T.C.
MUĞLA ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

İSTATİSTİK ANABİLİM DALI

BELİRLİ ÖLÇÜTLERE GÖRE İKİ DÜZEYLİ LOJİT VE PROBİT MODELLERİN KARŞILAŞTIRILMASI

YÜKSEK LİSANS TEZİ

SELEN ÇAKMAKYAPAN

EKİM 2011
MUĞLA

T.C.
MUĞLA ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

İSTATİSTİK ANABİLİM DALI

BELİRLİ ÖLÇÜTLERE GÖRE İKİ DÜZEYLİ LOJİT VE PROBİT MODELLERİN KARŞILAŞTIRILMASI

YÜKSEK LİSANS TEZİ

Selen ÇAKMAKYAPAN

MUĞLA 2011

T.C.
MUGLA ÜNİVERSİTESİ
Fen Bilimleri Enstitüsü

Yrd. Doç. Dr. Atilla GÖKTAŞ .danışmanlığında Selen ÇAKMAKYAPAN tarafından hazırlanan "Belirli Ölçütlere Göre İki Düzeyli Lojit ve Probit Modellerin Karşılaştırılması" başlıklı tez, ..25../..10../..2011.. tarihinde aşağıdaki jüri tarafından İstatistik Anabilim Dalı'nda yüksek lisans tezi olarak oybirliği/oyçokluğu ile kabul edilmiştir.

Başkan : Yrd. Doç. Dr. Andun Oben BALCE İmza :

Üye : Yrd. Doç. Dr. Atilla GÖKTAŞ İmza :

Üye : Yrd. Doç. Dr. Özge AKKUŞ İmza :

ÖNSÖZ

Gerek tez çalışmam, gerek lisans ve yüksek lisans eğitimim süresince bana destek ve yardımlarını esirgemeyen, değerli danışmanım Sayın Yrd. Doç. Atilla GÖKTAŞ'a,

Katıldığım dersi sayesinde tez konuma ilgi duymamda ve tez aşamasında emeği olan hocam Sayın Yrd. Doç. Dr. Özge AKKUŞ'a,

Değerli görüş ve önerileri için Sayın Yrd. Doç. Dr. Andım Oben BALCE'ye,

Üniversite eğitimim boyunca bana verdikleri özgüven ve destekleri için başta bölüm başkanımız Sayın Prof. Dr. Mustafa DİLEK olmak üzere tüm bölüm hocalarıma,

Benim her zaman yanımda olan ve her türlü manevi desteklerini esirgemeyen çok sevdiğim araştırma görevlisi çalışma arkadaşlarıma,

Emeklerini asla ödeyemeyeceğim her zaman her koşulda yanımda olan annem Sümran ÇAKMAKYAPAN ve abim Serkan ÇAKMAKYAPAN'a içtenlikle teşekürlerimi sunarım.

Selen ÇAKMAKYAPAN
Muğla, 2011

İÇİNDEKİLER

BELİRLİ ÖLÇÜTLERE GÖRE İKİ DÜZEYLİ LOJİT VE PROBİT MODELLERİN KARŞILAŞTIRILMASI

(Yüksek Lisans Tezi)

Selen ÇAKMAKYAPAN

MUĞLA ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

2011

ÖZET

Bağımlı değişkenin iki düzeyli olması durumunda Genelleştirilmiş Doğrusal Model ailesi üyelerinden Lojit ve Probit Modeller istatistiksel modellemede yaygın olarak kullanılmaktadır. Aynı veri seti için kullanılabilen iki modeli birbirinden ayıran özellik; kullandıkları varsayımlar ve buna bağlı olarak matematiksel fonksiyonlarıdır. Bu iki model arasında hangi koşullarda hangi modelin tercih edilmesi gerektiğine dair kesin bir yargı bulunmamaktadır. Bu çalışmada, iki düzeyli Lojit ve Probit Modellerin farklı koşullar altında veriye olan uyumları karşılaştırılarak, hangi koşulda hangisinin kullanılacağına dair bir sonuca ulaşmak amacıyla bir simülasyon çalışması gerçekleştirilmiştir. Simülasyon aşamasında çok değişkenli normal dağılıma sahip bağımlı ve açıklayıcı değişkenler üretilmiştir. Üretilen veri setinde bağımlı değişken sürekli bir değişken olduğundan Lojit ve Probit Modellere uygulanabilmesi için iki düzeyli olarak sınıflandırılmıştır. Elde edilen iki düzeyli bağımlı değişken ve açıklayıcı değişkenlere ait veriler Lojit ve Probit Modeller kullanılarak analiz edilmiştir. İki modele ilişkin uyum iyiliği sonuçları: artıklar, sapmalar ve iki düzeyli bağımlı değişken modellerinde kullanılan bazı yapay R^2 değerleri dikkate alınarak karşılaştırılmıştır. Bu işlemler bağımlı değişkeni sınıflandırmasında kullanılan iki ayrı kesim noktası, değişkenler arasında üç farklı ilişki düzeyi (çok, orta, yok) ve beş farklı örneklem büyüklüğü için gerçekleştirilmiştir. Yapılan analizler sonucunda her iki modelden elde edilen tahmini değerlerin birbirlerine oldukça yakın oldukları ve yapay R^2'lerin çoğuna

göre aralarında anlamlı bir farklılığın olmadığı saptanmıştır. Ancak artıklar incelendiğinde, 200 ve daha küçük örneklem büyüklüklerinde Probit Modelin, 200'den büyük örneklemlerde de Lojit Modelin tercih edilmesinin daha iyi olacağı sonucuna varılmıştır. Bunun yanı sıra, farklı kesim noktalarının ve ilişki düzeylerinin model seçiminde etkili olmadığı görülmüştür.

Anahtar Kelimeler : Doğrusal Olasılık Modeli, İki Düzeyli Lojit Model, İki Düzeyli Probit Model

Sayfa Adedi : 108

Danışman : Yrd. Doç. Dr. Atilla GÖKTAŞ

COMPARISON OF BINARY LOGIT AND PROBİT MODELS BASED ON SPECIFIC CRITERIA

(M. Sc. Thesis)

Selen ÇAKMAKYAPAN

MUĞLA UNIVERSITY
INSTITUTE OF SCIENCE

2011

ABSTRACT

Logit and Probit Models among Generalized Linear Models are widely used, when the dependent variable is observed to be binary. The differences between these two models that can use the same data set are due to the assumptions they use and hands their mathematical functions. There is no study specifying a certain judgement on the preference of these models to make a decision which model is better in what condition. A simulation study is conducted to make a comparison of the model fits to the generated data set under certain conditions to reach an end to which condition is better. In the process of simulation study, a dependent and explanotory variables are generated from multivariate normal distribution. Since the generated dependent variable is continuous, the variable are classified as binary to make dataset usable for Logit and Probit Models. Analysis has been made to the Logit and Probit Models for the generated dataset for both the binary dependent and the explanatory variables. Comparison has been made according to the goodness of fit test results related to both models: residuals, deviances and some pseudo R^2 used for binary dependent variables. These procedures have been performed for two different cut points used to classify response variables, three different relationship levels among variables (high, medium, none) and five different sample sizes. It has been found that there are no statisticaly significant differences among most pseudo R^2 and the estimated values from both models are considerable close according to the analysis results. However, residues examined, small sample sizes of 200 and more Probit Model, the Logit model for the choice of more than 200 large samples is concluded to be better. In

addition, the different cut-off levels and the relationship did not influence the choice of the model.

Key Words **:** Linear Probabilty Model, Binary Logit Model, Binary Probit Model

Page Number **:** 108

Adviser **:** Asst. Prof. Dr. Atilla GÖKTAŞ

ŞEKİLLER DİZİNİ

TABLOLAR DİZİNİ

KISALTMALAR DİZİNİ

EKK	En Küçük Kareler
DİSK	Diskriminant Analizi
PÇR	Poisson Çok Değişkenli Regresyonu
GDM	Genelleştirilmiş Doğrusal Modeller
DOM	Doğrusal Olasılık Modeli
EÇO	En Çok Olabilirlik
AEKK	Ağırlıklı En Küçük Kareler
kdf	Kümülatif dağılım fonksiyonu

1. GİRİŞ

Sosyal bilimlerde yoğunluklu olmak üzere, birçok alanda kullanılan istatistik araçlarından biri olan regresyon analizi, bağımlı değişken ile bağımsız değişkenler arasındaki istatistiksel ilişkiyi matematiksel bir fonksiyon elde edilerek açıklamaya ve değerlendirmeye çalışan bir analiz yöntemidir.

Bu analizin oldukça yaygın olarak kullanılmasının başlıca sebepleri vardır. Bunlardan biri, çok değişkenli yapısı sayesinde birden fazla bağımsız (açıklayıcı) değişkeni kullanabilme imkanı sağlamasıdır. Bir diğeri ise, kolay yorumlanabilir olması ve bazı varsayım ihlallerinde bile anlamlı sonuçlar verebilecek olmasıdır. Ancak yaygın olmasının yanı sıra en çok yanlış olarak kullanılan istatistiksel yöntemlerden biridir. Bazı varsayımlar altında regresyon analizinden elde edilen tahminler, hataya karşı güçlü olabilmelerine rağmen önemli varsayımların gerçekleşmemesi durumunda, yetersiz olabilmekte ve oldukça anlamsız tahminlerle karşılaşılabilmektedir. Bağımlı değişkenin sürekli ya da aralık düzeyinde ölçülmüş olması yerine nitel ölçümlü olması durumunda bu sorunla karşılaşılır. Bağımlı değişkenin kategorik olduğu regresyon tahminlerinde, bağımsız değişkenlerinin etkilerinin büyüklükleri tahmin edilemez. Hipotez testleri, güven aralıkları gibi istatistiksel sonuçların hepsi anlamsızlaşır ve regresyon tahminleri bağımsız değişkenlerin gözlemlenen belirli değerlerinde yüksek hassasiyete sahip olur (Aldrich ve Nelson, 1984).

