım olan paralel doğrular varsayımının kontrolü ile çalışmanın sınırları genişlemektedir. Böylece farklı alanlarda çalışmalar yapılabilmektedir. Oransal odds modelinde yer alan değişkenlerin bazılarının bahsedilen 'paralel doğrular varsayımını' sağlaması ve geriye kalan diğer değişkenleri sağlamaması sonucu kısıtlandırılmamış kısmi oransal odds modeli adı verilen yöntem kullanılmaktadır [9].

2. Oransal Odds Modeli (Proportional Odds Modeli)

Bir klasik doğrusal regresyon modelinde bağımsız değişken X ve bağımlı değişken Y arasındaki en basit ilişki şu şekilde ifade edilsin:

$$Y_i = \alpha + \beta X_i + \varepsilon_i \qquad i=1,\ldots,n \qquad (1)$$

Burada Y_i, bağımlı değişkeni; β eğim katsayısını; X, bağımsız değişkeni ve ε_i hata terimini göstermektedir [6]. Bağımlı değişken, nicel olabildiği gibi nitel değişken de olabilir.

Bağımlı değişkenin iki ya da daha çok kategoride gözlemlendiği durumlarda bağımlı değişken ile açıklayıcı değişken arasındaki ilişkiyi tanımlayan en uygun yöntemlerden biri, lojistik regresyon analizidir. Bağımlı değişkenin kategorik olduğu durumda bağımlı değişkenin düzey sayısının iki olması durumuna ikili (binary) lojistik regresyon, ikiden fazla olması durumunda ise çok düzeyli (polytomous) lojistik regresyon modelleri kullanılmaktadır. Çok düzeyli lojistik regresyon modellerinde bağımlı değişken ikiden fazla sıralı ölçme düzeyinde ise sıralı lojistik regresyon veya oransal odds modeli olarak adlandırılmaktadır [11]. Oransal odds modeli, Mc Cullagh tarafından 1980 yılında bağımlı değişkenin sıralı olduğu durumlar için önerilen en popüler modeldir. Yanıtları "kesinlikle katılıyorum", "katılıyorum" ve "kesinlikle katılmıyorum" şeklinde belirtilen anket soruları; "yüksek", "orta", "düşük" şeklinde ifade edilen gelir düzeyleri; "hiç çalışmayan", "yarı zamanlı çalışan", "tam zamanlı çalışan" şeklindeki istihdam durumları; öğrencilerin notlarının A dan F'ye kadar ölçeklerde değerlendirilmesi gibi 'likert ölçekli' bağımlı değişkenler örnek verilebilir [10].

Oransal odds modelinde, gözlemlenebilir Y bağımlı değişkeninin elde edilmesi için -∞ ve +∞ aralığında değerler alabilen ancak gözlemlenemeyen bir gizli Y^* değişkeni olduğu ve bu gizli Y^* değişkeninin Y' ye bilgi sağladığı düşünülür. Böylece gizli Y^* değişkeni ile sıralı kategorik değişken Y arasındaki ilişki şu şekilde ifade edilebilir [10].

$$\begin{aligned} &Y = 1 \text{ iken } -\infty = \alpha_0 \leq Y^* \leq \alpha_1 \\ &Y = 2 \text{ iken } \alpha_1 < Y^* \leq \alpha_2 \\ &Y = 3 \text{ iken } \alpha_2 < Y^* \leq \alpha_3 \\ &\ldots \\ &Y = J \text{ iken } \alpha_{J-1} < Y^* = \infty \end{aligned}$$

Burada j, sıralı düzeyli kategorik bağımlı değişkenin düzey sayısı olmak üzere α 'lar "eşik değerleri" (threshold) veya "kesim noktaları" (cut points) olarak ifade edilir. j sıralı düzeyli kategorik bağımlı değişkenin sınıflandırılmasında $j-1$ tane eşik değeri hesaplanır. Ayrıca $0 < \alpha_1 < \alpha_2 < \cdots < \alpha_j$ olacak şekilde bir sıralama mevcuttur. α 'lar modelde bulunan β'lar ile birlikte tahmin edilir.

Oransal odds modelinde elde edilen katsayıların yorumlanmasında odds ve odds oranından yararlanılır ve model literatürde kümülatif odds modeli olarak da geçmektedir. Odds, bir olayın olması olasılığının (P_i), olmaması olasılığına ($1-P_i$) oranıdır ve $odds = \frac{P_i}{1-P_i}$ şeklinde gösterilir. Odds, 0 ile $+\infty$ arasında değerler alabilir. Odds değerinin doğal logaritmasına ise lojit adı verilir ve $\ln[\frac{P_i}{1-P_i}]$ şeklinde gösterilir [2].

