Kocaeli Üniversitesi **Sosyal Bilimler** Enstitüsü **Dergisi**, 2004 / 2 : 185-208

Lojistik Regresyon Analizi: Tıp Verileri Üzerine Bir Uygulama

Hüdaverdi Bircan*

Özet: Bu çalışmanın amacı, ikili sonuç değişkeni ile hem sürekli hem de kesikli değişkenlerden oluşan bağımsız değişkenler kümesi arasındaki ilişkiyi tanımlayabilen lojistik regresyon analizinin incelenmesidir. Lojistik regresyon analizine bir uygulama göstermek amacıyla, çocuklarda doğum ağırlığını etkileyen önemli risk faktörlerini belirlemek için Tıp verileri üzerinde çalışılmıştır. Lojistik modele dahil edilecek bağımsız değişkenler, tek değişkenli lojistik regresyon analiziyle belirlendikten sonra, çok değişkenli modele dahil edilen her bir değişkenin önemliliği gösterilmelidir. Tek değişkenli modelde önemli bulunduğu halde çok değişkenli modelde önemsiz olan değişkenler model dışı bırakılmıştır. Böylece final model elde edilmiştir. Final modelin hem biyolojik olarak kabul edilebilir, hem de doğru sınıflama oranının yeteri kadar iyi olmasından dolayı, bu modelin risk faktörlerini belirleme de kullanılabileceği sonucuna varılmıştır.

Anahtar Kelimeler: Lojistik Regresyon Analizi, Odds oranı

Giriş

Gözlemleri verilerin yapısında bulunan olası gruplara atamak için birkaç yöntem vardır. Bu yöntemler;

1. Kümeleme
2. Diskriminant
3. Lojistik Regresyon Analizi

* Yrd. Doç. Dr. Hüdaverdi Bircan Cumhuriyet Üniversitesi İşletme Bölümünde öğretim üyesidir.

Kümeleme analizinde; verilerin yapısındaki grup sayısı bilinmemekte, gözlemler uzaklık veyahut benzerlik ölçütlerine göre kümelenmektedir. Burada amaç yalnızca gözlemlerin oluşturduğu kümenin yapısını bulmaktır.

Diskriminant ve Lojistik Regresyon Analizinde ise verilerin yapısındaki grup sayısı bilinmekte ve bu verilerden faydalanarak bir ayrımsama modeli elde edilmektedir. Kurulan bu model yardımı ile veri kümesine yeni alınan gözlemlerin gruplara atanması yapılmaktadır (Başarır, 1990: 1).

Lojistik Regresyon Analizinin kullanım amacı, istatistikte kullanılan diğer model yapılandırma teknikleri ile aynıdır. En az değişkeni kullanarak en iyi uyuma sahip olacak şekilde bağımlı ile bağımsız değişkenler arasındaki ilişkiyi tanımlayabilen ve biyolojik olarak kabul edilebilir bir model kurmaktır.

Lojistik modelin biyolojik deneylerin analizi için kullanımı ilk olarak Berkson (1944) tarafından önerilmiş, Cox (1970) bu modeli gözden geçirerek çeşitli uygulamalarını yapmış, özet gelişmeler ise ilk Andersson (1979, 1983) tarafından verilmiştir. Ayrıca verilerin lojistik modele uyumu ile ilgili birçok çalışmalar da yapılmıştır. Bunlar arasında Aranda-Ordaz (1981) ve Johnson(1985) tarafından yapılan çalışmalar en önemlileridir. Pregibon (1981) iki grup lojistik modelde etkin (influential), aykırı (outlier) gözlemleri ve belirleme ölçütlerini (diagnostic), Lesaffre (1986), Lesaffre ve Albert (1989) ise çoklu grup lojistik modellerde etkin ve aykırı gözlemlerle belirleme ölçütlerini incelemişlerdir.

Lojistik regresyon modellerinin yaygın bir şekilde kullanılır hale gelmesi, katsayı tahmin yöntemlerinin geliştirilmesi ve lojistik regresyon modellerinin daha ayrıntılı incelenmesine sebep olmuştur. Cornfield (1962), lojistik regresyondaki katsayı tahmin işlemlerinde diskriminant fonksiyonu yaklaşımını ilk kez kullanarak popüler hale getirmiştir. Lee (1984) basit dönüşümlü (cross-over) deneme planları için linear lojistik modeller üzerinde durmuştur. Bonney (1987) lojistik regresyon modelinin kullanımı ve geliştirilmesi üzerinde çalışmıştır. Robert ve ark. (1987) lojistik regresyonda standart Kikare, olabilirlik oran (G^2), "pseudo" en çok olabilirlik tahminleri, uyum mükemmelliği ve hipotez testleri üzerine çalışmalar yapmışlardır. Duffy (1990) lojistik regresyonda hata terimlerinin dağılışı ve parametre değerlerinin gerçek değerlere yaklaşımını incelemiştir. Başarır (1990) klinik verilerde çok değişkenli lojistik regresyon analizi ve ayrımsama sorunu üzerinde çalışmıştır. Hsu ve Leonard (1995) lojistik regresyon fonksiyonlarında Bayes tahminlerinin elde edilmesi işlemleri üzerine çalışmışlar ve lojistik regresyonda Monte Carlo dönüşümünün kullanılabileceğini göstermişlerdir. Akkaya ve Pazarlıoğlu (1998) lojistik regresyon modellerinin ekonomi alınında kullanımını örneklerle incelemişlerdir. Cox ve ark. (1998) kardiovasküler hastalıklar ve hipertansiyon arasındaki ilişkiyi incelemişlerdir.

Lojistik regresyon modelleri, son yıllarda biyoloji, tıp, ekonomi, tarım ve veterinerlik ve taşıma sahalarında yaygın olarak kullanılmaktadır.

Gardside ve Glueck (1995) insanlarda beslenme şekli, sigara ve alkol kullanımı, fiziksel aktivite gibi risk faktörlerinin kalp hastalığı üzerindeki etkilerini incelemiştir.

Kloiber ve ark (1996), Peoples ve ark. (1991), Buescher ve ark. (1993) kadınlarda düşük doğum ağırlığını etkileyen risk faktörlerini; Santos ve ark. (1998) kafein tüketimi ve düşük doğum ağırlığı arasındaki ilişkiyi; Sable ve Herman (1997) erken doğum ve düşük doğum ağırlığı arasındaki ilişkiyi incelemişlerdir.

Çeşitli varsayım bozulmaları olduğunda Lojistik Regresyon Analizi, diskriminant analizi ve çapraz tablo uygulamalarına alternatif olarak uygulanmaktadır. Kullanım nedeni olarak en temel yaklaşım doğrusal regresyon analizinde yapılabilir; bağımlı değişken 0 ve 1 gibi ikili (binary) ya da ikiden çok kategori içeren kesikli değişken olduğunda normallik varsayımı bozulmakta ve doğrusal regresyon analizi uygulanamamaktadır.

Lojistik regresyonu doğrusal regresyondan ayıran en belirgin özellik ise lojistik regresyonda sonuç değişkenin ikili veya çoklu olmasıdır. Lojistik regresyon ve doğrusal regresyon arasındaki bu fark hem parametrik model seçimine, hem de varsayımlara yansımaktadır.

Lojistik regresyonda da, doğrusal regresyon analizinde olduğu gibi bazı değişken değerlerine dayanarak tahmin yapılmaya çalışılır. Ancak bu iki yöntem arasında üç önemli fark vardır (Elhan, A.H. 1997: 4):

1. Doğrusal regresyon analizinde tahmin edilecek olan bağımlı değişken sürekli iken, Lojistik Regresyon Analizinde bağımlı değişken kesikli bir değer almaktadır.

2. Doğrusal regresyon analizinde bağımlı değişkenin değeri, Lojistik Regresyon Analizinde ise bağımlı değişkenin alabileceği değerlerden birinin gerçekleşme olasılığı tahmin edilir.

3. Doğrusal regresyon analizinde bağımsız değişkenin çoklu normal dağılım göstermesi şartı aranırken, Lojistik Regresyon Analizinde böyle bir şart yoktur.