Regresyon analizlerinde ve diğer çalışmalarda kategorik değişkenlerle sıkça karşılaşılmaktadır. Bu değişkenler nicel olabildikleri gibi nitel de olabilirler. Nitel değişkenler birçok araştırma için bağımlı ve bağımsız değişken olarak regresyon analizinde sıklıkla kullanılmaktadır. Ancak, bağımlı değişkenin iki ya da daha fazla düzeye sahip nitel bir değişken olması ya da en azından öyle gözlenmesi durumunda, etkin parametre tahminlerine ulaşmak için sıradan regresyon analizi yetersiz kalır ve regresyon analizinde kullanılan en küçük kareler yönteminin kullanılması birçok varsayım ihlalinden dolayı mümkün olmaz. Örneğin, ekonometride satın almak-almamak, siyasal bilimlerde oy verme-vermeme, sosyolojide medeni durum, mezuniyet derecesi, mutluluk düzeyi gibi birçok niteliksel seçimlerle ilgilenilir. Bu davranışların hepsi nitelikseldir ve çoğu iki düzeyli olarak temsil edilirler. Gerçekte

nicel olan bazı durumlar, tutum ve davranışlar da pratikte nitel olarak ölçülebilirler. Böyle durumlarda bilinen sıradan regresyon analizini kullanmak doğru ve yeterli sonuçlar vermeyebilir. Bu sakıncaları ortadan kaldırmak için farklı süreçler ve analizler kullanılmaktadır. Bunlardan en yaygın olarak kullanılanları; Doğrusal Olasılık Modeli , Lojit Model, Probit Model ve Tobit Modeldir.

Yapılan bu çalışmada, bağımlı değişkenin kategorik bir değişken olması ve daha özel olarak iki düzeyli ölçülmesi durumunda kullanılan iki düzeyli Lojit ve Probit Modeler üzerinde durulmuştur. Çalışanın temel amacı, bu iki modelin oldukça benzer tahmin değerleri elde etmelerine rağmen, hangi koşullarda hangisinin tercih edilmesinin daha doğru olacağı konusunda bir yargıya varmaktır.

Çalışmanın ilk bölümünde iki düzeyi Lojit ve Probit Modelin kullanım alanları ve modellerin karşılaştırılması ile ilgili yapılan önceki çalışmalara yer verilmiştir. İkinci bölümünde, kategorik değişkenler, iki düzeyli Lojit ve Probit Modelde kullanılan gizli değişken yaklaşımı, regresyon ve olasılık modelleri hakkında genel bilgilere değinilmiştir. Sonrasında Lojit ve Probit Modelde tahmin aşamasında kullanılan En Çok Olabilirlik yöntemi üzerinde durulmuştur. Bir sonraki bölümde iki düzeyli bağımlı değişken durumunda kullanılan Doğrusal Olasılık Modeli ve kullanılması durumunda karşılaşılan sorunlara yer verilmiştir. Ardından bu sorunlara çözüm olarak geliştirilen iki düzeyli Lojit ve Probit Model anlatılmış ve bu modellere ilişkin kullanılan uyum iyiliği ölçütleri üzerinde durulmuştur. Son bölüm olan uygulamada, yapılan simülasyon çalışmasıyla Lojit ve Probit Modellere ilişkin uyum iyiliği ölçüt değerleri tablolaştırılarak, gerekli sonuç ve yorumlara yer verilmiştir.

2. ÖNCEKİ ÇALIŞMALAR

İki düzeyli bağımlı değişkenlerle sosyal, sağlık, ekonomi ve daha birçok alanda karşılaşılmaktadır. Dolayısıyla, bu alanlarda yapılan araştırmalarda iki düzeyli cevap değişkenli modeller yaygın olarak kullanılmaktadır. Bu modeller arasında, en sık kullanılanlar ise Lojit ve Probit Modellerdir. Bu modellerin kullanıldığı ve incelendiği bazı çalışmalara aşağıda yer verilmiştir.

Ragsdale tarafından 1984 yılında, ulusal radyo ve televizyonda yayınlanan başkanlık konuşmalarının etkenleri ve sonuçları üzerine deneysel bir çalışma yapılmıştır. Bu çalışmada, başkan Truman'dan Carter'a kadar 1949-1980 yılları arasında ABD'de başkanlık yapan sekiz başkana ait ulusa sesleniş konuşmaları dikkate alınmıştır. Toplum davranışının, ulusal şartların ve olayların etkilerini temel alan iki düzeyli Probit analiz uygulanarak, bir ay olarak belirlenen zaman diliminde, başkan tarafından ulusa sesleniş konuşması yapılması olasılığı tahmin edilmiştir. Sonuçlar toplum onay oranlarındaki değişimin ve ulusal olayların varlığının bir başkanın konuşma yapması olasılığını arttırdığını göstermiştir. Diğer taraftan, kötü ekonomik durumlar (enflasyon ve işsizlik), askeri durumlar konuşma yapılması olasılığını azaltmıştır. Ayrıca, bir başkanın popularitesinin ulusa sesleniş konuşmasıyla anlamlı derecede yükseldiği görülmüştür.

Chin (2002) uçuş programları sıklıklarının hava seyahati taleplerinin üzerine etkisi ile ilgili bir çalışma gerçekleştirmiştir. Yapılan çalışmanın amacı; havayolları taleplerinde etkili olan faktörleri belirlemektir. Singapur Havayolları hizmetlerinin tercih edilme sebeplerinin açıklanmasında iki düzeyli seçim modelleri kullanılmıştır. Singapur havayollarının bireyler tarafından seçilmesini etkileyen en önemli faktör, bu hava yollarının diğer hava yollarına nazaran uygun bir uçuş programına sahip olması olarak bulunmuştur. Diğer anlamlı bir değişken ise, az ama pozitif etkili olan Krisflyer FFP (Frequent Flier Program) üyesi olmaktır. Çalışmada kullanılan örnek, farklı pazarlama bölümlerine ayrılarak sınıflandırılmıştır. Bu bölümler; iş seyahatine karşı serbest seyahat, uzak mesafe yolcusuna karşı kısa mesafe yolcusu, Krisflyer FFP üyesi olunmasına karşı Krisflyer FFP üyesi olunmaması ve FFP üyesi olunmasına karşı FFP üyesi olunmaması şeklindedir. Yapılan çalışma sonunda, her sınıflandırma bölümü arasında genel bir değişimin olduğu ifade edilmiştir.

Temple (2004) çalışmasında ileri yaştaki Avustralyalıların sağlık sigortası satın alma kararları üzerinde araştırma yapmıştır. Ekonomik, demografik ve sağlık faktörlerinin ileri yaştaki bireyin sağlık sigortası satın alma kararı ile ilgili olduğu bulunmuştur. Özellikle, düşük gelir ve eğitim düzeyine sahip bireylerden yalnız yaşayan ya da yurtdışında doğanların ileri yaşlarda sağlık sigortası alma ihtimalinin yüksek olduğu belirtilmiştir. Bu çalışmanın analiz aşamasında iki düzeyli Lojit ve Probit Modellerden faydalanılmıştır.

2006 yılında Gan, Clemes, Limsombunchai ve Weng tarafından yapılan çalışmada Yeni Zelanda'da tüketicilerin elektronik bankacılığını tercih etme durumları incelenmiştir. Bu çalışmada 1960 kişiden elde edilen anket sonuçları dikkate alınmıştır. Elektronik bankacılığı kullanma kararını etkileyen faktörler olarak; servis kalitesi, algılanan risk faktörleri, kullanıcı giriş faktörleri, fiyat faktörü, servis üretim karakteristikleri, bireysel faktörler ve yaş, cinsiyet, medeni durum, gelir gibi demografik faktörler dikkate alınmıştır. Bu çalışmada bağımlı değişken, bir bireyin elektronik banka kullanıcısı olup olmadığının bir ölçüsü olmuştur ve elde edilen verilerin analizinde Lojit Model kullanılmıştır. Analizler sonucunda, servis kalitesi, algılanan risk faktörleri, kullanıcı giriş faktörleri, iş ve eğitim elektronik bankacılığını kullanma seçimi üzerinde en baskın değişkenler olarak bulunmuştur.

Wiersema ve Bowen (2009) strateji araştırmacılarının dikkatlerinin, sabit performans üzerindeki strateji seçimlerinin etkisini açıklamaktan, sabit seviyedeki stratejik seçimleri belirleyen faktörleri açıklamaya döndüğünü ve bu durumda araştırmalarda sınırlandırılmış bağımlı değişkenlerle daha fazla karşılaşılacağını belirtmişlerdir. Çalışmada, sınırlandırılmış bağımlı değişkenler için Lojit ya da Probit gibi modellerin kullanılacağı belirtilerek, yaptıkları çalışmada, sınırlı bağımlı değişken modellerinden elde edilen sonuçların yorumu ve analizi için en genel yöntemler örneklenmiş ve ele alınmıştır.