Y bağımlı değişkeninin j sayıda sıralı bağımlı kategorisi olsun ve $1,2,\ldots,j$ değerlerini alsın. X bağımsız değişken olmak üzere oransal odds modeli şu şekilde ifade edilir [13].

$$\begin{aligned} \ln(Y_j) = \ln\left[\frac{P(Y \leq j|X_i)}{1-P(Y \leq j|X_i)}\right] &= \alpha_j \\ &+ (\beta_1 X_1 + \beta_2 X_2 + \cdots \\ &+ \beta_i X_i) \end{aligned} \qquad (2)$$

Modelde bulunan j adet sıralı kategoriden, bir kategori referans olarak seçilmekte (genellikle son kategori) ve bağımsız değişkenler mevcutken verilen Y yanıtının seçilen kategori ya da daha alt bir kategoriye düşme olasılığı yardımıyla lojitler hesaplanarak, kümülatif olasılıklar elde edilmekte ve

kategoriye düşme olasılıkları bulunmaktadır [3]. Kategoriye düşme olasılıkları şu şekilde elde edilmektedir:

$$P(Y=1) = P(\mathrm{Y} \leq 1|X_i)$$
$$P(Y=2) = P(\mathrm{Y} \leq 2|X_i) - P(\mathrm{Y} \leq 1|X_i)$$
$$P(Y=3) = P(\mathrm{Y} \leq 3|X_i) - P(\mathrm{Y} \leq 2|X_i)$$
$$\vdots$$
$$P(Y=j) = 1 - P(\mathrm{Y} \leq (j-1)|X_i)$$

Oransal odds modelinin uygulanabilmesi için paralel doğrular varsayımı olarak adlandırılan temel bir varsayımın sağlanması gerekmektedir. Bu varsayım, kesim noktası eşitliklerine karşı β' ların eşitliği olarak bilinmektedir. Yani modelde elde edilen eğim katsayısı β her lojit için aynı etkiye sahiptir ve bağımlı değişkenin düzey sayısına bağlı olarak elde edilen modellerde sabit katsayılar farklı olmasına rağmen eğim parametreleri aynıdır. Paralel doğrular varsayımı sağlanmadığı durumlarda ise kısıtlandırılmamış kısmi oransal odds modeli, söz konusu olabilmektedir [12].

3. Kısıtlandırılmamış Kısmi Oransal Odds Modeli (Unconstraint Partial Proportional Odds Modeli)

Oransal odds modelindeki bağımsız değişkenlerin oransal odds modelinin altında yatan varsayım olan paralel doğrular varsayımını kesinlikle sağlaması istenir. Ancak pratikte bu zordur [8]. Bazı açıklayıcı değişkenler bu varsayımı sağlamayabilir. Buradan hareketle paralel doğrular varsayımının bazı değişkenler için sağlanıp ve bazı değişkenler için sağlanmadığı durumlarda kısıtlandırılmamış kısmi oransal odds modeli kullanılmaktadır [12].

Kısıtlandırılmamış kısmi oransal odds modelinde, iki katsayı kümesi tahmin edilmektedir. İlk katsayı kümesi oransal odds değerinin bulunduğu yani, paralel doğrular varsayımının sağlandığı ve ikinci katsayı kümesi de oransal oddsun bulunmadığı yani paralel doğrular varsayımının sağlanmadığı kümedir. Böylece parametrelerin bazılarının paralel doğrular varsayımını sağlaması, bazılarının da sağlamaması nedeniyle modele 'kısıtlandırılmamış kısmi' oransal odds adı verilmektedir [14].

$\mathrm{q} \leq \mathrm{p}$ olmak üzere p açıklayıcı değişkenin q alt kümesi için paralel doğrular varsayımının sağlandığı bir model söz konusu olsun. Böyle bir modeli formülize etmek için, n bağımsız rasgele gözlemin örneklem olarak alındığı ve bu gözlemlerin yanıtlarının sıralı değişken olan Y'nin $\mathrm{k}+1$ kategorilerinde sınıflandırıldığı düşünülsün $(\mathrm{Y} = 0,1,\dots,\mathrm{k})$. Böylece her gözlem bağımsız çokterimli dağılıma sahip olmaktadır.

Bazılarının oransal oddsa sahip olduğu ve bazılarının da oransal olmayan oddsa sahip olduğu değişkenlerin bulunduğu kısıtlandırılmamış kısmi oransal odds modeli, kümülatif olasılıklar cinsinden,

$$C_{ij} = \Pr\left(Y \geq \frac{j}{X_i}\right) = \frac{1}{1+\exp\left(-\alpha_j - X_i'\beta - T_i'\gamma_j\right)}; \quad j = 1,\dots,k \qquad (3)$$

şekilde ifade edilir [14].

Burada;

- α_j: kesim noktasıdır. $\alpha_1 > \alpha_2 > \cdots > \alpha_k$
- X_i: p açıklayıcı değişkenin tüm kümesi üzerinde i gözlem değerlerini içeren $px1$ boyutlu vektördür.
- β: X_i'deki p değişkenlerle ilişkilendirilen regresyon katsayılarının $px1$ boyutlu vektörüdür.
- T_i: $q \leq p$ olmak üzere oransal odds varsayımının ya yapılmadığı ya da test edilmediği durumda, p açıklayıcı değişkenlerinin alt kümesi üzerinde i gözleminin değerlerini içeren $qx1$ boyutlu vektördür.
- γ_j: T_i'deki q değişkenlerle ilişkilendirilen regresyon katsayılarının $qx1$ boyutlu vektörüdür. Böylece sadece j. kümülatif lojit $j = 1,\dots,k$ ile ilişkilendirilen artıştır ve $\gamma_1 = 0$'dır.

β_l'nin elemanları $l = 1,\dots,p$ olmak üzere β_l ve γ_j'nin elemanları da $l = 1,\dots,q$ olmak üzere γ_{jl} ile gösterilir. Bu gösterim T_i'nin X_i'deki ilk q elemanlarına eşit olduğu anlamına gelir. Böylece oransal odds sadece X_i'deki son p-q değişkenleri için sağlanır.