1. Materyal ve Metot

Bu çalışmada, materyal olarak Sivas ili Doğumevi hastanesinde 2004 yılında yapılan anket çalışması sonucu elde edilen veriler kullanılmıştır. Çocuklarda doğum ağırlığını etkileyen risk faktörleri belirlenmiş, buna uygun olarak bir anket formu geliştirilmiştir. Anket çalışması 3 aylık bir dönemde yapılmıştır. Ankete katılan 250 gebe kadından çalışmamıza uygun veriler ayıklanarak 128 gebe kadına ait veriler analizde kullanılmıştır.

Normal bir doğum ağırlığı 2500gr-4500gr aralığında değişmektedir. Doğum ağırlığı 2.5 kg'a eşit ve küçük olanlar 1, 2.5 kg üzerindeki bebekler ise 2 şeklinde kodlanmıştır. Bağımlı değişken dolayısıyla iki şıklıdır. Bebeğin doğum ağırlığını etkileyen bir çok faktör vardır. Bunların bazıları kontrol altına alınabilir faktörlerdir. Doğum ağırlığını etkileyen veya düşük doğum ağırlığına neden olan risk faktörlerinin, doğum ağırlığını hangi düzeyde etkilediğinin bilinmesi gerekmektedir. Bu yolla kontrol altına alınabilir risk faktörleri dikkate alınarak düşük doğum ağırlığı oranı en aza indirilebilir.

"Doğum ağırlığını" etkileyen risk faktörleri olabilecek değişkenler aşağıda verilmiştir:

Cinsiyet, (CINS): Kesikli değişkendir, 1= Kız, 2= Erkek şeklinde kodlanmıştır.

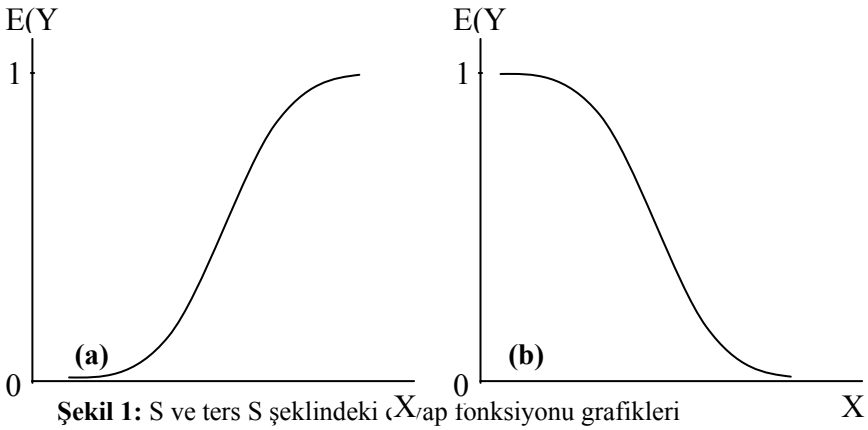
1. Anne Yaşı, (YAS): Süreli değişkendir.
2. Anne Boyu, (BOY): Sürekli değişkendir.
3. Gebelik Haftası, (GHAFTA): Kesikli değişkendir. 1 = 38 Hafta ≤ ; 2 = 38 Hafta >
4. Gebelik Öncesi Anne Kilosu, (GOAG): Sürekli değişkendir.
5. Gebelikte Alınan Kilo, (GAK): Sürekli değişkendir.
6. Son Doğumdan Buyana Gecen Süre, (SDGS): Sürekli değişkendir.
7. Annenin Hemoglobin Değeri, (HD): Sürekli değişkendir.
8. Annenin Hematokrit Değeri, (HTD): Sürekli değişkendir.
9. Annenin Çalışıp Çalışmadığı, (CDUR): Kesikli değişkendir, 1=Çalışıyor, 2=Çalışmıyor.
10. Annenin İstirahat Süresi, (ISUR)
11. Annenin Sigara Kullanıp Kullanmadığı, (SIG): Kesikli değişkendir, 1= Kullanıyor, 2= Kullanmıyor.
12. Annenin Alkol Kullanıp Kullanmadığı, (ALK): Kesikli değişkendir, 1=Kullanıyor, 2=Kullanmıyor.
13. Gravitası, (GR): Kesikli değişkendir.
14. Paritesi, (PR): Kesikli değişkendir.
15. Beslenme Durumu, (BES): Kesikli değişkendir, 1= Düzenli, 2= Düzensiz
16. Doğum Şekli, (DSEK): Kesikli değişkendir, 1= Normal, 2= Sezeryan.

Annenin Hemoglobin ve Hematokrit değerleri tüm gebe kadınlar için elde edilmediğinden bu değişkenler analiz dışı bırakılmıştır. Annenin alkol kullanma durumu ise, tüm fertler alkol kullanmadığından analiz dışı bırakılmıştır. Ayrıca Annenin istirahat süresi de, tüm bireyler için elde edilemediği için analiz dışı bırakılmıştır.

Lojistik regresyon modelleri zayıf ölçekle ölçülmüş değişkenler arasındaki ilişkinin şeklini ortaya koyan modellerdir. Yapılan bir çok çalışmada bağımlı değişken

sadece iki sonuca sahiptir. Genellikle üzerinde durulan olayın gerçekleşmesi 1 gerçekleşmemesi ise 0 ile gösterilir.

Hem teorik hem de deneysel incelemeler bağımlı değişken iki sonuçlu iken cevap fonksiyonunun şeklinin S veya ters S şeklinde olacağını göstermiştir. Bağımlı değişken, Şekil 1a ve 1b'de de görüldüğü gibi bitiş noktaları dışında yaklaşık olarak doğrusaldır. Bu cevap fonksiyonları 0 ile 1 değerlerinde X ve Y eksenlerine asimptottur.



Şekil 1: S ve ters S şeklindeki cevap fonksiyonu grafikleri

Şekil 1'de gösterilen cevap fonksiyonları, lojistik cevap fonksiyonları olarak bilinir. Lojistik fonksiyonun 0 ile 1 arasında bir değişim aralığına sahip olması lojistik fonksiyonun tercih edilmesindeki ilk önemli nedendir. Lojistik model, ortaya çıkacak riski 0 ile 1 arasında herhangi bir değer olarak tahmin etmeye yarar. Başka bir deyişle 1'in üstünde veya 0'ın altında bir risk olmaz. Bu durum her model için her zaman doğru olmamaktadır (Hosmer – Lemeshow, 1980: 1043-1069).

Araştırmacı bazen bağımsız değişkenler üzerinde denetime sahiptir. Böyle bir imkan söz konusu olduğunda X_i değerlerine karşılık gelen hücrelerdeki birim sayısının asgari 30 olması regresyonun verilere uygunluğunu önemli ölçüde yükseltir. (Pindyck – Rubinfeld, 1985: 273)

Lojistik regresyon fonksiyonu,

$$\pi(x) = \frac{\exp(\beta_0 + \beta_1 X)}{1 + \exp(\beta_0 + \beta_1 X)} \qquad 1$$

şeklindedir. Bu ifadenin bir diğer şekli ise,

$$\pi(x) = \left[1 + \exp(-\beta_0 - \beta_1 X)\right]^{-1} \qquad 2$$

olarak yazılabilir (Agresti, 1996: 103). $\pi(x) = E(Y/x)$ değeri şartlı ortalama olarak bilinir.

Şartlı ortalamanın, modelde yer alan parametrelerle ($\beta_o+\beta_1$) doğrusal hale dönüştürülmesi için, transformasyona tabi tutulması gerekir. Bu transformasyona Logit transformasyon adı verilir ve aşağıdaki şekilde gösterilir:

$$g(x) = \ln\left[\frac{\pi(x)}{1-\pi(x)}\right] = \beta_0 + \beta_1 x \qquad 3$$

Transformasyon değişkeni g(x), modeldeki parametrelerle doğrusaldır, süreklidir ve $-\infty$, $+\infty$ aralığında değişen değerler alır. $\pi(x)$ arttıkça g(x)'te artar ve eğer $\pi(x)<0.5$ ise g(x) negatif, $\pi(x)>0.5$ ise g(x) pozitif değerler alır (Hosmer ve Lemeshow, 1989:307).