İki düzeyli verilerin analizinde, lojistik ve Probit regresyonlar başta olmak üzere birçok model kullanılabilmektedir. Bu nedenle, aynı veri seti için kullanılabilen bu farklı yaklaşımların karşılaştırılması söz konusu olmuştur. Araştırmacı hangi sürecin veriye daha uygun olduğunu ve hangi modelle daha iyi sonuçlar elde edeceğini bilmek ister. Bu amaçla literatürde, iki düzeyli veri analizinde kullanılan birçok yöntemin karşılaştırıldığı çalışmalar yer almaktadır. Bu

çalışmalardan biri, Malthora (1983) tarafından gerçekleştirilmiştir. Öncesinde yapılan benzer çalışmalar ise, Malthora (1983) tarafından aşağıdaki tablo ile özetlenmiştir. Aşağıdaki tabloda, "EKK", en küçük kareler regresyonunu; "DİSK", diskriminant analizini; "PÇR", poisson çok değişkenli regresyonu ifade etmektedir. Malthora (1983) çalışmasında Tablo 2.1'de yer alan çalışmalar hakkında aşağıdaki bilgilere yer vermiştir:

Yapılan çalışmalarda genel olarak tüm süreçlerden benzer sonuçlar elde edilmiştir. Diğer taraftan, marjinal farklılıkların bulunması ile Lojit yaklaşımı bir adım öne çıkarılmıştır (Kobashigawa ve Berki, 1977; McCoy ve Manicke, 1978; Press ve Wilson, 1978; Talvitie, 1972). Lojitin farklılığı Powers vd. (1978) tarafından da belirtilmiştir. Deneysel kanıtlar ışığında, Probitin Lojitten üstünlüğü Werner, Wendling ve Budde (1978)'in çalışması ile sınırlı kalmıştır. McFadden (1974) ve Werner, Wendling, and Budde (1978) çalışmaları hariç, bu çalışmalarda örneklem hacminin farklı süreçlerin performansları üzerindeki etkisi açıklanmamıştır. McFadden'ın (1974) çalışmasında olasılık düzeylerinin etkisi üzerinde durulsa da, diğer çalışmalarda olasılık düzeylerinin etkisi sistematik olarak açıklanmamıştır.

Tablo 2.1'de yer alan çalışmalar, Bayes yaklaşımına dayanmayan tahmin süreçleriyle ilgili yapılan çalışmalardır. Malthora (1983), bu çalışmalar dışında, literatürde iki düzeyli veri için Bayesci yaklaşımı kullanan birçok çalışmanın yer aldığını ve tam ve yaklaşık Bayesci sonuç süreçlerini içeren bir çalışmanın da Zellner ve Rossi (1992) tarafından yapıldığını belirtmiştir.

Tablo 2.1 Konuyla ilgili yapılan bazı önceki çalışmalar (Malthora, 1983)

Yazar	Süreç	Örnek Hacmi	Kriter	Sonuçlar
Grablowsky ve Talley (1976)	PROBİT DİSK	Belli değil	Doğru sınıflandırma yüzdesi	Genel olarak, PROBİT kredi başvurularının mali bakımdan daha etkili bir sınıflandırma sağlar.
Gunderson (1980)	EKK, LOJİT PROBİT	1596	Subjektif karşılaştırma	LOJİT ve PROBİTten elde edilen sonuçlar çok benzerdir ve özellikle 0,5 olasılığı için EKK ile karşılaştırılabilir
Kobashigawa ve Berki (1977)	LOJİT, EKK, DİSK, PÇR	10000	Sınıflandırma uyum iyiliği	LOJİT ve DİSK, EKK ve PÇR' den marjinal olarak üstündür.
McCoy ve Manicke (1978)	LOJİT, CP REGRESYON	5192, 6167	Anlamlı değişkenliğin tanımlanması	LOJİTin daha etkili olduğu bulunmuştur.
McFadden (1974)	LOJİT, EKK	5- 200	Tahminlerin yanlılığı ve varyans	MLE(LOJİT) 60 ve daha büyük örneklem büyüklüğünde daha iyidir.
Powers vd. (1978)	LOJİT, PROBİT, DİSK.	585	Karışıklık (Confusion) matrisleri	LOJİT anlamlı derecede daha iyidir.
Press ve Wilson (1978)	LOJİT, DİSK.	40, 115	Sınıflandırma oranı	LOJİT marjinal olarak DİSK'ten daha iyidir.
Talvitie (1972)	LOJİT, PROBİT, DİSK	20	Sınıflandırma, Beklenen değer	LOJİT biraz farklı olmasına rağmen, süreçler karşılaştırılabilir sonuçlar sağlamıştır.
Werner,Wendling ve Budde (1978)	LOJİT, PROBİT, DİSK, EKK	25, 75, 250, 500	Yan, varyans ve HKO	PROBİT tüm örneklem büyüklüğünde LOJİTten iyidir. PROBİT, EKK ve DİSKten küçük örneklemler hariç hepsinde üstündür. Küçük örneklemlerde EKK ve DİSK daha iyidir.
Wilensky ve Rossiter (1978)	LOJİT, EKK	857, 573	Tahmin edilen katsayıların karşılaştırılması	İki süreç de benzer sonuçlar vermiştir.

Malthora'nın (1983) kendi çalışmasında ise, tüketici tercihlerini temsilen iki düzeyli verinin analizinde Lojit, Probit ve EKK regresyonunun performanslarını karşılaştırmayı amaçlanmıştır. Çalışmasında, bu üç sürece ait tanımlamalara, açıklamalara yer verilmiştir. Süreçleri karşılaştırmak için iki farklı deneysel veri seti kullanılmış ve sonrasında bir simülasyon çalışması gerçekleştirilmiştir.

İlk deneysel uygulama olarak, 1978 yılı Ulusal Enerji Tasarruf Politikası Yasasına göre belirlenen enerji sınırını aşan müşterilere enerji denetimi sunulması gerektiğinden, bu denetim önerilerine tüketici cevaplarını değerlendirmek için bir

çalışma yapılmıştır. Bu çalışmada, denetim önerilerine cevap vermede tüketicinin gönüllüğünü etkileyen beş önemli faktör dikkate alınmıştır. Bu faktörler; yatırım, finansman, geri ödeme periyodu, enerji tasarrufu ile yatırımın geri dönüşü ve vergi kredisidir. Bağımlı değişken ise gönüllü olmak ve gönüllü olmamak olarak iki düzeyli ifade edilmiştir. Üç modelin de veriye uygulanması sonucunda elde edilen bilgiler ışığında, EKK ve Probitin daha iyi olduğu sonucuna varılmıştır. EKK'nın Probitten daha iyi sonuç verdiği ancak bu farklılığın anlamlı olmadığı belirtilmiştir.

Yapılan diğer deneysel çalışmada ise, tüketicilerin ürün kalite anlayışları incelenmiştir. Temel olarak, tüketicinin kalite faktörlerini içeren kalite fonksiyonu ya da tüketicinin faydasının tahmini ele alınmıştır. Yapılan çalışmada; faydanın ya da kalitenin fonksiyonunun EKK, Lojit ve Probit Modelleri kullanılarak tahmin edilmesine ve bu süreçlerin doğru tahmin performanslarının karşılaştırmasına önem verilmiştir. Veri seti 1977 yılında Buffalo'da, New York Devlet Üniversitesine kayıt yaptıran 260 öğrenciden elde edilmiştir. Ayakkabıda kalite anlayışı incelenmek üzere, kalite anlayışını etkileyen faktörler; fiyat, taban malzemesi, renk, satış yeri ve üst malzemesi olarak belirlenmiştir. Yapılan analizler sonucunda, EKK, Probit ve Lojit benzer sonuçlar vermiştir. Süreçler arasında anlamlı farklılık görülmemiştir.

Son olarak Malhora'nın çalışmasında, farklı süreçlerin sistematik özelliklerini keşfetmek amacıyla bir simülasyon çalışması gerçekleştirilmiştir. Örneklem büyüklükleri 27, 54, 81, 108, 135, 162, 189, 216 olarak belirlenmiştir. Gözlenme olasılıkları olarak 0.1, 0.2, 0.5, 0.8 ve 0.9 değerleri dikkate alınmıştır ve her bir olasılık düzeyi ile örneklem hacmine karşılık 12 tekrar yapılmıştır. Simülasyon çalışmasında, 11 adet parametre tahmin edilmiştir.

Malthora (1983) yaptığı çalışmada, genel itibariyle; parametre sayısı 5 ile 10 arasında ve örneklem hacmi 54 ya da daha küçük olduğunda, EKK regresyonun iki düzeyli veri analizinde kullanışlı olduğu sonucuna varmıştır. Örneklem hacmi 108 ya da daha büyük olduğunda Lojit sürecinin en uygun süreç olduğunu ifade etmiştir. 54 ile 108 arasındaki örneklem büyüklüğünde iki sürecin tercih edilirliğinin değiştiği sonucuna varmıştır.

Bir modelin veriyi ne kadar iyi temsil ettiğinin ifadesi için ya da aynı veriye uygulanabilecek olan birden fazla modelin değerlendirilmesinde uyum iyiliği ölçütlerinden sıklıkla yararlanılmaktadır. Probit ve Lojit Modellerin de aynı verilere

uygulanabilir olmasından dolayı, iki modelin karşılaştırılmasında bu ölçütlere gerek duyulur. Ancak, EKK regresyondan farklı olarak Probit ve Lojit Modeller için bilinen R^2 istatistiği doğru bir istatistik değildir. Bu nedenle, bu modellerin değerlendirilmesinde farklı birçok yapay R^2'ler önerilmiştir. Bu aşamada, değerlendirmede önerilen ölçütlerden hangisinin tercih edilmesi gerektiği sorusuyla karşılaşılmıştır.

Tardiff (1976) tarafından yapılan çalışmada seçim modellerinde kullanılan uyum iyiliği ölçütleri incelenmiştir. Bu istatistikler içerisinden olabilirlik oran indeksi üzerinde durulmuştur. İki düzeyli ve çok düzeyli durumlar için, bu indeksin diğer uyum iyiliği istatistikleriyle karşılaştırmalı olarak matematiksel ve arzu edilir özelliklerinden bahsedilmiştir. İndeksin, yapılan önceki çalışmalarda, minimum değerinin sıfır olmaması ve örnek oranlarına bağlı olmasından ötürü farklı örneklerden elde edilen indeks sonuçlarını karşılaştırılmasını önlemesi gibi istenmeyen özelliklere sahip olduğu görülmüştür. Tardiff'in çalışmasında, bu zorlukları gidermek adına düzeltilmiş olabilirlik oran indeksine yer verilmiş ve bu ölçütün seçim modellerini değerlendirmede daha uygun bir indeks olduğu ileri sürülmüştür.