Lojit, çokterimli lojistik regresyonda verilen bir kategori ile referans kategorisi arasındaki karşılaştırmalara dayanmaktadır. Böylece kısıtlandırılmamış kısmi oransal odds modelinde lojit, oransal odds modeline benzer olarak, j'den yüksek kategorilere karşı j'ye eşit veya daha az kategorinin odds değerini referans almaktadır.

Bu modelden açıklayıcı değişkene ait, tek bir β katsayısı ve $k-2\gamma$ adet katsayı elde edilir. Burada k, Y bağımlı değişkenindeki kategori sayısıdır ve γ (gamma) katsayısı oransallıktan sapmayı ifade etmektedir [16]. Eğer tüm gammalar sıfıra eşitse, model oransal odds modeline dönüşür [1]. Modelde

yer alan γ (gamma) parametresinin birçok avantajı vardır:

- γ parametresi modelde odds değerinin ne kadar arttığını göstermektedir. Yani oransal olmayan bir model için odds oranlarındaki artıştaki değişimi tahmin eder [7].
- Paralel doğrular varsayımını açıklamada araştırmacıya farklı bir yol sunmaktadır. Eğer gamma parametresi bir değişken için sıfıra eşitse, varsayımın sağladığı sonucu elde edilir. Eğer tüm j için $\gamma_j = 0$ ise, bu modelde oransallık sağlanmakta sonucu elde edilir ve model oransal odds modeline dönüşür. [14].
- Parametre açısından değerlendirildiğinde modelde daha sıkı bir düzenin varlığından söz edilir. Yani daha az parametre ile model elde edilebilir ve aynı zamanda modelde potansiyel olarak görülen problemlerin ilk etapta belirlenmesini sağlayabilir [15].
- Gamma parametresinin incelenmesiyle paralel doğrular varsayımının tam olarak nerede ihlal edildiği saptanabilir. Gamma parametresinin bulunduğu değişkenler varsayımın sağlanmadığı değişkenlerdir [15].
- Kısıtlandırılmamış kısmi oransal odds modelinde lojitlerin elde edilmesinde γ (gamma) parametresinden yararlanılarak hesaplamalar yapılır.

4. Simülasyon Çalışması (Simulation Study)

Bu bölümde kısıtlandırılmamış kısmi oransal odds modeli, oransal odds modeli, doğrusal diskriminant analizi ve karesel diskriminat analizinin sınıflandırma performanslarını değerlendirmek amacıyla simülasyon çalışmalarına yer verilmiştir. Dört modelin sınıflandırma performanslarını kıyaslamak için her modele ait doğru sınıflama oranı elde edilmiştir. Analiz için dört modelin doğru sınıflama oranlarının elde edilmesini içeren simülasyon kodu STATA 12 programında hazırlanmıştır. Veri üretiminin aşamaları aşağıdaki gibi gerçekleşmiştir:

Adım 1: Kısıtlandırılmamış kısmi oransal odds modelinde bağımlı değişkene ait düzey sayısının en az üç olması gerektiği için, sırasıyla bağımlı değişken düzey sayısı üç, dört ve beş olarak seçilmiştir. Bağımsız değişkenlerin türü sürekli olarak belirlenmiştir. Ayrıca örneklem büyüklüğü bağımsız değişken sayısı ve bağımlı değişkene ait düzey sayısı için sırasıyla $100, 500, 1000, 5000$ ve $10\,000$ olarak belirlenmiştir. α_j ve β_l katsayıları keyfi olarak verilen sayılarla belirlenerek (düzey sayısı–1) adet model denklemi elde edilmiştir. Elde edilen model denklemleri için paralel doğrular varsayımını test eden istatistikler (Wald testi, olabilirlik oran testi vb.) kontrol edilmiştir. Ayrıca modelin ve elde edilen katsayıların anlamlılık kontrolleri de test edilmiştir.

Adım 2: Diskriminant analizinde grup sayısı, bağımlı değişkenin düzey sayısı gibi düşünülüp, her grupta bulunan birimlerin ortalama vektörü tanımlanmıştır. (Değişken sayısı ve grup sayısına göre bu değerler değişmektedir.) Her gruba ait varyans-kovaryans matrisi belirlenmiştir. (Değişken sayısı ve grup sayısına göre bu değerler değişmektedir.) Karesel diskriminant analizinde ise, analiz başında belirlenen varyans-kovaryans matrisleri alınmıştır. Kısıtlandırılmamış kısmi oransal odds modeli için belirlenen bağımlı değişkene ait düzey sayısı, bağımsız değişken sayısı ve örneklem büyüklüğü gibi bilgiler hem karesel hem de doğrusal diskriminant analizinde kullanılmıştır.

Adım 3: Simülasyon çalışması için algoritma belirlendikten sonra, STATA programında simülasyon için modellerin algoritmasını içeren bir kod hazırlanmıştır. STATA programında, kısıtlandırılmamış kısmi oransal odds modeli ve oransal odds modeli için doğru sınıflama oranının elde edilmesi aşamasında sırasıyla; her analiz için kategori olasılıklarının belirlenmesi, kategori olasılıklarının saptanması ve elde edilen olasılıklara göre kategorilere atama işleminin yapılmıştır. Elde edilen veriler için çapraz tablo kurularak doğru sınıflama oranına ulaşılmıştır.