Modelin sonuç değişkeninin sınırlarını genişletmek için uygulanan Logit transformasyonunun bazı özellikleri şöyle sıralanabilir:

(1) p arttıkça lojit(p) de artar.

(2) p, 0 ile 1 arasında iken lojit(p) reel sayılar doğrusu üzerinde değerler alabilir.

(3) $p < 0.5$ olduğunda lojit(p) < 0 ve $p > 0.5$ olduğunda lojit(p) > 0 olur.

Doğrusal regresyon modelinde bağımlı değişkene ait bir gözlem $y = E(Y/x) + \varepsilon$ şeklinde gösterilebilir. ε hata terimi olarak isimlendirilir ve gözlemin koşullu olasılıktan ne kadar saptığını gösterir. ε'nin ortalamasının sıfır ve varyansının ise bağımsız değişkenin her düzeyinde sabit olacak şekilde normal dağılım göstereceği genel bir varsayımdır. Bu varsayım bağımlı değişken iki düzey içerdiği zaman geçerli değildir. Bu tür durumlarda x verildiğinde sonuç değişkeninin değeri $y = \pi(x) + \varepsilon$ ile gösterilir. Ve ε'nin mümkün olan iki değerden başka değer alamayacağı varsayılır. Eğer y = 1 ise, $\pi(x)$ olasılıkla $\varepsilon = 1 - \pi(x)$ değerini alır ve eğer y = 0 ise, $1 - \pi(x)$ olasılıkla $\varepsilon = - \pi(x)$ değerini alır. Böylece ε, sıfır ortalamalı ve $\pi(x)[1 - \pi(x)]$'a eşit varyanslı binomiyal bir dağılım gösterir.

1.1. Lojistik Regresyon Analizinde Parametre Tahmini

Lojistik modelde parametrelerin tahmin edilmesi için çeşitli yöntemler ortaya atılmıştır. Bu çalışmada parametrelerin tahmin edilmesinde en çok olabilirlik (maximum likelihood) tahmin yöntemi kullanılacaktır.

Genel olarak en çok olabilirlik yöntemi, gözlenen veri kümesini elde etmenin olasılığını maksimum yapan bilinmeyen parametrelerin değerlerini verir. Bu metodu uygulamak için öncelikle, en çok olabilirlik fonksiyonunun oluşturulması gerekmek-

tedir. Bu fonksiyon gözlenen verilerin olasılıklarını, bilinmeyen parametrelerin bir fonksiyonu olarak açıklar. Bu parametrelerin en çok olabilirlik tahmin edicileri, fonksiyonu maksimum yapan değerleri bulacak şekilde seçilir. Böylece sonuçta elde edilen tahminleyiciler, gözlenen verilerle çok yakın değerlere sahiptir.

Eğer y, 0 ve 1 olarak kodlandıysa, bu durumda 1 numaralı eşitlikte verilen $\pi(x)$ ifadesi, verilen x değeri için y'nin 1'e eşit olma koşullu olasılığını vermektedir. Bu olasılık $\pi(x) = P(y = 1/x)$ sembolüyle gösterilir. Buradan hareketle, $[1-\pi(x)]$ ifadesi de, y'nin 0 değerini alma koşullu olasılığını göstermektedir. Bu olasılık da $[1-\pi(x)] = P(y = 0/x)$ şeklinde gösterilir. (x_i, y_i) çifti için $y_i = 1$ olduğunda olabilirlik (likelihood) fonksiyonuna katkısı $\pi(x_i)$ iken $y_i = 0$ olduğunda olabilirlik fonksiyonuna katkısı $1- \pi(x_i)$ kadardır. (x_i, y_i) çiftinin olabilirlik foksiyonuna katkısını ifade etmenin güvenilir bir yolu aşağıda verilmiştir:

$$\zeta(x_i) = \pi(x_i)^{y_i}[1-\pi(x_i)]^{1-y_i} \qquad 4$$

Gözlemlerin birbirinden bağımsız oldukları varsayıldığı için, olabilirlik fonksiyonu Eşitlik 4'deki terimlerin çarpılmasıyla elde edilir.

$$\iota(\beta) = \Pi\zeta(x_i) \qquad 5$$

En çok olabilirliğin temel ilkesinde β kestiriminin Eşitlik 5'deki ifadeyi maksimum yaptığı vurgulanır. Matematiksel olarak Eşitlik 5'in logaritmasıyla çalışmak daha kolay olacağından log-olabilirlik fonksiyonu aşağıdaki gibi elde edilir:

$$L(\beta) = \ln[\iota(\beta)] = \sum_{i=1}^{n}\{y_i \ln[\pi(x_i)] + (1-y_i)\ln[1-\pi(x_i)]\} \qquad 6$$

$L(\beta)$'yi maksimum yapan β değerlerini bulmak için, $L(\beta)$'nin β_0 ve β_1'e göre türevi alınarak sıfıra eşitlenir. Elde edilecek eşitlikler:

$$\sum_{i=1}^{n}[y_i - \pi(x_i) = 0 \qquad 7$$

$$\sum_{i=1}^{n}x_i[y_i - \pi(x_i) = 0 \qquad 8$$

şeklindedir. Bu eşitlikler olabilirlik eşitlikleri (likelihood equations) olarak adlandırılır.

Lineer regresyon analizinde β'ya göre türevinden elde edilen olabilirlik eşitlikleri, bilinmeyen parametreleri içeren doğrusal ifadelerdir, bu nedenle kolayca çözümlenebilir. Lojistik regresyon için elde edilen 7 ve 8 Eşitlikleri β_0 ve β_1'de lineer değildir. Bundan dolayı bu eşitliklerin çözümlenmesi için özel yöntemlere ihtiyaç vardır. Bu denklemlerin çözümleri genelleştirilmiş ağırlıklı en küçük kareler yöntemi ile elde edileceği gösterilmiştir. Bu metotlar iteratiftir (Hosmer ve Lemeshow: 1989).

1.2. Lojistik Regresyon Parametrelerinin Önem Testi

Lojistik regresyonda gözlenen ve beklenen değerlerin karşılaştırılması log olabilirlik fonksiyonu ile yapılmaktadır.

$$D = -2\ln\left[\frac{\text{Şu andaki modelin olabilirligi}}{\text{Doymuş modelin olabilirliği}}\right] \qquad 9$$

Eşitlik 9'da parantez içerisinde verilen ifade olabilirlik oranı "likelihood ratio" olarak adlandırılır. (–2ln) katının alınması, matematiksel olduğu kadar dağılımı bilinen bir değer elde etmektir. Bu değer hipotez testi amacıyla kullanılmaktadır. Böyle bir teste olabilirlik oran testi adı verilmektedir. Eşitlik 2.6 kullanılarak, Eşitlik 10 aşağıdaki gibi verilir:

$$D = -2\sum_{i=1}^{n}\{y_i \ln\left(\frac{\hat{\pi}_i}{y_i}\right) + (1-y_i)\ln\left(\frac{1-\hat{\pi}_i}{1-y_i}\right) \qquad 10$$

Bağımsız bir değişkenin önemine karar vermek için denklemde bağımsız değişkenin olduğu ve olmadığı durumlardaki D değerleri karşılaştırılır. Bağımsız değişkeni kapsamasından dolayı ortaya çıkan D'deki değişim aşağıdaki gibidir:

G = D(Değişkensiz Model için) – D(Değişkenli Model için) 11

Hesaplanan bu istatistik de, doğrusal regresyonda kullanılan F testindeki pay kısmı ile aynı rolü üstlenir. G'yi hesaplamak için farkı alınacak D değerlerinin her ikisi için de doymuş modelin olabilirlikleri ortak olduğundan G istatistiği aşağıdaki şekli alır:

$$G = -2\ln\left[\frac{\text{Değişkensiz modelin olabilirliği}}{\text{Değişkenli modelin olabilirliği}}\right] \qquad 12$$

Tek bağımsız değişkenli özel durumlarda, değişkenin modelde olmadığı zamanda ki β_0'ın ençok olabilirlik tahmini $\ln(n_1/n_0)$'dır. ($n_1 = \Sigma y_i$ ve $n_0 = \Sigma(1-y_i)$). Tahmin değeri sabittir, n_1/n. G istatistiği aşağıdaki gibi hesaplanır:

$$G = -2\ln\left[\frac{\left(\frac{n_1}{n}\right)^{n_1}\left(\frac{n_0}{n}\right)^{n_0}}{\prod_{i=1}^{n}\hat{\pi}_i^{y_i}(1-\hat{\pi}_i)^{(1-y_i)}}\right] \qquad 13$$

ya da

$$G = 2\left\{\sum_{i=1}^{n}\left[y_i\ln(\hat{\pi}_i) + (1-y_i)\ln(\hat{\pi}_i)\right] - \left[n_1\ln(n_1) + n_0\ln(n_o) - n\ln(n)\right]\right\} \qquad 14$$

dır. $\beta_1 = 0$ hipotezi altında, G istatistiği 1 serbestlik derecesiyle χ^2 dağılımı gösterir. Tüm değişkenleri içeren model ile kestirilen modele ilişkin olabilirlik oran değerlerinin farkına dayanan ölçütlerin Ki-kare dağılacağı düşüncesinden hareketle kurulan modelin geçerliliği sınanmaktadır (Elhan, A.H. 1997: 10).