Gessner ve diğerleri (1988), iki düzeyli bağımlı değişken durumunda kullanılan, matematiksel olarak benzer birçok modeli dikkate alarak, deneysel bir araştırma yapmışlardır. Deneysel sonuçlar ışığında, iki düzeyli cevap değişkenli modeller için gerekli olan varsayımların gerçekleşmemesi durumunda parametre tahminlerinin, uyum iyiliği istatistikleri bu durumdan etkilenmemiş olsalar bile, yanıltıcı olabileceğini göstermişlerdir.

Hagle ve Mitchell'in (1992) Probit ve Lojit Modellerde kullanılan uyum iyiliği ölçütleriyle ilgili çalışmalarında; sürekli bir değişkenin, iki düzeyli hale getirilmesinden kaynaklanan bilgi kaybının yapay R^2'ler üzerindeki etkisi, bilinen R^2 ile karşılaştırılarak incelenmiştir. Lojit ve Probit Modellerde yaygın olarak kullanılan dört yapay R^2'yi (McKelvey ve Zavoina, Aldrich ve Nelson, Dhrymes, Achen) ve bilinen R^2'yi karşılaştırmak için bir simülasyon çalışması yapılmıştır. 100 adet veri seti üretilmiştir. Her veri seti 11 değişken ve 100 gözlem içermektedir. Bağımlı değişken, ortalaması sıfır, varyansı bir olan standart normal dağılımdan

türetilerek, sonrasında da 0'dan büyükse 1, küçükse 0 değeri verilerek iki düzeyli hale getirilmiştir. Bağımsız dokuz adet değişken ise sürekli bağımlı değişken ile farklı ilişki katsayılarına sahip olarak türetilmiştir (sırasıyla 0.1,0.2…). Ayrıca, sürekli değişkenin dağılımının az çarpık (25 adet bağımlı sürekli değişken, $-\infty$ ile 0 aralığında ya da 1.65 ile ∞ aralığında) ve daha çarpık (25 adet bağımlı sürekli değişken, $-\infty$ ile 0 aralığında ya da 1.96 ile ∞ aralığında) bir dağılıma sahip olmasının etkisi de dikkate alınmıştır. Yine benzer şekilde iki modlu bir dağılım olması durumu da dikkate alınmıştır (25 adet bağımlı sürekli değişken, $-\infty$ ile -1.65 aralığında ya da 1.65 ile ∞ aralığında bir moda sahip ve 25 tanesi $-\infty$ ile -1.96 aralığında ya da 1.96 ile ∞ aralığında bir moda sahip). Toplamda 200 adet veri setine önce EKK regresyon ve iki düzeyli hale getirilerek Lojit ve Probit uygulanmıştır. Sonuç olarak, McKelvey ve Zavoina'nın önerdiği R^2'nin, gerçek R^2 ile daha paralel sonuçlar verdiği bulunmuştur. Ayrıca, Aldrich-Nelson'un R^2'si için bir düzeltme önerilmiştir. Bu düzeltme ile elde edilen R^2'nin daha iyi sonuçlar verdiği belirtilmiştir.

Cox ve Wermuth (1992) çalışmasında, modelin uyumunun ifadesi olan belirlilik katsayısının, Lojit ve Probit regresyon için kullanılması halinde karşılaşılan durumlar belirtilmiştir. Lojit ve Probitten elde edilen sonuçların olasılık ifadesi olmaları sebebiyle sınırlı değerler alacağını ve dolayısıyla belirlilik katsayısı R^2'nin de sınırlanacağı ifade edilmiştir. Bunun sonucu olarak, önemli bir ilişki söz konusu olsa dahi, iki düzeyli cevap değişkenli modellerin geçerliliğini sınamada R^2'nin uygun olmayacağı belirtilmiştir.

Vell ve Zimmermann (1994) iki düzeyli Probit Modeller için kullanılan yapay R^2'leri değerlendirmişlerdir. Sınırlandırılmış verilerin analizinde genel olarak kabul edilmiş bir yapay R^2 olmayışının dezavantajlı olduğu belirtilmişlerdir. Kimi ölçütlerin, iki düzeyli bağımlı değişkenli regresyonların tahmini olasılıklarını ya da artıklarını değerlendirirken, bazılarının olabilirlik oran test istatistiğini temel aldığını ya da analiz için tahmin-gerçekleşme tabloları kullandığını ifade etmişlerdir. Çalışmanın amacı, yapay R^2 seçimi ve yorumlamasında araştırmacılara rehberlik etmektir. Bu amaçla çalışmada, simülasyon tekniğiyle elde edilen verilere iki düzeyli

Probit analizi uygulanmış ve McKelvey-Zavonia, Cragg-Uhler, McFadden, Lave, Aldrich-Nelson, normalleştirilmiş Aldrich-Nelson yapay R^2'leri ve korelasyon katsayısı bilinen R^2 ile karşılaştırılmak üzere hesaplanmıştır. Elde edilen sonuçlara göre, bilinen R^2'ye en yakın R^2'nin McKelvey ve Zavoina'nın önerdiği R^2 olduğu belirtilmiştir. Ayrıca Aldrich-Nelson'nun ölçüsünün önerilen normalleştirilmiş halinin hemen hemen McKelvey ve Zavoina'nın R^2'si kadar iyi sonuçlar verdiği ve kolay hesaplanabilir olduğu ifade edilmiştir.

3. GENEL BİLGİLER

İstatistiksel analizlerde model seçiminde değişkenlerin yapısı önem taşımaktadır. Örneğin; Lojit ve Probit Modellerde kullanılan bağımlı değişken kategorik olma özelliğine sahiptir. Bu bölümde, özellikle bu modellere ilişkin iki düzeyli bağımlı değişkenler üzerinde durulmuştur. Bu bölümde ilk olarak kategorik değişkenler hakkında bilgiler verilmiştir. Sonrasında ise iki düzeyli Probit ve Lojit Modellerde kullanılan gizli değişken yaklaşımı ele alınmıştır. Lojit ve Probit Modelleri içermesi sebebiyle, Genelleştirilmiş Doğrusal Modeller hakkında da genel bilgilere yer verilmiştir. Son olarak, ilgilenilen bu iki modelin tahmin sürecinde kullanılan En Çok Olabilirlik yöntemi üzerinde durulmuştur.

Kategorik değişkenler için yöntemlerin gelişimi sosyal ve biomedikal bilimlerdeki araştırma çalışmalarıyla hız kazanmıştır. Kategorik değişkenler ya da veriler sosyal ve biomedikal bilimlerde yaygın olmalarına rağmen, kullanımları sadece bu alanlarla sınırlı değildir. Bu değişkenler, davranış bilimleri, epidemiyoloji, halk sağlığı, genetik, zooloji, eğitim ve pazarlama alanlarında da sıklıkla kullanılmaktadır. Mühendislik ve endüstriyel kalite alanlarında dahi kategorik değişkenlerle karşılaşmak mümkündür (Agresti, 2002).

Değişkenler aldıkları değerlere bağlı olarak, sürekli ya da kesikli değişken olarak sınıflandırılır. Tüm değişkenlerin gerçek ölçümleri, ölçülürken sınırlandırılmalarına bağlı olarak kesikli bir yapıda olurlar. Pratikte, sürekli-kesikli sınıflandırma, değişkenleri çok değer alması ya da az değer almasına göre ayırt eder. Örneğin, istatistikçiler, aldığı değerlerin sayısı fazla olan kesikli aralık düzeyindeki değişkenleri (test skorları gibi) sürekli olarak alırlar ve onlar için sürekli cevap modelleri kullanırlar (Agresti, 2002).

Değişkenler nitel ve nicel değişkenler olarak ölçülebilirler. Nicel değişkenler gerçek sayısal değerleri ile ifade edilen değişkenlerdir. Nitel değişkenler ise cinsiyet, meslek gibi kategorileri olan ve gerçekte sayılarla değil nitel terimlerle ifade edilebilen değişkenlerdir. Nicel değişkenler sürekli ve kesikli değişkenler olarak sınıflandırılır. Sürekli değişkenler aralık düzeyli değişkenler olarak da adlandırılırlar. Nitel değişkenler kesikli değişkenlerdir. Ayrıca sınıflı, ya da sıralanmış olarak sınıflandırılabilirler. Sıralı değişken sıralanmış kategorilere sahiptir. Örneğin çok küçük, küçük, orta ve geniş kategorileri ile otomobil genişliği, yüksek, orta, düşük

kategorileri ile sosyal sınıf gibi. Anket sorularında, bireyin bir konu hakkında görüşü de, katılmıyorum, kısmen katılıyorum, katılıyorum şeklindeki cevaplarla sıralı olarak gruplanmış olabilir. Ayrıca bir olayın olma sıklığı da sık sık, genelde, nadiren ve asla olarak sıralı kategorik değişken olabilir. Sıralı değişkenlerde kategoriler arasındaki uzaklık bilinmez. Bu tür değişkenleri kullanan yöntemler, onların sırasından yararlanırlar.