Adım 4: Simülasyon çalışmasında, beş farklı örneklem büyüklüğü üretilerek, her örneklem büyüklüğü için doğru sınıflama oranları elde edilmiştir. Ele alınan her örneklem büyüklüğü için iterasyon sayısı $10\,000$ olarak alınmıştır. $10\,000$ iterasyon ile yapılan analiz sonucunda her modele ait ortalama doğru sınıflama oranı elde edilmiştir [5].

Çizelge 1: Sürekli bağımsız değişkenler ile üretilen denklemler

Simülasyon Çalışması No	Bağımlı Değişkenin Düzeyi (y)	Bağımsız Değişken Sayısı (x_i)	Üretilen Gözlem Sayısı (n)	Model Denklemleri	Bağımlı/Bağımsız Değişken Açıklaması
1	3	3	100	$y_1 = 2 - 12x_1 + 11x_2 + 2x_3 + \varepsilon$ $y_2 = 3 - 12x_1 + 11x_2 + 2.25x_3 + \varepsilon$	y bağımlı değişkeni 1,2,3 değerlerini alan sıralı bir yapıdadır. $x_1 \; ve \; x_2$ paralel doğrular varsayımını sağlamakta, x_3 paralel doğrular varsayımını sağlamamaktadır.
2			500		
3			1000		
4			5000		
5			10000		
6		5	100	$y_1 = 2 - 12x_1 + 10x_2 + 2x_3 + 0.3x_4 + 0.1x_5 + \varepsilon$ $y_2 = 3 - 12x_1 + 10x_2 + 2.1x_3 + 0.3x_4 + 0.1x_5 + \varepsilon$	y bağımlı değişkeni 1,2,3 değerlerini alan sıralı bir yapıdadır. $x_1, x_2, x_4 \; ve \; x_5$ paralel doğrular varsayımını sağlamakta, x_3 paralel doğrular varsayımını sağlamamaktadır.
7			500		
8			1000		
9			5000		
10			10000		
11		7	100	$y_1 = 2 - 12x_1 + 10x_2 + 2x_3 + 0.3x_4 + 0.1x_5 + 0.2x_6 + 0.1x_7 + \varepsilon$ $y_2 = 3 - 12x_1 + 10x_2 + 2.1x_3 + 0.3x_4 + 0.1x_5 + 0.2x_6 + 0.1x_7 + \varepsilon$	y bağımlı değişkeni 1,2,3 değerlerini alan sıralı bir yapıdadır. $x_1, x_2, x_4, x_5, x_6 \; ve \; x_7$ paralel doğrular varsayımını sağlamakta, x_3 paralel doğrular varsayımını sağlamamaktadır.
12			500		
13			1000		
14			5000		
15			10000		

Çizelge 2: Sürekli bağımsız değişkenler ile üretilen denklemler

Simülasyon Çalışması No	Bağımlı Değişkenin Düzeyi (y)	Bağımsız Değişken Sayısı (x_i)	Üretilen Gözlem Sayısı (n)	Model Denklemleri	Bağımlı/Bağımsız Değişken Açıklaması
16			100		
17			500		
18		3	1000	$y_1 = 2 - 12x_1 + 11x_2 + 2x_3 + \varepsilon$ $y_2 = 3 - 12x_1 + 11x_2 + 2.25x_3 + \varepsilon$ $y_3 = 4 - 12x_1 + 11x_2 + 2.29x_3 + \varepsilon$	y bağımlı değişkeni 1,2,3,4 değerlerini alan sıralı bir yapıdadır. x_1 ve x_2 paralel doğrular varsayımını sağlamakta, x_3 paralel doğrular varsayımını sağlamamaktadır.
19			5000		
20			10000		
21			100		
22			500		
23	4	5	1000	$y_1 = 2 - 12x_1 + 10x_2 + 2x_3 + 0.3x_4 + 0.1x_5 + \varepsilon$ $y_2 = 3 - 12x_1 + 10x_2 + 2.1x_3 + 0.3x_4 + 0.1x_5 + \varepsilon$ $y_3 = 4 - 12x_1 + 10x_2 + 2.29x_3 + 0.3x_4 + 0.1x_5 + \varepsilon$	y bağımlı değişkeni 1,2,3,4 değerlerini alan sıralı bir yapıdadır. x_1, x_2, x_4 ve x_5 paralel doğrular varsayımını sağlamakta, x_3 paralel doğrular varsayımını sağlamamaktadır.
24			5000		
25			10000		
26			100		
27			500		
28		7	1000	$y_1 = 2 - 12x_1 + 10x_2 + 2x_3 + 0.3x_4 + 0.1x_5 + 0.2x_6 + 0.1x_7 + \varepsilon$ $y_2 = 3 - 12x_1 + 10x_2 + 2.1x_3 + 0.3x_4 + 0.1x_5 + 0.2x_6 + 0.1x_7 + \varepsilon$ $y_3 = 4 - 12x_1 + 10x_2 + 2.29x_3 + 0.3x_4 + 0.1x_5 + 0.2x_6 + 0.1x_7 + \varepsilon$	y bağımlı değişkeni 1,2,3,4 değerlerini alan sıralı bir yapıdadır. x_1, x_2, x_4, x_5, x_6 ve x_7 paralel doğrular varsayımını sağlamakta, x_3 paralel doğrular varsayımını sağlamamaktadır.
29			5000		
30			10000		