1.3. Çoklu Lojistik Regresyon Modeli

Çoklu lojistik regresyonda, bağımsız değişkenler değişik ölçüm biçimlerinde olabilmektedir. Kesikli ve nominal ölçekli bağımsız değişkenleri modele dahil etmek için dizayn değişkenleri kullanılması gerekir. Öncelikle modeldeki tüm bağımsız değişkenlerin her birinin en az aralık ölçekli olduğunu düşünelim.

X' = ($x_1, x_2, x_3,.....,x_p$) vektörü ile gösterilsin. Sonuç değişkeninin mevcut olduğu (Y=1) zaman ki koşullu olasılık, P(Y = 1/x) = π(x)'e eşit olacaktır. Çoklu lojistik regresyon modelinin logiti aşağıdaki denklem ile gösterilir:

$$g(x) = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p \qquad 15$$

Bu durum da,

$$\pi(x) = \frac{e^{g(x)}}{1 + e^{g/(x)}} \qquad 16$$

olarak bulunmuştur.

Bağımsız değişkenler kesikli, nominal ölçekli ise, o zaman bu değişkenler yerine dizayn (kukla) değişkenlerinin bu değişkenleri temsil etmesi için kullanılması gerekir. Genel olarak nominal değişken k kategoriye sahipse, o zaman k-1 dizayn değişkenine ihtiyaç vardır. J. Bağımsız değişken (x_j), k_j kategoriye sahip olsun. K_j –1 dizayn değişkeni D_{ju} olarak ve katsayıları da β_{ju}, u = 1,2,.....,k_j – 1 olarak belirtilirse, j. Değişken kesikli olan p değişkenli model için logit aşağıdaki gibidir:

$$g(x) = \beta_0 + \beta_1 X_1 + ... + \sum_{u=1}^{k_j - 1} \beta_{ju} D_{ju} + \beta_p X_p \qquad 17$$

Birbirinden bağımsız n tane (x_i, y_i), i=1,2,....,n gözlem çiftinin olduğu olduğunu düşünelim. Tek değişkenli modelde olduğu gibi modelin kurulması için tahmin vektörünün $\beta' = (\beta_0, \beta_1, ..., \beta_p)$ elde edilmesi gerekir. Çok değişkenli durumda, tek değişkenli durumda olduğu gibi tahmin metodu en çok olabilirlik metodu olacaktır. Olabilirlik fonksiyonu Eşitlik 5'de verildiği gibidir. Tek değişiklik p(x)'in Eşitlik 16'da tanımlandığı gibi olmasıdır. Log olabilirlik fonksiyonu p+1 katsayıya göre türevi alınarak, 7 ve 8 numaralı eşitliklere benzer p+1 tane olabilirlik denklemi elde edilir.

$$\sum_{i=1}^{n} [y_i - \pi(x_i) = 0 \qquad 18$$

$$\sum_{i=1}^{n} x_{ij} [y_i - \pi(x_i) = 0 \qquad j=1,2,3,...,p \qquad 19$$

$\hat{\beta}$, 19 ve 20 numaralı eşitliklerin çözümünü göstersin. Çoklu lojistik regresyon modelinde tahmin edilen değerler $\pi(\hat{x}_i)$ dir. Eşitlik 16'daki ifadenin değeri, $\hat{\beta}$ ve x_i'yi kullanarak hesaplanmıştır. En çok olabilirlik tahmin teorisi, log olabilirlik fonksiyonunun ikinci dereceden türevlerinden oluşan matristen tahmin değerlerinin elde edileceğini vurgular.

Logaritmik olabilirlilik fonksiyonunun β_0, β_1, ..., β_{p-1} parametrelerine göre ikinci dereceden kısmi türevlerinin matrisini G ile gösterelim. G matrisi,

$$\underset{pxp}{G} = (g_{ij}) \qquad i= 0, 1, 2, ..., p-1; j = 0, 1, 2, ..., p-1 \qquad 20$$

şeklinde gösterilebilir. Bu şekilde

$$g_{00} = \frac{\partial^2 \log_e L(\beta)}{\partial \beta_0^2}, \quad g_{01} = \frac{\partial^2 \log_e L(\beta)}{\partial \beta_0 \partial \beta_1} \quad , ... , \qquad 21$$

değerleri elde edilir. Bu matris, Hessiyan matrisi olarak adlandırılır. Hessiyan matrisdeki ikinci derece kısmi türevleri, $\beta = b$ olarak; yani, en çok olabilirlilik tahmincileri olarak görmek gerekir. En yüksek ihtimal tahmini için kullanıldığında Lojistik regresyondan tahmin edilen regresyon katsayılarının tahmini yaklaşık varyans-kovaryans matrisi şu eşitlikten elde edilir.

$$s^2\{b\} = \left[(-g_{ij})_{\beta=b}\right]^{-1} \qquad 22$$

Örnek hacmi yeterince büyük olduğunda basit veya çoklu lojistik regresyon modellerindeki regresyon katsayılarının anlamlı olup olmadığını test ederken aşağıdaki değere bağlı olarak karar verilir.

$$\frac{b_k - \beta_k}{s\{b_k\}} \approx Z, \qquad k = 0, 1, 2, ..., p\text{-}1 \qquad 23$$

Formül 23'deki Z değeri standart normal değerdir. $S\{b_k\}$ değeri Formül 22'den elde edilen b_k'nın tahmini standart sapmasıdır.

Çoğu kez çoklu lojistik regresyon modelindeki X değişkenlerinin alt grupları ile ilişkili regresyon katsayılarının önemli olup olmadığı araştırılır. Kullanılacak test prosedürü en çok olabilirlilik tahmininin genelleştirilmiş bir şeklidir. Büyük örnekler durumunda uygulanabilen bu test olabilirlilik oranı testi olarak adlandırılır. Genel model,

$$\pi = \left[1 + \exp\left(-\beta' X\right)\right]^{-1} \qquad 24$$

şeklindedir. Bu modelde,

$$\beta' X = \beta_0 + \beta_1 X_1 + ... + \beta_{p-1} X_{p-1} \qquad 25$$

olur. Model için bulunacak en çok olabilirlilik tahminlerini b_F ile gösterelim. Olabilirlilik fonksiyonunu $L(\beta)$ ile gösterdiğimizde $\beta = b_F$ olur. Genel modelde olabilirlilik fonksiyonunun bu değerini L(F) ile gösterelim.

Test edilecek hipotezler,

H_0: $\beta_q = \beta_{q+1} = ... = \beta_{p-1} = 0$

H_1: En az bir β_k değeri sıfırdan farklıdır.

En son p-q katsayılarını test etmek için model düzeltilir. Kısaltılmış model,

$$\pi = \left[1 + \exp\left(-\beta_R' X\right)\right]^{-1} \qquad 26$$

şeklindedir. Bu modelde

$$\beta_R' = \beta_0 + \beta_1 X_1 + ... + \beta_{q-1} X_{q-1} \qquad 27$$

olarak yazılır.