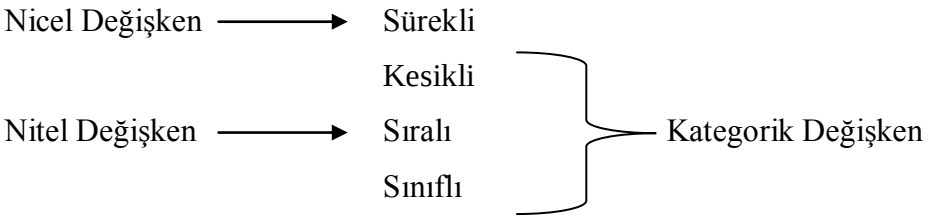
Sınıflı değişkenler sıralı olmayan çok düzeyli nitel değişkenlerdir. Değişkenler kendinden bir sıralamaya sahip olmaksızın kategoriktirler. Örneğin; ırk (beyaz, siyah, melez), medeni durum (bekar, evli, boşanmış, dul), siyasi partiler, markalar gibi. Sınıflı değişkenler için, kategorilerin sırası önemsizdir. Kategoriler arasındaki uzaklık sayısal değildir. İstatistiksel analizler de sıraya bağlı olmazlar. Sınıflı ve sıralı değişkenler arasındaki kesin çizgi her zaman çok açık olmamakla beraber araştırma sorularına bağlıdır. Aynı değişken kimi araştırmalarda sıralı olurken, bazı araştırmalar için sınıflı olabilirler.

Bir olayın meydana gelme sıklığını ifade eden değişken ise sıklık değişkenidir. Örneğin; bir bireyin yıl içerisinde doktora gitme sayısı, bir bilim adamının yıl içerisinde yayınlanan makale sayısı, bir ailenin sahip olduğu çocuk sayısı gibi.
Kategorik bir değişkenin ölçü düzeyi her zaman açık ve kesin olmayabilir. Örneğin eğitim değişkeni, yüksek eğitim mezunu olmak ya da daha alt düzey mezunu olmak olarak iki düzeyli bir değişken olabilir. Eğitim, değişken düzeyleri eğer yüksek eğitim, lisans düzeyi, orta öğretim şeklinde alınırsa, bu durumda sıralı bir değişken olacaktır. Yine kişinin aldığı eğitim süresi yıl olarak dikkate alınırsa, eğitim değişkeni sıklık değişkeni olarak ifade edilebilir (Long, 1997).

Uygulamalarda kullanılan bazı değişkenlerin açıkça kategorik değişken olduğu görülür. Örneğin; ırk, cinsiyet, medeni durum, doğum ve ölüm gibi. Ancak bazı durumlarda sürekli değişkenler kategorik olabilirler. Sürekli bir değişken kategorik bir değişken gibi davranması, kategorileştirme ya da ayrıklaştırma olarak adlandırılır. Kategorileştirme genellikle pratikte gereklidir. Örneğin yaş kavramsal olarak sürekli olmasına rağmen, açık ve pratik olması sebebiyle kategorik olarak dikkate alınır (Powers ve Xie, 1999). Bir araştırmada yaş değişkeninin 0-25, 25-60, 60 ve üstü şeklinde aralıklara bölünmesi bu değişkenin kategorik değişken olarak ele alındığını ifade eder.

Kategorik ve sürekli değişkenler birçok ortak özelliğe sahip olmalarına rağmen bazı farklılıklara da sahiptirler. Aralarındaki farklılık bağımlı değişken olduklarında dikkati çeker. Bağımlı bir değişken (cevap, çıktı ya da açıklanan değişken) bir çalışmada açıklanan ilgili kitle karakteristiğini temsil eder. Bunun tersine, açıklayıcı değişken olarak kullanıldıklarında farklılık küçüktür. Açıklayıcı (bağımsız) değişken ise bağımlı değişkendeki değişimi açıklamak için kullanılan değişkendir. (Powers ve Xie,1999).

Değişkenlerin yapılarına göre sınıflandırılması ve kategorik değişkenlerin ne tür değişkenleri içerdirği aşağıdaki diyagram ile özetlenmiştir.



Kategorik bir değişken özel olarak, iki kategoriye sahipse iki düzeyli(binary) değişken olarak adlandırılır. İki düzeyli değişken; bir olayın olup olmaması ya da bir özelliğe sahip olup olma durumunu temsil edebilir. İki düzeyli değişken araştırmacının yorumuna göre, kesikli, sıralı ya da sınıflı olabilir. Örneğin; bir bireyin bir malı satın alıp almama kararı, bir ailenin çocuk sahibi olup olmaması gibi. Cinsiyet gibi değişkenler iki düzeyli ve nitel değişkenlerdir. Bu değişken modele katılabilmek için 0 ya da 1 değerlerini alan yapay değişken olarak modele eklenirler. Bu gibi değişkenlere gölge (dummy) değişken adı verilir. Gölge değişkene alternatif olarak kukla, gösterge, iki değerli, gizli kategorik ya da dikotomi değişken adları da verilmektedir (Gujarati, 1999)

Analizlerde, açıklayıcı değişken ya da değişkenlerin koşulundaki bağımlı değişkenin (ya da dönüştürülmüş halinin) kitle ortalaması dikkate alınır. Bunun anlamı; bağımlı değişkenin açıklayıcı değişkenlere bağlı olmasıdır. Bu bağlılık, bağımlı değişkenin açıklayıcı değişkenlerin bir fonksiyonuyla ifade edilebilir. Analizlerde kullanılan farklı modeller bağımlı ve açıklayıcı değişkenler arasındaki bu fonksiyonların farklılığını da beraberinde getirebilir.

Bağımlı değişkenin sansürlenmiş, kesilmiş, iki düzeyli, sıralı, sınıflı ya da sıklık düzeyli olduğu durumlarla sıkça karşılaşılmaktadır. Bu şekildeki değişkenler, kategorik ya da kısıtlı bağımlı değişken olarak dikkate alınabilir. Kategorik değişkenler bir dizi kategorilerden oluşan bir ölçüm ölçeğine sahiptirler. Böyle değişkenlere uygun olarak farklı regresyon modelleri geliştirilmiştir.

Doğrusal regresyon modelleri birçok bilim dalında yaygın olarak kullanılan istatistiksel yöntemlerdir. Regresyon modelinde, bağımlı değişkenin sürekli olduğu ve normal dağıldığı varsayılır. Ancak, gerçekte ilgilenilen çoğu bağımlı değişkeni sürekli olmayabilir ya da normal dağılmayabilir (Long, 1997).

Bağımlı değişken olarak kullanılacak değişkenin ölçüm tipini belirlemek, uygun analiz yöntemini belirmede önemli bir rol oynar. Eğer seçilen model doğru ölçü düzeyini dikkate almıyorsa, elde edilen tahminler yanlı, etkin olmayan ya da tamamen tutarsız olur. Kategorik ve kısıtlı bağımlı değişkenler için özellikle tasarlanmış çok fazla model vardır. Örneğin; iki düzeyli Lojit ve Probit Modeller iki düzeyli çıktılar için uygundur. Sıralı Lojit ve Probit Modeller, bağımlı değişkenin sıralı olduğu durumu ele alırlar. Çok düzeyli Lojit Model, sınıflı çıktılar için uygundur. Tobit model sansürlenmiş veriler için tasarlanmıştır. Poisson ve negatif binom regresyon gibi çeşitli modeller sıklık düzeyli çıktılar için kullanılabilirler (Long, 1997).

Doğrusal regresyon modelinde, açıklayıcı değişkenin nitel ya da nicel olması ile ilgili bir kısıt yokken, bağımlı değişkenin sürekli olma şartı vardır. Çünkü x_k , β_k ya da u_i üzerinde sınır olmaması, Y_i'nin eksi sonsuz ile artı sonsuz arasında değerler almasını gerektirir. Aslında pratikte Y_i'nin daha küçük bir aralıkta değerler alacak şekilde sınırlandırılması süreklik koşulunu bozmayabilir. Örneğin, Y_i bağımlı değişkeni kişi gelirini temsil edecek olursa, sadece belirli bir aralıkta değerler alabilecektir. Bunun sebebi, açıklayıcı değişkenlerin de benzer şekilde sınırlı değerlere sahip olmasıdır. Sonuç olarak Y_i'nin sınırlı da olsa belirli bir aralıkta sürekli olması varsayım ihlaline sebep olmaz. Ancak, Y_i'nin nitel olması ya da bir gölge değişken olması durumu süreklilik varsayımının ihlaline sebep olur. Bu durumda doğrusal regresyon modelini kullanmak doğru değildir. Onun yerine, Lojit ve Probit Modeller kullanılabilir.

3.1. Gizli Değişken Yaklaşımı

Çoğunlukla davranışsal bilimlerde karşılaşılmak üzere, pratikte doğrudan gerçek değerleri ölçülemeyen bazı kavramlara ilişkin bilgi sahibi olunmak ve istatistiksel analizler yapılmak istenebilir. Bunun yanı sıra bazı iki düzeyli değişkenler için, sürekli olan başka bir değişkenin etkisi altında olduğu durumlar olabilir. Bu gibi durumlar gizli (latent) bir değişken yaklaşımını ortaya çıkarır.

Bu yaklaşımdaki temel düşünce, gözlenen kategorik değişkenin ardında gözlenemeyen ya da gizli bir sürekli değişkenin varlığıdır. Gizli değişken bir eşik değeriyle kesilirse, gözlenen kategorik değişken farklı bir değer alır. Gizli değişken yaklaşımına bağlı olarak, kısmen gözlenebilen sürekli dağılan değişkenlerden farklı kategorik değişkenler elde edilir. Bu yüzden gözlenen kategorik değerler ile gizli değişkenin gerçek değerleri hakkında yorum yapılamaz. Bu sebeple ekonometriciler kategorik değişkenleri kısıtlı bağımlı değişken olarak isimlendirmişlerdir (Maddala, 1983).

Araştırmacıların gizli değişken yaklaşımında teorik olarak ilgilendikleri açıklayıcı değişkenin sürekli olan bağımlı değişken üzerindeki etkisidir. Örneğin; iktisadi bilimlerde, kişilerin tercihlerini ortaya çıkarmaya yönelik olarak "fayda" kavramından yararlanılır. Kişi tercihlerini dikkate alan çalışmalarda, fayda değişkeni gizli bir değişkendir ve gerçekte gözlemlenemez. f faydanın simgesini, e ise kriter ya da eşik değerini göstersin. Söz gelimi, bir kişinin bir ürünü alıp-almama tercihini yapma durumu dikkate alınsın. Bu durumda, kişi tercih yaparken ürünü satın almada elde ettiği faydaya göre karar verir ve bu fayda belirli bir eşik değerinin üstünde olması durumunda alma kararını verecektir.