Çizelge 3: Sürekli bağımsız değişkenler ile üretilen denklemler

Simülasyon Çalışması No	Bağımlı Değişkenin Düzeyi (y)	Bağımsız Değişken Sayısı (x_i)	Üretilen Gözlem Sayısı (n)	Model Denklemleri	Bağımlı/Bağımsız Değişken Açıklaması
31			100	$y_1 = 2 - 12x_1 + 11x_2 + 2x_3 + \varepsilon$ $y_2 = 3 - 12x_1 + 11x_2 + 2.25x_3 + \varepsilon$ $y_3 = 4 - 12x_1 + 11x_2 + 2.27x_3 + \varepsilon$ $y_4 = 4 - 12x_1 + 11x_2 + 2.38x_3 + \varepsilon$	y bağımlı değişkeni 1,2,3,4,5 değerlerini alan sıralı bir yapıdadır. x_1 ve x_2 paralel doğrular varsayımını sağlamakta, x_3 paralel doğrular varsayımını sağlamamaktadır.
32			500		
33		3	1000		
34			5000		
35			10000		
36			100	$y_1 = 2 - 12x_1 + 11x_2 + 2x_3 + 0.3x_4 + 0.1x_5 + \varepsilon$ $y_2 = 3 - 12x_1 + 11x_2 + 2.25x_3 + 0.3x_4 + 0.1x_5 + \varepsilon$ $y_3 = 4 - 12x_1 + 11x_2 + 2.27x_3 + 0.3x_4 + 0.1x_5 + \varepsilon$ $y_4 = 5 - 12x_1 + 11x_2 + 2.38x_3 + 0.3x_4 + 0.1x_5 + \varepsilon$	y bağımlı değişkeni 1,2,3,4,5 değerlerini alan sıralı bir yapıdadır. x_1, x_2, x_4 ve x_5 paralel doğrular varsayımını sağlamakta, x_3 paralel doğrular varsayımını sağlamamaktadır.
37			500		
38	5	5	1000		
39			5000		
40			10000		
41			100	$y_1 = 2 - 12x_1 + 11x_2 + 2x_3 + 0.3x_4 + 0.1x_5$+$0.2x_6 + 0.1x_7 + \varepsilon$ $y_2 = 3 - 12x_1 + 11x_2 + 2.25x_3 + 0.3x_4 + 0.1x_5$+$0.2x_6 + 0.1x_7 + \varepsilon$ $y_3 = 4 - 12x_1 + 11x_2 + 2.27x_3 + 0.3x_4 + 0.1x_5$+$0.2x_6 + 0.1x_7 + \varepsilon$ $y_4 = 5 - 12x_1 + 11x_2 + 2.38x_3 + 0.3x_4 + 0.1x_5$+$0.2x_6 + 0.1x_7 + \varepsilon$	y bağımlı değişkeni 1,2,3,4,5 değerlerini alan sıralı bir yapıdadır. x_1, x_2, x_4, x_5, x_6 ve x_7 paralel doğrular varsayımını sağlamakta, x_3 paralel doğrular varsayımını sağlamamaktadır.
42			500		
43		7	1000		
44			5000		
45			10000		

5. Veri Üretimi Sonuçları

Çizelge 1, Çizelge 2 ve Çizelge 3, sürekli bağımsız değişkenler ile üretilen denklemlere ait simülasyon senaryolarıdır. Çizelge 1, Çizelge 2 ve Çizelge 3 ile verilen algoritmaya uygun olarak kısıtlandırılmamış kısmi oransal odds modeli, oransal odds modeli, doğrusal diskriminant analizi ve karesel diskriminant analizinin performanslarını değerlendirmek amacıyla 10 000 iterasyon sonucunda her modele ait doğru sınıflandırma oranlarını veren simülasyon sonuçları elde edilmiştir. Analizde hata teriminin 0 ortalama ve 1 standart sapma ile normal dağıldığı kabul edilmiş ve elde edilen simülasyon sonuçları Çizelge 4 ile Çizelge 6 arasında verilmiştir.

Çizelge 4'de verilen simülasyon sonuçları bağımlı değişken düzeyinin üç olması durumunda elde edilen doğru sınıflama oran sonuçlarını içermektedir.

Çizelge 5'te verilen simülasyon sonuçları bağımlı değişken düzeyinin dört olması durumunda elde edilen doğru sınıflama oran sonuçlarını içermektedir.

Çizelge 6'da verilen simülasyon sonuçları bağımlı değişken düzeyinin beş olması durumunda elde edilen doğru sınıflama oran sonuçlarını içermektedir.

Çizelge 4: Hataların normal dağıldığı durumda doğru sınıflandırma oranları

Bağımlı Değişken Düzeyi					
3					
Bağımsız Değişken Sayısı	Örnekleme Büyüklüğü	Doğru Sınıflandırma Oranı			
		Kısıtlandırılmamış Kısmi Oransal Odds Modeli	Oransal Odds Modeli	Karesel Diskriminant Analizi	Doğrusal Diskriminant Analizi
3	100	0,9347	0,9329	0,9402	0,942
	500	0,9374	0,934	0,922	0,918
	1000	0,9395	0,936	0,906	0,884
	5000	0,9392	0,936	0,903	0,8794
	10000	0,9392	0,936	0,9029	0,8793
5	100	0,9782	0,9752	0,9595	0,9357
	500	0,9743	0,9734	0,9545	0,9348
	1000	0,9715	0,9708	0,9533	0,9339
	5000	0,9701	0,9697	0,9518	0,9328
	10000	0,97	0,9696	0,9517	0,9327
7	100	0,988	0,988	0,8692	0,8444
	500	0,981	0,981	0,8682	0,8413
	1000	0,978	0,9771	0,8561	0,837
	5000	0,9758	0,9754	0,8535	0,8358
	10000	0,9756	0,9752	0,8533	0,8392