Şimdi kısaltılmış model için maksimum olabilirlilik tahminlerinin elde edilişi izah edilecektir. Maksimum olabilirlilik tahminleri b_R ile gösterilir. $\beta_R = b_R$ olduğunda q adet parametre ihtiva eden kısaltılmış model için olabilirlilik tahmini tarif ede-

biliriz. Olabilirlilik fonksiyonunun bu değeri L® ile gösterilecektir. L® değeri L(F) değerini hiçbir zaman geçemez. Bu sebeple L® değeri L(F) değerine yaklaştığında ilave parametreler olabilirliliği fazlaca artırmayacakları için H_0 hipotezinin doğru olduğuna karar verilir. L® değeri L(F) değerinden yeterince küçük olursa H_1 hipotezinin doğru olduğuna karar verilir.

Test istatistiği X^2 gösterildiğinde,

$$X^2 = 2\log_e\left[\frac{L(R)}{L(F)}\right] = -2\left[\log_e L(R) - \log_e L(F)\right] \qquad 28$$

olur. Örnek hacmi yeterince büyük olduğunda H_0 hipotezi doğru ise X^2 istatistiği yaklaşık olarak $\chi^2_{(1-\alpha;p-q)}$ şeklinde dağılım gösterir. Serbestlik derecesi, v = (n – q) –(n–p) şeklindedir. Böylece $X^2 \le X^2_{(1-\alpha;p-q)}$ olduğunda H_0 kabul edilirken $X^2 > X^2_{(1-\alpha;p-q)}$ olduğunda H_1 kabul edilir.

Regresyon katsayılarının önemli olup olmadığını test etmede kullanılabilecek ikinci test Wald testidir. Wald testine ait test istatistiğinin dağılımı standart normal dağılıma yaklaşır. Her değişken için listedeki standart hatalar kullanılarak Z testi yapılır. Wald testi, örnek hacminin büyük olması durumunda anlam kazanır. (Buse, 1982: 153)

Eğim parametresinin en yüksek ihtimal tahmincisi standart hatasının tahmini değeri ile mukayese edilir. $\beta_1 = 0$ iken test istatistiğinin dağılımı standart normal dağılıma uygundur. Bu teste ait test istatistiği,

$$W = \frac{\hat{\beta}_1}{S\hat{E}(\hat{\beta}_1)} \qquad 29$$

formülü ile elde edilir.

1.4. Lojistik Regresyon Katsayısının Yorumlanması

Lojistik cevap fonksiyonunda tahmin edilen b_1 regresyon katsayısının yorumlanması lineer bir regresyon modelindeki kadar kolay değildir. X eksenindeki başlangıç noktasına göre hazırlanan lojistik regresyon modelinde X değişkenindeki bir birimlik artışın tesirini ölçmek zordur. B_1 katsayısı yorumlanırken X'deki bir birimlik artış için $\frac{\pi}{1-\pi}$ odds tahmini ile $\exp(b_1)$ çarpılarak elde edilen lojistik cevap fonksiyonundan yararlanılır.

$X = X_j$ değeri için $\hat{\pi}'(X_j) = b_0 + b_1 X_j$ yazılırsa [Burada $\hat{\pi}'(X_j)$ ifadesi, tahmini değerle ilişkili X seviyesini göstermektedir.] $X = X_j + 1$ için $\hat{\pi}'(X_j + 1) = b_0 + b_1(X_j + 1)$ olarak yazılabilir. İki tahmin arasındaki fark $\hat{\pi}'(X_j + 1) - \hat{\pi}'(X_j) = b_1$ şeklindedir.

Bir önceki eşitliğe göre $\hat{\pi}'(X_j)$ 'nin değeri $X = X_j$ iken tek sayılar oranı (odds) tahmininin tabii logaritmasıdır. Yani $\log_e(\text{odds}_1)$ 'dir. Aynı şekilde $\hat{\pi}'(X_j + 1)$ 'in değeri de $X = X_j + 1$ iken $\log_e(\text{odds}_2)$ 'dir. Buradan:

$$\log_e(\text{odds}_2) - \log_e(\text{odds}_1) = \log_e \frac{\text{odds}_2}{\text{odds}_1} = b_1 \qquad 30$$

yazılabilir.

Lojistik modeldeki etkiler odds'a dayanır. X'in bir değerinde kestirilen odds'un, diğer değerinde kestirilen odds'a oranı olarak verilmektedir. Bu istatistik x=1 olan bireylerin x=0 olan bireylere nazaran bağımlı değişkenin kaç kat daha fazla 1 olarak görüldüğü sonucunu verir.

1.5. Uyum İyiliği İstatistikleri

Kurulan modelin uyum iyiliği testi Hosmer-Lemeshow'un hem onlu risk grupları hem de sabit kesim noktası yöntemine göre hesaplanmaktadır.

Uyum iyiliğine karar vermek için onlu risk grupları yöntemine göre hesaplanmak isteniyorsa, Hosmer-Lemeshow $\hat{C}_g^*$ istatistiği hesaplanır.

$$\hat{C}_g^* = \sum_{k=0}^{1} \sum_{l=1}^{10} \frac{(o_{kl} - e_{kl})^2}{e_{kl}} \qquad 31$$

Hosmer-Lemeshow $\hat{C}_g^*$ istatistiği, t-2 serbestlik dereceli ki-kare dağılımı göstermektedir.

Kestirilen modelin uyum iyiliği testi sabit kesim noktası yöntemiyle hesaplanmak istendiğinde ise, Hosmer-Lemeshow $\hat{H}_g^*$ istatistiği kullanılmaktadır.

$$\hat{H}_g^* = \sum_{k=0}^{1}\sum_{l=1}^{10}\frac{(o'_{kl} - e'_{kl})^2}{e``_{kl}} \qquad 32$$

Hosmer-Lemeshow $\hat{H}_g^*$ istatistiği, t-2 serbestlik dereceli ki-kare dağılımı göstermektedir.

2. Bulgular ve Tartışma

Tablo 2.1'de bağımlı değişken olan doğum ağırlığı ile ilişkili olabileceği düşünülen olası değişkenlerin tek değişkenli lojistik regresyon sonuçları verilmiştir.

Tablo 2.1'de, verilen değişkenlere ait modelde yalnız o değişken bulunurken kestirilen eğim katsayısı ($\hat{\beta}$), kestirilen eğim katsayısının standart hatası ($\hat{SE}(\hat{\beta})$), kestirilen odds oranı (ψ), kestirilen odds oranı için %95 güven sınırları, model için -2log-olabilirlik değeri, eğim katsayısının sıfıra eşit olup olmadığını test eden olabilirlik-oran test istatistiği (G) ve P değerleri verilmiştir. Olabilirlik-oran test istatistiği, modelde yalnız sabit terim bulunurken hesaplanan -2log-olabilirlik değeriyle, modelde test edilmek istenen değişkenin olduğu zaman hesaplanan -2log-olabilirlik değeri arasındaki farka eşittir. Örneğin CINS değişkeni için olabilirlik-oran test istatistiği şöyle hesaplanır;

Tablo 2.1:.Doğum Ağırlığı ile İlgisi Olabileceği Düşünülen Değişkenlerin Tek Değişkenli Lojistik Regresyon Model Sonuçları