Gerçekte f ve e doğrudan gözlemlenebilir değişkenler değillerdir. Bu gizli değişkenlerin yerine $Y=1$ ya da $Y=0$ değerlerini alan bağımlı değişkenler gözlenir. Bağımlı değişkeninin 1 olarak gözlenmesi bireyin fayda fonksiyonunun eşik değerinden yukarıda olması anlamına gelir. Dolayısıyla bağımlı değişkenin 1'e eşit olma olasılığı ile fayda fonksiyonunun kriterden büyük olma olasılığı eşitlenebilir.

$$P(Y=1)=P(f>e) \tag{3.1}$$

Gizli değişkenler ve açıklayıcı değişkenler arasındaki ilişki doğrusaldır ve aşağıdaki gibi ifade edilebilir:

$$f = \sum_{k=0}^{K} \beta_k^f x_k + u^f \tag{3.2}$$

$$e = \sum_{k=0}^{K} \beta_k^e x_k + u^e \tag{3.3}$$

Burada β_k^f ve β_k^e sırasıyla, f ve e bağımlı değişkenleri için *k*. bağımsız katsayıyı ve u^f ile u^e eşitlikler için hatayı temsil etmektedir. (3.2) ve (3.3) eşitlikleri (3.1)'de yerlerine konursa eşitlik (2.4) elde edilir

$$\begin{aligned} P(Y=1) &= P(f > e) \\ &= P(u^e - u^f < \sum_{k=0}^{K} \beta_k^f x_k - \sum_{k=0}^{K} \beta_k^e x_k) \\ &= P(u^e - u^f < \sum_{k=0}^{K} (\beta_k^f - \beta_k^e) x_k) \end{aligned} \tag{3.4}$$

Burada karşılaşılan temel problem f ve e'nin kendilerinin değil bileşenlerindeki farkın tanımlanabiliyor olmasıdır. Modeli tanımlamak için, bazı sınırlamaların getirilmesi gerekir. Bunu yapmanın en kolay yolu; f ya da e'yi 0'a eşitlemektir.

$$\beta_k^f = 0 \text{ , } (k = 1, ..., K) \qquad u^f = 0 \tag{3.5}$$

ya da

$$\beta_k^e = 0 \text{ , } (k = 1, ..., K) \qquad u^e = 0 \tag{3.6}$$

Eşitlik (3.5) sabit kriterli rasgele faydalı bir modele karşılık gelir. Yine benzer şekilde (3.6) eşitliği; değişen kriterli sabit faydalı bir model oluşturur. Bu iki tanımlama farklı yorumlara sebep olsa da, istatistiksel olarak ayrılamazlar. Bu ikisi bir eşitlikte birleştirilerek ikisini de kapsayacak şekilde aşağıdaki gibi ifade edilebilir (Powers ve Xie, 1999);

$$u = u^e - u^f \tag{3.7}$$

$$\beta_k = \beta_k^f - \beta_k^e \text{ , } (k = 1, ..., K) \tag{3.8}$$

(3.7) ve (3.8) eşitlikleri (3.4) eşitliğinde yerine konularak aşağıdaki ifade elde edilir.

$$P(Y=1)=P(u<\sum_{k=0}^{K}\beta_k x_k)=F(\sum_{k=0}^{K}\beta_k x_k) \quad (3.9)$$

Eşitlikte yer alan $F=(.)$, u'ya ait kümülatif dağılım fonksiyonunun (kdf) ifadesidir. β parametrelerinin değerlerini belirlemek için u'nun ortalamasının ve varyansının standartlaştırılması gerekir. Bundan dolayı, iki düzeyli bağımlı değişkenler sürekli bir değişkenden farklı olarak doğal bir ölçüye sahip değildir. İki düzeyli bağımlı değişken durumunda β parametre vektörünün elemanları, u'nun ortalaması ve varyansına bağlıdır. u'nun ortalaması ve varyansı bilinmediğinden, $u^*=(u-a)/b$ olarak tekrar tanımlanabilir. Burada a ve b seçilen sabitlerdir. Eşitlik (3.9)'da u yerine u^* konularak

$$P(Y=1)=F^*(\frac{\sum_{k=0}^{K}\beta_k x_k-a}{b}) \quad (3.10)$$

elde edilir. $F^*=(.)$ u^*'ın kdf'sidir. Eğer $a=\mu_u$, $b=\sigma_u$ ve $F^*=\Phi(.)$ eşitlenirse (3.10) eşitliği Φ standart kümülatif normal dağılımı göstermek üzere, ileride daha detaylı olarak ele alınacak olan Probit Modele dönüşür. Benzer olarak, Lojit Model de u'nun lojistik dağılıma sahip olduğu varsayılarak elde edilir (Powers ve Xie, 1999).

3.2. Gizli Değişken Yaklaşımının Genişletilmesi

Gizli değişken yaklaşımı genellikle birey düzeyli verilerin analiziyle ilgilidir. Bir i bireyi için, gizli sürekli değişken Y_i^* tanımlanmış olsun. Bu değişken fayda değişkenine karşılık gelmektedir. Bu değişkenin x_{ik} açıklayıcı değişkenleri ve u_i artıklarıyla olan doğrusal fonksiyonu genel ifadeyle aşağıdaki gibi verilebilir:

$$Y_i^*=\sum_{k=0}^{K}\beta_k x_{ik}+u_i \quad (3.11)$$

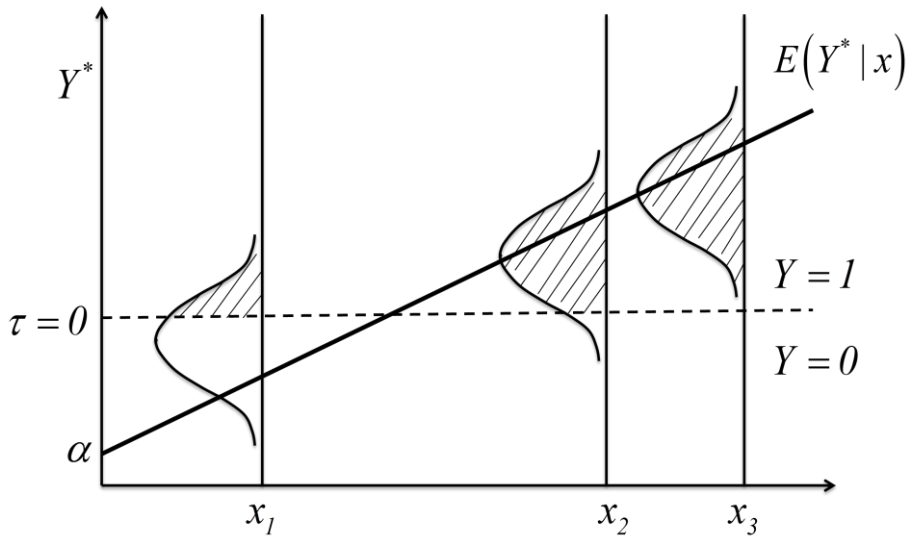
Burada Y_i^* gizli değişkendir, yani gözlenemeyen bireyin faydasını temsil eder. Bu eşitlik tamamen regresyon eşitliğine benzerdir ve tüm bileşenlerin yorumu aynıdır.

Ama β_k gözlenen kesikli değişken Y_i'nin değil, sürekli ve gözlenmeyen değişken Y_i^* üzerinde x_{ik}'nın değişim etkisini ölçmektedir (Aldrich ve Nelson,1984).

Gizli değişken, gözlenen iki düzeyli Y değişkeni ile aşağıdaki bağıntıya sahiptir ve τ eşik noktasını ifade etmektedir.

$$Y_i = \begin{cases} 1 & Y_i^* > \tau \\ 0 & Y_i^* \leq \tau \end{cases} \tag{3.12}$$

Y^* ve gözlenen Y arasındaki ilişki $Y^* = \alpha + \beta x + u$ modeli için Şekil 3.1'de gösterilmiştir. Burada gizli değişken, y ekseninde yer almış ve eşik noktası da kesikli çizgilerle ifade edilmiştir.



Şekil 3.1. İki düzeyli modelde verilen *x* için Y^*'ın dağılımı (Long, 1997)

Şekil 3.1'de Y^*'ın dağılımı çan eğrisi olarak ifade edilmiştir. Y^*'ın eşik değerini geçtiği noktalarda $Y = 1$ değerini alır ve bu kısım taralı alanlara karşılık gelir. Örneğin, x_1'in yaklaşık %25'i, x_2'nin yaklaşık %90'ı, x_3'ün tamamı 1 değerini alır. Y^*'ın eşik değerini geçemediği durumda ise $Y = 0$ olacaktır.