Çizelge 5: Hataların normal dağıldığı durumda doğru sınıflandırma oranları

Bağımlı Değişken Düzeyi					
4					
Bağımsız Değişken Sayısı	Örnekleme Büyüklüğü	Doğru Sınıflandırma Oranı			
		Kısıtlandırılmamış Kısmi Oransal Odds Modeli	Oransal Odds Modeli	Karesel Diskriminant Analizi	Doğrusal Diskriminant Analizi
3	100	0,91	0,91	0,975	0,965
	500	0,9382	0,9332	0,9269	0,9071
	1000	0,9366	0,932	0,9267	0,9069
	5000	0,9347	0,9308	0,9254	0,9066
	10000	0,9346	0,9306	0,9253	0,9065
5	100	0,9744	0,9735	0,975	0,971
	500	0,9743	0,9734	0,9545	0,9348
	1000	0,9614	0,957	0,9642	0,9431
	5000	0,959	0,9554	0,9634	0,9429
	10000	0,9588	0,9552	0,9631	0,9425
7	100	0,99	0,99	0,89	0,87
	500	0,97	0,972	0,8766	0,8585
	1000	0,9734	0,9696	0,8706	0,8543
	5000	0,9697	0,9672	0,8642	0,8514
	10000	0,9693	0,967	0,8633	0,851

Çizelge 6: Hataların normal dağıldığı durumda doğru sınıflandırma oranları

Bağımlı Değişken Düzeyi					
5					
Bağımsız Değişken Sayısı	Örnekleme Büyüklüğü	Doğru Sınıflandırma Oranı			
		Kısıtlandırılmamış Kısmi Oransal Odds Modeli	Oransal Odds Modeli	Karesel Diskriminant Analizi	Doğrusal Diskriminant Analizi
3	100	0,9475	0,939	0,9291	0,9112
	500	0,9425	0,9381	0,9262	0,9077
	1000	0,9406	0,9361	0,9251	0,9074
	5000	0,9388	0,935	0,9242	0,9065
	10000	0,9386	0,9349	0,9238	0,9064
5	100	0,988	0,9796	0,951	0,9352
	500	0,981	0,9793	0,9506	0,9316
	1000	0,9815	0,978	0,944	0,9235
	5000	0,9793	0,9747	0,9395	0,9223
	10000	0,9777	0,9742	0,9425	0,9232
7	100	0,9915	0,9891	0,8354	0,8256
	500	0,9911	0,9889	0,8391	0,8298
	1000	0,9901	0,9884	0,8387	0,8296
	5000	0,9881	0,9861	0,8336	0,8221
	10000	0,9874	0,9856	0,8313	0,8208

Elde edilen çizelge 4, çizelge 5 ve çizelge 6 incelendiğinde aşağıdaki sonuçlar elde edilmiştir:

- Bağımlı değişken düzeyinin üç ve bağımsız değişken sayısının üç olması halinde en yüksek doğru sınıflandırma oranı, örneklem büyüklüğünün en düşük olduğu (n=100) doğrusal diskriminant analizinde (0,942) elde edilmiştir.
- Bağımlı değişken düzeyinin üç ve bağımsız değişken sayısının beş olması halinde en yüksek doğru sınıflandırma oranı, örneklem büyüklüğünün en düşük olduğu (n=100) kısmi oransal odds modelinde (0,978) elde edilmiştir.
- Bağımlı değişken düzeyinin üç ve bağımsız değişken sayısının yedi olması halinde ise en yüksek doğru sınıflandırma oranı, örneklem büyüklüğünün en düşük olduğu (n=100) kısmi oransal odds modeli ile oransal odds modelinde (0,988) elde edilmiştir.
- Bağımlı değişken düzeyinin üç ve bağımsız değişken sayısının üç olması halinde en yüksek doğru sınıflandırma oranı, örneklem büyüklüğünün en düşük olduğu (n=100) doğrusal diskriminant analizinde (0,942) elde edilmiştir.
- Bağımlı değişken düzeyinin dört ve bağımsız değişken sayısının üç olması halinde en yüksek doğru sınıflandırma oranı, örneklem büyüklüğünün en düşük olduğu (n=100) karesel diskriminant analizinde (0,975) elde edilmiştir.
- Bağımlı değişken düzeyinin dört ve bağımsız değişken sayısının beş olması halinde en yüksek doğru sınıflandırma oranı, örneklem büyüklüğünün en düşük olduğu (n=100) karesel diskriminant analizinde (0,975) elde edilmiştir.
- Bağımlı değişken düzeyinin dört ve bağımsız değişken sayısının yedi olması halinde ise en yüksek doğru sınıflandırma oranı, örneklem büyüklüğünün en düşük olduğu (n=100) kısmi oransal odds modeli ile oransal odds modelinde (0,99) elde edilmiştir.
- Bağımlı değişken düzeyinin beş ve bağımsız değişken sayısının üç olması halinde ise en yüksek doğru sınıflandırma oranı, örneklem büyüklüğünün en düşük olduğu (n=100) kısmi oransal odds modelinde (0,947) elde edilmiştir.
- Bağımlı değişken düzeyinin beş ve bağımsız değişken sayısının beş olması halinde ise en yüksek doğru sınıflandırma oranı, örneklem büyüklüğünün en düşük olduğu (n=100) kısmi oransal odds modelinde (0,988) elde edilmiştir.
- Bağımlı değişken düzeyinin beş ve bağımsız değişken sayısının yedi olması halinde ise en yüksek doğru sınıflandırma oranı, örneklem büyüklüğünün en düşük olduğu (n=100) kısmi oransal odds modelinde (0,99) elde edilmiştir.
- Kısıtlandırılmamış kısmi oransal odds modeli için elde edilen doğru sınıflandırma oranları incelendiğinde, bağımlı değişken düzeyinin üç ve bağımsız değişken sayısının üç olması halinde, örneklem büyüklüğünün (n=5000) ve n=10000 olduğu durumlarda doğru sınıflandırma oranları 0,93 olarak elde edilmiştir. Bağımlı değişken düzeyinin beş ve bağımsız değişken sayısının beş olması durumunda, örneklem büyüklüğünün (n=5000) ve n=10000 olduğu hallerde doğru sınıflandırma oranları 0,959 olarak bulunmuştur. Bununla birlikte bağımlı değişken düzeyinin yedi ve bağımsız değişken sayısının yedi olması halinde, örneklem büyüklüğünün (n=5000) ve n=10000 olduğu durumlarda doğru sınıflandırma oranları aynı değer (0,98) elde edilmiştir.
- Kısıtlandırılmamış kısmi oransal odds modelinde örneklem büyüklüğü, bağımlı değişkenin düzeylerine göre arttıkça doğru sınıflandırma oranlarının arttığı sonucu elde edilmiştir. Örneğin, bağımlı değişken düzeyinin üç olması halinde kısıtlandırılmamış kısmi oransal odds modelinde örneklem büyüklüğü arttıkça doğru sınıflandırma oranlarının arttığı; buna rağmen bağımlı değişken düzeyinin beş olması halinde kısıtlandırılmamış kısmi oransal odds modelinde örneklem büyüklüğü arttıkça doğru sınıflandırma oranlarının azaldığı sonucu elde edilmiştir.
- Kısıtlandırılmamış kısmi oransal odds modelinde bağımlı değişkenlerin tüm düzeyleri göz önünde bulundurulduğunda, bağımsız değişken sayısının bazı durumlarında doğru sınıflandırma oranının arttığı ortaya çıkmıştır.
- Kısıtlandırılmamış kısmi oransal odds modelinde bağımlı değişkenlerin tüm düzeyleri göz önünde bulundurulduğunda, bağımsız değişken sayısının beş olması halinde örneklem büyüklüğü arttıkça doğru sınıflandırma oranının arttığı azaldığı sonucu

elde edilmiştir. Bu durum bağımsız değişken sayısının üç ve yedi olması halinde elde edilmemiştir.

- Bağımlı değişken düzeyinin üç olması halinde bağımsız değişken sayısının üç, beş ve yedi olması durumunda örneklem büyüklüğü arttıkça kısıtlandırılmamış kısmi oransal odds modeli ve oransal odds modelinde doğru sınıflandırma oranları artmaktayken, doğrusal diskriminant analizi ve karesel diskriminant analizi için elde edilen doğru sınıflandırma oranları azalmaktadır. Bu durum bağımlı değişken düzeyinin beş ve yedi olması halinde de aynı sonuçları vermiştir.
- Örneklem büyüklüğünün küçük olduğu (n=100) bazı durumlarda doğrusal diskriminant analizinin karesel diskriminant analizine göre daha iyi bir performans sergilediği gözlemlenmiştir.
- Bağımlı değişken düzeyinin üç, beş ve yedi olduğu durumlarda en iyi performansı kısıtlandırılmamış kısmi oransal odds modeli sergilemiştir. Bununla birlikte kısıtlandırılmamış kısmi oransal odds modeli oransal odds modeline göre, karesel diskriminant analizi de doğrusal diskriminant analizine göre daha yüksek doğru sınıflandırma oranlarına sahip olmuştur.

6. Sonuçlar (Conclusion)

Bu çalışmada, bağımlı değişken düzeyi, bağımsız değişken sayısı ve örneklem büyüklüğü verildiğinde sınıflama tekniklerinden olan kısıtlandırılmamış kısmi oransal odds modeli, oransal odds modeli, karesel diskriminant analizi ve doğrusal diskriminant analizi bir simülasyon çalışması ile karşılaştırılmaya çalışılmıştır.

Modellerin sınıflandırma performanslarını değerlendirmek amacıyla farklı örneklem büyüklükleri ile çalışmalar yapılmış ve doğru sınıflama oranları ortaya koyulmuştur. Simülasyon çalışması ile bağımsız değişkenlerin sürekli olması durumunda elde edilen temel sonuçlar şu şekilde özetlenmektedir:

Analizde elde edilen sonuçlara göre, bağımlı değişkenin farklı düzeyleri ve her farklı düzeye ait farklı bağımsız değişken sayısı ele alındığında ilgilenilen dört model içerisinde kısıtlandırılmamış kısmi oransal odds modelinin her durumda en yüksek doğru sınıflama oranına sahip olan model olduğu ve bu modelin en iyi sınıflandırma performansına sahip olduğu sonucu elde edilmiştir.