Değişkenler	$\hat{\beta}$	$\hat{SE}(\hat{\beta})$	ψ	%95 Güven Sınırları	-2log-Olabilirlik	Wald	G	P
Sabit					100,275			
CINS	1,064	0,544	2,899	0,999-8,414	96,241	3,834	4,034	0,050
YAS	0,106	0,068	1,111	0,973-1,269	97,528	2,434	2,747	0,119
BOY	0,103	0,053	1,109	0,999-1,230	96,316	3,780	3,959	0,052
GHAFTA	2,376	0,574	10,766	3,492-33,190	82,928	17,347	18,259	0,000
GOAG	0,073	0,033	1,076	1,008-1,148	94,502	4,784	5,773	0,029
GAK	0,081	0,054	1,084	0,9761,205	97,755	2,279	2,520	0,131
SDGS	0,219	0,133	1,245	0,960-1,614	96,544	2,724	3,731	0,099
CDUR	-0,565	1,079	0,568	0,069-4,706	99,960	0,275	0,315	0,600
SIG	0,005	0,809	1,005	0,206-4,903	100,275	0,000	0,000	0,995
GR	0,081	0,181	1,084	0,760-1,545	100,066	0,199	0,209	0,656
PR	0,171	0,237	1,186	0,746-1,886	99,706	0,522	0,569	0,470
BES	0,148	0,546	0,172	0,059-0,506	90,272	10,242	10,003	0,001
DSEKLI	0,148	0,544	1,159	0,399-3,365	100,200	0,074	0,0075	0,786

$$G = -\,Ln\left[\frac{(\frac{n_1}{n})^{n_1}(\frac{n_0}{n})^{n_0}}{\prod_{i=1}^{n}\hat{\pi}_i^{y_i}(1-\hat{\pi}_i)^{(1-y_i)}}\right]$$

$$G = 100.275 - 96{,}241 = 4{,}034$$

Olabilirlik oran test sonucunda, Tablo 2.1'in son kolonunda verilen ihtimal düzeyi (P) 0.25'in altında (P<0.25) bulunan değişkenler çok değişkenli model için aday değişkenler olarak belirlenmiştir. Böylece CINS, BOY, GHAFTA, GOAG ve BES değişkenleri bağımlı değişkenle anlamlı derecede ilişkili olduklarından; YAS, GAK ve SDGS değişkenleri de, bağımlı değişkenle en azında orta derecede ilişkili olduklarından dolayı çok değişkenli modele dahil edilmiştirler. Aday değişkenlerle kurulan çok değişkenli modele ilişkin sonuçlar Tablo 2.2'de verilmiştir.

Tablo 2.2:.Tek Değişkenli Modelde Aday Değişken Olarak Alınan Değişkenleri Kapsayan Çok Değişkenli Model Sonuçları

	$\hat{\beta}$	ŞE ($\hat{\beta}$)	Wald	SD	P	ψ Exp(B)	%95 Güven Sınırları	
							Alt	Üst
CINS	1,956	0,851	5,283	1	0,022	7,070	1,334	37,473
YAS	-0,025	0,124	0,041	1	0,839	0,975	0,764	1,244
BOY	0,131	0,078	2,832	1	0,092	1,140	0,979	1,328
GHAFTA	3,473	0,967	12,908	1	0,000	32,247	4,848	214,503
GOAG	0,048	0,051	0,887	1	0,346	1,049	0,949	1,160
GAK	0,145	0,086	2,810	1	0,094	1,156	0,976	1,369
SDGS	0,132	0,209	0,400	1	0,527	1,141	0,758	1,719
BES	-0,977	0,764	1,634	1	0,201	0,377	0,084	1,684
DSEKLI	1,217	0,845	2,074	1	0,150	3,377	0,645	17,689
Sabit	-32,137	13,903	5,343	1	0,021	0,000		

-2 log-olabilirlik = 58,952

Tablo 2.2'de verilen sonuçlara göre YAS, BOY, GOAG, GAK, SDGS ve DSEKLİ değişkenlerinin modelde önemli olmadığı görülmektedir. Öncelikle en yüksek önemsizliğe sahip (P=0,839) YAS değişkenin modelde kalıp kalmayacağı, YAS değişkenini kapsayan ve kapsamayan modeller için olabilirlik-oran test istatistiğinin karşılaştırılması sonucunda karar verilecektir. YAS değişkeni model dışı bırakıldıktan sonra, geriye kalan değişkenlerle kurulan çok değişkenli lojistik regresyon modeline ait -2log-olabilirlik değeri 60,155 olarak bulunmuştur.

YAS değişkenini kapsayan ve kapsamayan modelleri karşılaştıran olabilirlik-oran test istatistiği (G) değeri, bu iki modele ait -2log-olabilirlik değerlerinin farkına eşittir;

$$G = 60,155 - 58,952 = 1,203$$

Bu değer 1 serbestlik derecesi ile ki-kare cetvel değeriyle karşılaştırılır. Hesaplanan (G) istatistiği, 1 serbestlik derecesindeki ki-kare cetvel değerinden küçük olduğu için ($G = 1,203 < \chi^2_{(1)} = 3,841$) YAS değişkeninin modelde kalmasına gerek kalmadığına karar verilir. YAS değişkeninin modele herhangi bir katkısının olmadığı sonucuna varılır.

Benzer işlem diğer önemsiz değişkenler içinde yapılarak en uygun çok değişkenli lojistik regresyon modeli Tablo 2.3'de verilmiştir.

Tablo 2.3:. En Uygun Çok Değişkenli Lojistik Regresyon Modeli

	$\hat{\beta}$	ŞE ($\hat{\beta}$)	Wald	SD	P	ψ Exp(B)	%95 Güven Sınırları	
							Alt	Üst
CINS	1,951	0,758	6,621	1	0,010	7,037	1,592	31,106
BOY	0,137	0,068	4,009	1	0,045	1,147	1,003	1,312
GHAFTA	2,806	0,762	13,578	1	0,000	16,547	3,719	73,611
BES	-1,547	0,687	5,072	1	0,024	0,213	0,055	0,818
Sabit	-25,480	11,472	4,933	1	0,026			

-2 log-olabilirlik = 65,566

Etkileşim terimlerine karar vermeden önce, hem sürekli değişkenlerin logitle doğrusal bir ilişki içinde olup olmadıklarını, hem de bu değişkenlerin modele doğru ölçekle girip girmediklerini kontrol etmek gerekmektedir. Modelimizde bulunan 4 değişkenden birisi sürekli değişkendir. BOY sürekli değişkenlerin logitle doğrusallığı test edilecektir. BOY değişkeni dört eşit gruba bölündükten sonra (kuartiller kullanılarak) en küçük grup referans grup olarak belirlenip, üç dizayn değişkeni ile (sürekli BOY değişkeni yerine) modele dahil edilecektir. Yeni değişkenimiz BOY1 olarak isimlendirilecektir (Tablo 2.4.).

Tablo 2.4: BOY Değişkeni İçin Kuartil Analiz Sonuçları

	$\hat{\beta}$	ŞE ($\hat{\beta}$)	Wald	SD	P	ψ Exp(B)	%95 Güven Sınırları	
							Alt	Üst
CINS	1,809	0,737	6,021	1	0,014	6,106	1,439	25,906
GHAFTA	2,626	0,728	13,012	1	0,000	13,825	3,318	57,600
BES	-1,474	0,673	4,792	1	0,029	0,229	0,061	0,857
BOY1			4,901	3	0,179			
BOY1(1)	-,208	1,869	0,012	1	0,911	0,812	0,021	31,643
BOY1(2)	-1,634	1,233	1,757	1	0,185	0,195	0,017	2,186
BOY1(3)	,070	1,419	0,002	1	0,960	1,073	0,066	17,314
Constant	-2,193	2,193	0,999	1	0,318	0,112		

Eğer logit BOY değişkeniyle doğrusal ise, o zaman kestirilen katsayılarda doğrusal artan ya da azalan bir eğilim olması beklenir. BOY1 değişkeni için Tablo 2.4'de verilen kuartil analizi sonucunda, 4. grubun kestirilen odds oranı, 1. grubun kestiri-

len odds oranına yakın olduğu görülmektedir. 2. grubun kestirilen odds oranının ise, diğer iki gruptan daha küçük olduğu görülmektedir. Bu durum BOY değişkeninin logitle doğrusal bir ilişki içerisinde olmadığının göstergesidir. Box-Tidwell yaklaşımı kullanılarak, sürekli dağılıma sahip BOY değişkenini kapsayan modele BOY[ln(BOY)] değişkeninin eklenmesi sonucunda, BOY[ln(BOY)] değişkeninin istatistiksel olarak önemli bir eğime sahip olması BOY değişkeninin logitle doğrusal olmadığı varsayımını desteklemektedir. Bu nedenden dolayıdır ki, BOY değişkeni modele dizayn değişkeni kullanarak, kesikli bir değişken olarak girecektir. Üç grubun kestirilen odds oranları birbirine yakın olduğundan (0,812-0,195-1,073) her üç grup birleştirilecek (Hosmer ve Lemeshow, 1989) ve kuartil analizinde kullanılan kesim noktaları yerine Tablo 2.5'de verilen kesim noktaları kullanılacaktır.