Y^* sürekli olduğu için, model seçimi olarak Doğrusal Olasılık Modeli nin tercih edilmesi durumunda, daha sonra detaylı olarak ele alınacak olan, bu modelde

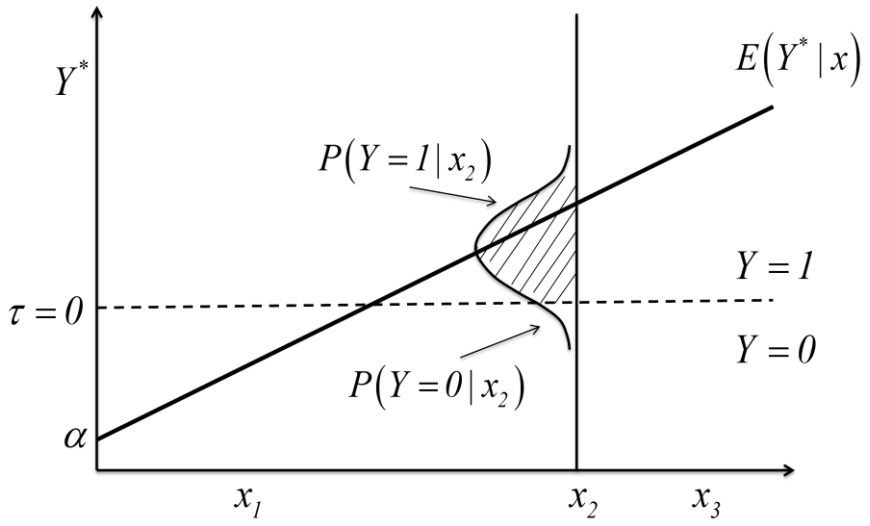
meydana gelebilcek sorunlarla karşılaşılmaz. Ancak, bağımlı sürekli Y^* değişkeninin gözlenemediği durumlarda modelin tahmini en küçük kareler tekniği ile yapılamaz. Bunun yerine, hata dağılımı ile ilgili varsayımları gerektiren En Çok Olabilirlik tahmini kullanılır. Çoğunlukla hataların normal dağılıma sahip olduğu varsayıldığında buna uygun olarak kullanılacak model Probit Model ve lojistik dağılım olduğunda ise kullanılacak model Lojit Modeldir. Doğrusal regresyon modelinde olduğu gibi hatanın beklenen değeri $E(u \mid x) = 0$'dır (Long, 1997).

Doğrusal olmayan tüm olasılık modellerinde gözlenemeyen Y^* ve x_{ik} arasında doğrusal bir ilişki temel alınır. Y^* aralık düzeyinde ya da daha genel ifadeyle sürekli bir değişken olarak gözlemlenemez ya da en azından ölçülemez. Onun yerine Y^*'ın bir fonksiyonu olarak iki düzeyli Y_i ölçülebilirdir. Seçilecek modelin hangi moldel olacağı hata terimi u_i'nin dağılımına bağlıdır. Bu nedenle, özel olarak doğrusal olmayan bir model seçmek aynı zamanda u_i'nin dağılımını seçmektir. Problemin doğası gereği hatanın gerçek dağılımının ne olduğu bilinemez ve genellikle seçilen dağılıma sahip olduğu varsayılır. Bu dağılımlar arasında en çok kullanılan dağılımlar normal ve lojistik dağılımlarıdır (Aldrich ve Nelson, 1984).

Y_i^* gözlemlenebilir olmadığı için, doğrusal regresyon modelinde olduğunun aksine hataların varyansları tahmin edilemez. Probit Modelde $Var(u \mid x) = 1$ olduğu, Lojit Modelde ise $Var(u \mid x) = \pi^2 / 3 \approx 3.29$ olduğu varsayılır. Bunun sebebi şu şekilde açıklanabilir; Y_i^* doğrudan gözlemlenmediğinden, ölçüsüne karar verilemez. Yukarıdaki eşitlik, Y_i^*'ın işaretini değiştirmeksizin keyfi pozitif sabit bir sayıyla çarpılırsa, böyle bir değişim nitel çıktı değişkeni Y_i'nin gözlemi üzerine herhangi bir etki yaratmayacaktır. Çünkü Y_i'nin alacağı değer Y_i^*'ın büyüklüğüyle değil işaretiyle ilgilidir. Böylece hem Probit, hem Lojit analiz, keyfi ölçümlerin dışındakileri belirlemek için normalleştirmeyi kullanırlar. Probitte uygun normalleştirme 1'dir, yani u_i'nin standart sapmasını 1 olarak belirler. Lojit analizde ise 1.8138 $(\pi / \sqrt{3})$ 'dir (Aldrich ve Nelson, 1984).

Normal ve lojistik dağılımlar çok benzer dağılım eğrilerine sahiptirler. Dolayısıyla, Probit ve Lojit Modeller de oldukça benzer modellerdir. İki model

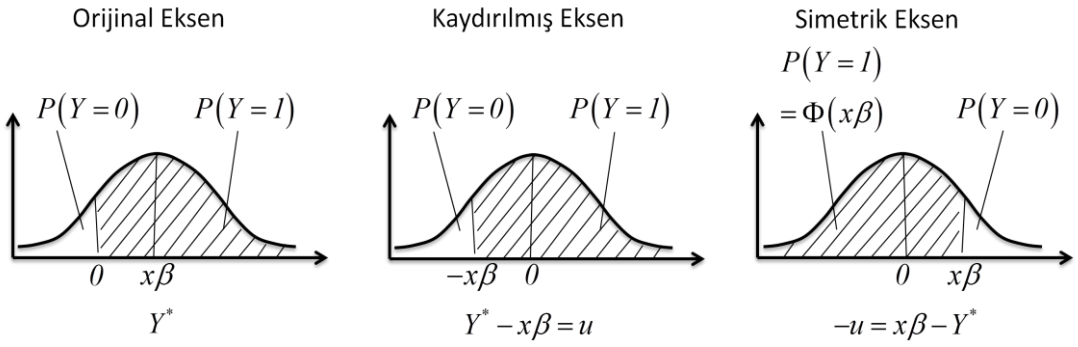
arasında birini diğerine tercih etmek için teorik bir alt yapı veya sebep yoktur. Lojit, olasılık dağılım ve yoğunluk fonksiyonun Probite göre daha basit olmasından dolayı daha çok tercih edilmektedir. Probit Modelde odds oranları için basit bir kapalı form yoktur. Lojit Model log odds oranlarını kullanarak terimlerin daha kolay yorumlanabilmesini de sağladığından daha çok kullanılmaktadır (Powers ve Xie, 1999).



Şekil 3.2 İki düzeyli cevap değişkenli modelde gözlenen değerlerin olasılıkları

(Long, 1997)

Verilen bir x için, u'ya ilişkin bir dağılım varsayılarak, $Y=1$ olma olasılığının hesaplanması mümkündür. Şekil 3.2 incelendiğinde, u $E(Y^* \mid x)=\alpha+\beta x$ etrafında lojistik ya da normal dağılır. Şekilde, $Y=1$ değerleri u dağılımının kesim doğrusu üzerinde kalan taranmış bölgede gözlenecektir. $E(Y^* \mid x)$, $Y=1$ olduğu taralı bölgede olsa bile, eğer u büyük ve negatifse, 0 olarak gözlemlenebilir. Negatif hatalar, Y^*'ı eğrinin taranmamış bölgesine kaydırır.



Şekil 3.3 İki düzeyli cevap değişkenli modelde $P(Y=1|x)$'nın hesaplanması (Long, 1997)

Şekil 3.3, $P(Y=1|x)$ olasılığını hesaplamak için bu düşüncenin dönüşümünü göstermektedir. Birinci şekil hata dağılımını göstermektedir. $Y^* > 0$ olduğu zaman $Y=1$ değerini alır ve aşağıdaki şekilde ifade edilir:

$$P(Y=1|x) = P(Y^* > 0|x) \tag{3.13}$$

$Y^* = x\beta + u$ yerine konulursa;

$$P(Y=1|x) = P(x\beta + u > 0|x) \tag{3.14}$$

Eşitsizliğin her iki tarafından $x\beta$ çıkarılması ile kaydırılmış eksen şekli ortaya çıkar. Olasılıksal eşitlik ise aşağıdaki gibi ifade edilir:

$$P(Y=1|x) = P(u > -x\beta|x) \tag{3.15}$$

Kdf, bir rasgele değişkenin verilen bir değerden küçük ya da eşit olma olasılığının ifadesi olduğu için eşitsizliğin yönünün değiştirilmesi gerekir. Normal ve lojistik dağılımlar simetrik dağılımlar olduğundan, şekil 3.3'teki ikinci eğrideki $-x\beta$ 'dan büyük taranmış alan ile üçüncü eğrideki $x\beta$'dan küçük taranmış alanlar birbirine eşittir. Sonuç olarak (3.15) eşitliği (3.16) eşitliği ile ifade edilebilir.

$$P(Y=1|x) = P(u \le x\beta|x) \tag{3.16}$$

Hata dağılımının kdf'si $x\beta$'ya göre değerlendirilir ve eşitlik (3.17)'deki gibi yazılabilir.

$$P(Y=1|x) = F(x\beta) \tag{3.17}$$

Burada F Probit Model için normal dağılım fonksiyonu (Φ) ve Lojit Model için lojistik dağılım fonksiyonu (Λ) olmuştur. Verilen x için bir olayın gözlenmesi olasılığı, $x\beta$'da değerlenmiş kümülatif yoğunluktur (Long, 1997). Bu ifade iki düzeyli regresyon modeli için tek açıklayıcı değişken varken eşitlik (3.18)'deki gibi yazılır:

$$P(Y=1|x)=F(\alpha+\beta x) \tag{3.18}$$

3.3. Dönüşümsel Yaklaşım ve Genelleştirilmiş Doğrusal Modeller

Dönüşümsel ya da istatistiksel yaklaşımda, kategorik verinin doğası gereği kategorik olduğu ve bu şekilde modellenebileceği düşünülür. Bu yaklaşımda, ilgilenilen kitle parametreleri ve örnek istatistikleri arasında doğrudan bire bir uyuşma vardır. Amaç, örnek istatistiğine karşılık gelen kitle parametrelerini tahmin etmektir. Ele alınan değişken gizli ya da gözlenmemiş değildir. Dönüşümsel yaklaşımda istatistiksel modelleme; belirli dönüşümlerden sonra, kategorik bağımlı değişkenin beklenen değerinin açıklayıcı değişkenlerin doğrusal bir fonksiyonu olarak ifade edilmesi ile gerçekleşir (Powers ve Xie, 1999).