Bilindiği üzere modele eklenen her yeni bir bağımsız değişken R^2 değerini (belirleme katsayısı); yani bağımsız değişkenlerce bağımlı değişkeni açıklama yüzdesini arttırmaktadır. Bu nedenle modele eklenecek her bir bağımsız değişkenin doğru sınıflama oranını arttırması beklenen bir durumdur. Analiz sonuçlarına göre, kısıtlandırılmamış kısmi oransal odds modeli ile oransal odds modelinde her bağımlı değişken düzeyinde bağımsız değişken sayısı arttırıldığında doğru sınıflama oranının arttığı sonucu elde edilmiştir. Buna rağmen, hem karesel hem de doğrusal diskriminant analizinde her bağımlı değişken düzeyinde bağımsız değişken sayısı arttırıldığında ise doğru sınıflama oranının azaldığı sonucu elde edilmiştir.

Oransal odds modeli ve diskriminant analizinin kendi içerisinde değerlendirilmesi istendiğinde, hemem hemen tüm durumlarda kısıtlandırılmamış kısmi oransal odds modelinin oransal odds modeline göre daha yüksek doğru sınıflama oranına sahip olduğu görülmektedir. Bununla birlikte karesel diskriminant analizi de doğrusal diskriminant analizine göre daha iyi bir sınıflama performansına sahiptir. Fakat örneklem büyüklüğünün küçük olduğu bazı durumlarda (n=100) doğrusal diskriminant analizinin karesel diskriminant analizine göre daha iyi sonuç verdiği ortaya çıkmıştır.

Kaynaklar (References)

[1] Abdel A. and Wang X. (2008). Analysis of left-turn crash injury severity by conflicting pattern using partial proportional odds models. Department of Civil&Environmental Engineering, University of Central Florida, Orlando, United States.

[2] Adeleke K.A. and Adepoju A.A. (2009). Ordinal logistic regression model: an application to pregnancy outcomes journal of mathematics and statistics. International Journal of Epidemilogy, Great Britain. 279-285, 2010 ISSN 1549-364.

[3] Ananth C. and Kleinbaum D.G. (1997). Regression models for ordinal responses: A review of methods and applications. International Journal of Epidemilogy, Great Britain.

[4] Chen C.K. and Hughes J. (2004). Using ordinal regression model to analyze student satisfaction questionnaires. Association for Institutional Research, Volume1.

[5] Dağlıoğlu H.(2014). Kısıtlandırılmamış Kısmi. Oransal Odds Modelinin Doğru Sınıflandırma

Performansı Üzerine Bir Çalışma, Gazi Üniversitesi, Fen Bilimleri Enstitüsü, Doktora Tezi, Ankara.

[6] Damodar G. (1995). Basic econometrics. İstanbul, Third Edition. 1995, pp.541.

[7] Fujimoto K. (2005). From women's college to work: inter-organizational networks in the Japanese female labor market. Social Science Research 34 (4), 651–681.

[8] Lall R., Campbell M. J., Walters S. J., Morgan K. (2002). A review of ordinal regression models applied on health-related quality of life assessments. Statistical Methods in Medical Research 11: 49–67.

[9] Liao T. F. (1994). Interpreting probability models: lojit, probit, and other generalized linear models. Sage Publications, Thousand Oaks, CA

[10] Long S. J. (1997). Regression models for categorical and limited dependent variables. Sage Publications, Thousand Oaks, CA.

[11] McCullagh P. (1980) Regression models for ordinal data. Journal of the Royal Statistical Society. Series B Volume 42, Issue 2, 109-142.

[12] McCullagh P. and Nelder, J. A. (1989). Generalized Linear Models, Second Edition, Chapman and Hall, London.

[13] Peterson B.L. (1986). Proportional odds and partial proportional odds models for ordinal response variables. Department of Biostatistics, University of North Coralino, at Chapel Hill Institute of statistics mimeo series no, October 1986

[14] Peterson B. and Harrell, F. E. (1990). Partial proportional odds models for ordinal response variables. Applied Statistics, 39, 205-217.

[15] Williams R. (2005) Gologit2: a program for generalized logistic regression/ partial proportional odds models for ordinal variables. Retrieved May 12, 2005.

[16] Williams R. (2006). Generalized ordered lojit/partial proportional odds models for ordinal dependent variables. Stata Journal 6 (1), 58–82

Hatice DAĞLIOĞLU[*]

Dr. Hatice DAĞLIOĞLU, 1983 yılı Erzurum doğumludur. 2006 yılı Ege Üniversitesi İstatistik bölümünde lisans, 2008 yılı Hacettepe Üniversitesi Fen Bilimleri Enstitüsü İstatistik Anabiliminde de master derecesini almıştır. 2014 yılında Gazi Üniversitesi Fen Bilimleri Enstitüsü İstatistik Anabilim dalında doktorasını tamamlamıştır. 2006-2008 yılları arasında TÜBİTAK-2210 yurt içi yüksek lisans bursu, 2008-2014 yılları arasında da TÜBİTAK-2211 yurt içi doktora bursunu almıştır. 2008-2010 yılları Ziraat Bankası Genel Müdürülüğünde görev aldıktan sonra, 2010 yılında TÜBİTAK SAGE'de çalışmaya başlamıştır. Halen aynı kurumda çalışmalarına devam etmektedir.