Tablo 2.5: BOY Değişkeninin Kesikli Değişken Olarak Kodlanması

Gruplar	Kod	YBOY
BOY≤160	1	1
BOY>160	0	0

Yeni oluşturulan desen değişkeni (YBOY), BOY değişkeni yerine modele dahil edilir. Oluşturulan yeni çoklu lojistik regresyon modelinin analiz sonuçları Tablo 2.6'da verilmiştir:

Tablo 2.6: YBOY Değişkenin Modele Dahil Edildiği Çok Değişkenli Lojistik Regresyon Modeli

	$\hat{\beta}$	ŞE ($\hat{\beta}$)	Wald	SD	P	ψ Exp(B)	%95 Güven Sınırları	
							Alt	Üst
CINS	1,738	0,735	5,595	1	,018	5,684	1,347	23,985
GHAFTA	2,657	0,727	13,368	1	,000	14,249	3,430	59,194
BES	-1,501	0,677	4,925	1	,026	0,223	0,059	0,839
YBOY	1,578	0,799	3,899	1	,048	4,844	1,012	23,189
Constant	-5,190	2,273	5,213	1	,022			

-2log-olabilirlik: 65,268

Son aşamada kurulan model için etkileşim terimlerinin incelenmesi gerekir. Kurulan modele her bir etkileşim terimi ilave edilerek meydana gelen yeni modelin -

2log-olabilirlik değerleri, olabilirlik oran test istatistiği (G), serbestlik derecesi (SD) ve P değerleri Tablo 2.7'de verilmiştir.

Tablo 2.7: Herbir Etkileşim Teriminin Ana Etkiler Modeline Eklenmesiyle Elde Edilen Yeni Model İçin -2log-olabilirlik Değeri, G Değeri, Serbestlik Derecesi ve P Değeri

Etkileşim Terimleri	-2log-olabilirlik	G	SD	P değeri
Ana Etkiler Modeli	65,268			
CINS x BES	65,204	0,064	1	P>0,05
CINS x YBOY	64,653	0,615	1	P>0,05
CINS x YGHAFTA	64,842	0,786	1	P>0,05
BES x YBOY	65,111	0,157	1	P>0,05
BES x YGHAFTA	62,981	2,287	1	P>0,05
YBOY x YGHAFTA	65,121	0,147	1	P>0,05

Tablo 2.7'de de görüldüğü gibi tüm etkileşim terimleri istatistiksel olarak anlamlı bulunmamıştırlar. Modele herhangi bir etkileşim terimi dahil edilmeyecektir. Tablo 2.6'da verilen modelimiz en uygun modeldir.

2.1. Modelin Uyum İyiliği

Kurulan modelin uyum iyiliği testi Hosmer-Lemeshow'un hem onlu risk grupları hem de sabit kesim noktası yöntemine göre hesaplanacaktır. Tablo 2.8'de onlu risk grupları yöntemiyle hesaplanan beklenen ve gözlenen değerler verilmiştir.

Tablo 2.8: Sabit Denek Sayılı Onlu Risk Grupları İçin Gözlenen ve Beklenen Frekanslar

İkili Bağımlı Y Değişkenin Değeri		Risk Grupları										Toplam
		1	2	3	4	5	6	7	8	9	10	
Y = 1	Gözlenen	0	0	0	1	1	0	5	8	2	0	17
	Beklenen	0	0,125	0,092	1,618	0,827	0,613	4,743	8,178	1,804	0	
Y = 0	Gözlenen	0	24	4	24	22	6	24	7	0	0	111
	Beklenen	0	23,875	3,908	24,382	22,173	5,387	24,257	6,822	0,196	0	
Toplam		0	24	4	25	23	6	29	15	2	0	128

Uyum iyiliğine karar vermek için, Denklem 31'de verilen formül yardımıyla Tablo 2.8'deki veriler kullanılak, Hosmer-Lemeshow $\hat{C}_g^*$ istatistiği hesaplanır.

$$\hat{C}_g^* = \sum_{k=0}^{1}\sum_{l=1}^{10}\frac{(o_{kl}-e_{kl})^2}{e_{kl}} = \frac{(0-0{,}125)^2}{0{,}125} + \frac{(0-0{,}092)^2}{0{,}092} + \ldots + \frac{(0-0{,}196)^2}{0{,}196} = 1{,}424$$

Hosmer-Lemeshow $\hat{C}_g^*$ istatistiği, $\alpha = 0.05$ yanılma düzeyi ve 6 serbestlik dereceli ki-kare dağılımıyla karşılaştırılır. $\hat{C}_g^* < \chi^2_{0.05,6} = 12{,}592$ olduğundan model uyumunun oldukça iyi olduğu sonuçuna varılır.

Kestirilen modelin uyum iyiliği testi sabit kesim noktası yöntemiyle hesaplanmak istendiğinde (Tablo 2.9) Denklem 32'de verilen Hosmer-Lemeshow $\hat{H}_g^*$ istatistiği kullanılmaktadır.

Tablo 2.9: Sabit $P(x_i)$ Aralıklı Onlu Risk Grupları İçin Gözlenen ve Beklenen Frekanslar

İkili Bağımlı Y Değişkenin Değeri		Risk Grupları										Toplam
		0-0.1	.11-.20	.21-30	.31-.40	.41-50	.51-.60	.61-.70	.71-.8	.81-.90	.91-1.0	
Y = 1	Gözlenen	2	3	2	2	0	0	6	0	0	2	17
	Beklenen	1,662	3,259	2,097	2,357	0,000	0,000	5,821	0,000	0,000	1,804	
Y = 0	Gözlenen	74	24	6	4	0	0	3	0	0	0	111
	Beklenen	74,338	23,741	5,903	3,643	0,000	0,000	3,000	0,000	0,000	0,000	
Toplam		76	27	8	6	0	0	9	0	0	2	128

$$\hat{H}_g^* = \sum_{k=0}^{1}\sum_{l=1}^{10}\frac{(o'_{kl}-e'_{kl})^2}{e``_{kl}} = \frac{(2-1{,}662)^2}{1{,}662} + \frac{(3-3{,}259)^2}{3{,}259} + \ldots\ldots\ldots + \frac{(3-3)^2}{3} = 0{,}442$$

Hosmer-Lemeshow $\hat{H}_g^*$ istatistiği, $\alpha = 0.05$ yanılma düzeyi ve 4 serbestlik dereceli ki-kare dağılımıyla karşılaştırılır. $\hat{H}_g^* < \chi^2_{0.05,4} = 9{,}488$ olduğundan modelin uyumunun oldukça iyi olduğu sonuçuna varılır.

Tablo 2.6'da verilen en uygun modele ait sınıflandırma tablosu, Tablo 2.10'da verilmiştir. Tablodan da görüldüğü gibi doğru sınıflandırma oranı %90,6 şeklinde bulunmuştur. Sınıflama tablosunu kullanmak analizin amacı sınıflama olduğu zaman doğrudur. Yoksa uyuma karar vermek için yeterli değildir.

Tablo 2.10: Uydurulan Model için Sınıflandırma Tablosu

		Kestirilen Doğum Ağırlığı		Doğrulama Oranı
		≤2500	>2500	
Gözlenen	≤2500	8	9	47,1
Doğum Ağırlığı	>2500	3	108	93,3
Ortalama %				90,6

3. Sonuçlar

Anket türü verilerde kesikli yanıt olması ve açıklayıcı değişkenlere sıkça rastlanması, amaç regresyon analizi olduğunda regresyon modelini kurmayı, ayrımsama olduğunda da diskriminant fonksiyonunu oluşturmayı yetersiz kılmaktadır. Böyle durumlarda alternatif olarak Lojistik Regresyon Analizi kullanılmaktadır. Bu çalışmada lojistik regresyon konusu literatürden genel hatları ile araştırılmış, iki sonuçlu bağımlı değişken, kalitatif ve kantitatif bağımsız değişkenler irdelenmiştir. Bağımsız değişkenlerin önemli olup olmadıklarını anlamak amacıyla lojistik regresyon analizi teknikleri kullanılmıştır.

Düşük doğum ağırlığında risk faktörü olan 13 değişken (CINS, YAS, BOY, GHAFTA, GOAG, GAK, SDGS, CDUR, SIG, GR, PR, BES ve DSEKLI) incelenmiştir. Tablo 2.6'da görüldüğü gibi, bu değişkenlerden sadece CINS, GHAFTA, BES ve YBOY değişkenleri ile en uygun model oluşturulmuştur. Düşük doğum ağırlığında cinsiyet, gebelik haftası, beslenme durumu ve anne boyu değişkenlerinin anlamlı olduğu görülmüştür.

Tahmin edilen lojistik regresyon modelinden çıkartılacak yorumlar, modeldeki risk faktörleri için tahmin edilen odds oranları ile yapılır. Final modelimiz hiçbir interaksiyon terimi içermediğinden CINS, GHAFTA, BES ve YBOY risk faktörleri için odds oranları, tahmin edilen katsayıların üsteli alınarak elde edilir.

Tablo 2.6'da Exp(B) sütununda odds oranları(ψ) verilmiştir. CINS değişkenine ait odds oranı 5,684 olarak bulunmuştur. Bu katsayısı, doğacak bebeğin kız olması durumunda düşük doğum ağırlığında olması olasılığının, erkek olması durumuna göre 5,684 kat daha yüksek olduğunu gösterir. GHAFTA değişkenine ait odds oranı ise 14,249 olarak bulunmuştur. Bu katsayısı doğacak bebeğin ≤38 haftalıktan az olması durumunda düşük doğum ağırlığında olması olasılığının, 38 haftalıktan fazla olması durumuna göre 14,249 kat daha yüksek olduğunu gösterir. YBOY değişkenine ait

odds oranı ise 4,844 olarak bulunmuştur. Anne boy uzunluğu ≤160 cm'den küçük olması durumunda düşük doğum ağırlığında olması olasılığının, 160 cm'den büyük olması durumuna göre 4,844 kat daha yüksek olduğunu gösterir.

Tablo 2.10'da verildiği gibi CINS, GHAFTA, BES ve YBOY değişkenleri ile kurulan modelin doğru sınıflandırma oranı %90,6 olarak bulunmuştur.

Final modelin uyum iyiliği, Hosmer ve Lemeshow'un hem onlu risk grupları hem de sabit kesim noktası yöntemine ile test edilerek oldukça iyi bir uyum ortaya koyduğu görülmüştür.

Studies of the Logistic Regression Analysis and Its Application on the Medical Data

Abstract: The purpose of this study is to evaluate logistic regression model which is able to define the relation between dichotomous outcome variable and the set of independent variables that contains both continuous and discrete variables. To describe the application of logistic regression analysis, it was studied on medical data to determine important risk factors effecting the real birth weight of children. After variables which will be included in logistic model were described by univariate logistic regression, the importance of each variable included in the multivariate model should be verified. The variables which are not statistically significant in multivariate model were removed from the model despite it's significant in univariate model. Since the final model is both biologically acceptable and its rate of correct classification is good enough, final model can be used for both determining risk factors.

Key Words: Logistic Regression analysis, Odds ratio

Kaynakça

Akaya, Ş., M.V. Pazarlıoğlu (1998), *Ekonometri*. Erkan Matbaacılık, İzmir,

Anderson, J. A. (1979), *Multivariate Logistic Compounds*, Biometrika, 66: 17–191 .

Anderson, J. A. (1983), *Robust İnference Using Logistic Models*, Bulletion of ınternational Statistical Institute, 48: 35-53.

Aranda – Ordaz, F. J. (1981), *On two Families of Transformations to Additivity for Binary Responce Data*, Biometrika, 68: 357–363.

Başarır, G. (1990), *Çok Değişkenli Verilerde Ayrımsama Sorunu ve Lojistik Regresyon Analizi*. (Uygulamalı istatistik doktora tezi.) H.Ü., 1-36, Ankara.

Bonney, G.E. (1987), *Logistic Regression for Dependent Binary Observations*. Biometrics, 43: 951-973.

Buescher, P.A., Larson, L.C., Nelson, M.D., Lenihan, A.J. (1993). *Prenatal WIC Participation Can Reduce Low Birth Weight and Newborn Medical Costs: A Cost Benefit Analysis of Wic Participation in Nort Carolina*, Journal of the American Dietetic Association, 93:163-166.

Buse, A. (1982), *The Likelihood Ratio, Wald and Lagrange Multiplier Tests: An Expository Note*, The American Statistician,Vol. 36, No. 3, Part 1.

Cornfield, J. (1962), *Joint Dependence of the Risk of Coronary Heart Disease on serum Cholesterol and Sistolic Blood Pressure: A Diskrimant Function Analysis*, Federation Proceedings, 21: 58-61.

Cox, B.D., Whichelow, H.J., Prevost, A.T. (1998). *The Development of Cardiovascular disease in relation to anthropometric indices and hypertension in British adults*. International Journal of Obesity, 22: 97330-10966.

Cox, D. R. – Snell, E. S., 1989, *Analysis of Binary Data*, London.

Duffy, D.E. (1990). *On Continuity-corrected Residuals in Logistic Regression*, Biometrika, 77: 287-293.

Elhan, A.H. (1997), *Lojistik Regresyon Analizinin İncelenmesi ve Tıpta Bir Uygulaması*. (Biyoistatistik Yüksek Lisans Tezi) A.Ü.,4-29, ANKARA

Gardside, P.S., Glueck, C.J. (1995). *The Important Role of Modifiable Dietary and Behaviour Characteristic in the Causation and Prevention of Coronary Heart Disease Hospitalization and Mortality*. Journal of American College of Nutrition, 14: 71-79.

Hosmer, D. W. – Lemeshow, S. (1980), *Goodness of Fit Tests for the Multiple Logistic Regression Model*, Communications in Statistics, Seri A 9. 1043-1069

Hosmer, D. W. – Lemeshow, S., (1989), *Applied Logistic Regression*, John Willey & Sons.,

Hsu, J.S., Leonard, T. (1995), *Hierarchical Bayesian Semiparametric Procedures for Logistic Regression*. Biometrika, 84: 85-93.

Johnson, W. (1985), *Influence Measures for Logistic Regression, Another Point of View*, Biometrika, 72, 1, 59–65 .

Kloiber, L.L., Winn, N.J., Shaffer, S.G., Hassanein, R.S. (1996), *Late Hyponatremia in very Low Birth Weight Infants: Incidence and Associated Risk Factors*. Journal of the American Dietetic Association, 96: 880-884..

Lee, C.T. (1984), *Logistic Models for Cross-over Designs*. Biometrika, 71: 216-217.

Lesaffre, E. (1986), *Logistic Discriminant Analysis With Applications İn Electrocardiography*, PhD thesis, Katholieke Universiteit Leuven, Belgium, 354 p. (unpublished) .

Lesaffre, E., A. Albert (1989), *A Multiple Group Logistic Regression Diagnostics*, Applied Statistics, 38, 3, 425–440.

Peoples, M.D., Siegel, E., Suchindran, C.M., Origasa, H., Ware, A., Barakat, A. (1991), *Characteristics of Maternal Employment During Pregnancy: Effects on Low Birtweight*. American Journal of Public Health. 81: 1007-1012.

Pindyck, R. – Rubinfeld, D. (1985), *Econometric Models*, Mc Graw Hill, Singapore.

Pregibon, D. (1981), *Logistic Regression Diagnostics*, The Annals of Statistics, 9, 705–724.

Robert, G., Rao, N.K., Kumar, S. (1987), *Logistic Regression Analysis of Sample Data*, Biometrika, 35: 58:79.Survey

Sable, M.R., Herman, A.A. (1997), *The Relationship Between Prenatal Health Behaviour Advice and Low Birthweight.* Public Health Reports. 112: 332-339.

Santos, I.S., Victoria, C.G., Huttly, S., Carvalhal, J.B. (1998), *Caffeine Intake and Low Birth Weight: A Population Based Case Control Study*. American Journal of.M. (1988), *The Retreat From Class: A New True Socialism*, London: Verso.