Kesikli (sıklık) verilerinin analizinde beklenen frekanslar (ya da hücre sıklıkları) negatif olmamalıdır. Regresyon modellerinden elde edilen tahmini değerlerin bu kısıta uyduğundan emin olmak için, bağımlı değişkenin beklenen değeri doğal logaritma (ya da log bağ) fonksiyonu kullanılarak dönüştürülür. Bu sayede, logaritması alınan sıklık, açıklayıcı değişkenlerin doğrusal bir fonksiyonu olarak ifade edilebilir. Bu logaritmik doğrusal dönüşüm iki amaca hizmet eder. Tahmin edilen değerlerin sıklık verisine uygun olduğunu garanti eder ve bilinmeyen regresyon parametrelerinin tümüyle gerçek uzay (parametre uzayı) içinde yer almasına izin verir (Powers ve Xie, 1999).

İki düzeyli (binomial) cevap değişkenli modellerde, tahmini olasılıklar [0,1] aralığında olmak zorundadır. Eğer açıklayıcı değişkenlerin değerlerine kısıt getirilmezse her hangi bir doğrusal fonksiyon, aralığın ihlal edilmesine sebep olabilir. Bu aralıkta olasılıkları doğrudan kullanan bir model kurmak yerine, olasılıkların dönüşümlerinin $(-\infty,+\infty)$ aralığında olmasını sağlayan bir model kurulabilir. Olasılıkları dönüştüren çeşitli yollar vardır. Örneğin, Lojit dönüşüm,

$\log[p/(1-p)]$ olasılıkların dönüşümünde kullanılabilir. Benzer şekilde Probit dönüşümü $\Phi^{-1}(p)$ de bu amaçla kullanılır. Probit bağ beklenen olasılıkların $(-\infty,+\infty)$ aralığına dönüştürülmesinde standart normal dağılım fonksiyonundan faydalanır. Lojit ve Probit dönüşümlerinin her ikiside tahmini olasılıkların tüm olası parametre ve açıklayıcı değişken değerleri için uygun aralıkta olduğunu garanti eder Kategorik bağımlı bir değişkenin yapısı nedeniyle kurulmak istenen regresyon fonksiyonu doğrusal olamaz. Doğrusal olmama problemi, kategorik değişkenin beklenen değerini açıklayıcı değişkenlerin doğrusal bir fonksiyonuna dönüştüren doğrusal olmayan fonksiyonlarla ele alınır. Böyle dönüşüm fonksiyonları bağ (link) fonksiyon olarak adlandırılırlar. Bağ fonksiyonuyla doğrusal modellere dönüştürülen modellere Genelleştirilmiş Doğrusal Modeller (GDM) adı verilir (Powers ve Xie, 1999).

GDM doğrusal regresyon ve varyans analizi modellerini, nitel verilerde kullanılan Lojit, Probit Modellerini, log-doğrusal modelleri, sıklıklar için kullanılan çok düzeyli modelleri ve yaşam verilerinde yaygın olarak kullanılan modelleri içerir. Bu modeller doğrusallık gibi ortak özelliklere sahiptirler (McCullagh ve Nelder, 1989).

GDM klasik doğrusal modellerin genişletilmiş halidir ve tüm olasılık modelleri GDM başlığı altında sınıflandırılabilir.

3.3.1. Genelleştirilmiş doğrusal modellerin bileşenleri

GDM üç bileşenden oluşurlar. Birincisi rasgele bileşendir. Bağımlı değişken Y 'yi ve onun olasılık dağılımını tanımlar. İkinci bileşen sistematik bileşendir ve doğrusal tahmin edici fonksiyonunda kullanılan açıklayıcı değişkenleri tanımlar. Üçüncü bileşen bağ fonksiyonudur ve sistematik bileşeni modele eşitleyen $E(Y)$ 'nin fonksiyonunu tanımlar (Agresti, 2002). Bağ fonksiyonu, rasgele ve sistematik bileşen arasındaki bağdır.

GDM'nin rasgele bileşeni üstel ailesine ait bir dağılımdan, $(Y_1,...,Y_N)$ bağımsız gözlemleri ile Y bağımlı değişkenini içerir. Bu ailenin olasılık yoğunluk fonksiyonu aşağıdaki fonksiyonel yapıya sahiptir.

$$f(Y_i;\theta_i) = a(\theta_i)b(Y_i)\exp(Y_iQ(\theta_i))$$

θ_i parametresinin değeri i 1'den N'ye değişmek üzere, açıklayıcı değişkenlerin değerlerine bağlı olarak değişir. $Q(\theta)$ terimi doğal parametre olarak adlandırılır (Agresti, 2002).

Bir Y değişkeni için K adet bilinmeyen parametre ve açıklayıcı değişkenin doğrusal birleşimi $E(Y) = \mu = \sum_{k=1}^{K} \beta_k x_{ik}$ şeklinde verilir. Bu eşitlik bilinen doğrusal modeli ifade eder ve doğrusal regresyon modelini çağrıştırır. Daha genel bir model oluşturmak için $\sum_{k=1}^{K} \beta_k x_{ik}$ fonksiyonunu μ'ye bağlayan η değişkeni tanımlanır ve bunun doğrusal bir şekli olmasına gerek yoktur. Genel olarak, η açıklayıcı değişenler tarafından üretilen doğrusal bir tahmin edici olarak tanımlanır. Modelin türüne bakılmaksızın, açıklayıcı değişkenler kümesi her zaman doğrusal olarak η değişkenini üretir (Liao, 1994).

GDM'nin sistematik bileşeni η'dür ve açıklayıcı değişkenlerle arasındaki ilişki aşağıdaki şekilde ifade edilir. Açıklayıcı değişkenlerin eşitlikteki doğrusal birleşimi doğrusal tahmin edici olarak adlandırılır.

$$\eta_i = \sum_k \beta_k x_{ik}, \ i = 1, ..., N \tag{3.19}$$

Bu eşitlikte, tüm i değerlerinde $x_{ik} = 1$ olması modeldeki sabit katsayıyı verir. Sabit katsayının ayrıca α olarak gösterimi de kullanılmaktadır.

GDM'nin üçüncü bileşeni olan bağ fonksiyonu, rasgele ve sistematik bileşenleri birbirine bağlar. $\mu_i = E(Y_i), \ i = 1, ..., N$ olmak üzere, model μ_i'yi η_i'ye $\eta_i = g(\mu_i)$ şeklinde bağlar. Burada g monotonik, türevlenenebilen bir fonksiyondur ve bu fonksiyon $E(Y_i)$'yi açıklayıcı değişkenlere eşitlik (3.20) ile bağlar.

$$g(\mu_i) = \sum_k \beta_k x_{ik}, \ i = 1, ..., N \tag{3.20}$$

$g(\mu) = \mu$ bağ fonksiyonu, özdeş bağ olarak adlandırılır. Burada $\eta_i = \mu_i$'dir. Bağ fonksiyonun ortalamayı doğal parametreye dönüştürmesine kanonik bağ denir ($g(\mu_i) = Q(\theta_i)$ ve $Q(\theta_i) = \sum_k \beta_k x_{ik}$) (Agresti, 2002).

Y rasgele değişkeni ile doğrusal bir model kullanarak çalışılıyorsa, bu değişkenin beklenen değeri K adet açıklayıcı değişkenin doğrusal bir fonksiyonu olarak tanımlanır ve aşağıdaki gibi ifade edilir.

$$E(Y) = \mu = \sum_k \beta_k x_{ik} \tag{3.21}$$

Bu eşitlik doğrusal regresyon modeli için geçerlidir. Daha genel bir modeli elde etmek için, doğrusal bir biçimi olması gerekmeyen, $\sum_k \beta_k x_{ik}$ fonksiyonunu μ'ye bağlayan η değişkeni kullanılır. Genelde, η $x_1, x_2, ..., x_K$ açıklayıcı değişkenleri tarafından üretilen doğrusal bir tahmin edici olarak tanımlanır. Modelin türüne bakılmaksızın açıklayıcı değişkenler doğrusal olarak, Y'nin tahmin edicisi η'yü üretirler. η ve x değişkenleri arasındaki ilişki eşitlik (3.22) şeklinde verilir.

$$\eta = \sum_k \beta_k x_{ik} \tag{3.22}$$

η ve μ arasındaki bağ, genel doğrusal model üyeleri arasında farklılık gösterir. η ve μ arasında birçok olası bağ fonksiyonu vardır. Bazı bağ fonksiyonları aşağıdaki gibi verilebilir (Liao, 1994).

1- Doğrusal(Özdeş) Bağ Fonksiyonu

Bu bağ fonksiyonu eşitlik (3.21) ve (3.22)'in birbirine eşit olduğu anlamına gelir ve bilinen klasik doğrusal modellerin tanımlanmasında kullanılır.

$\eta = \mu$

2- Lojit Bağ Fonksiyonu

Lojit bağ fonksiyonu, eşitlik (3.21)'e uygulandığında, iki düzeyli bağımlı değişkenine sahip bir Lojit Model tanımlanmış olur.

$\eta = \log[\mu / (1 - \mu)]$

3- Probit Bağ Fonksiyonu

Eşitlikte kullanılan Φ^{-1}, standart normal kdf'nin tersini ifade etmektedir. Bu bağ fonksiyonu benzer şekilde bir Probit Modeli tanımlar.

$$\eta = \Phi^{-1}(\mu)$$

4- Logaritmik Bağ Fonksiyonu

Bu bağ fonksiyonunda doğal logaritma kullanılır ve poisson regresyon modelini ifade eder.

$$\eta = \log \mu$$

5- Çok düzeyli (Multinomial) Lojit Bağ Fonksiyonu

J toplam düzey sayısını, j, j. düzeyi temsil etmek üzere $(j = 1,..., J)$, bu bağ fonksiyonu Lojit bağ fonksiyonunun genel halidir.

$$\eta_J = \log(\mu_j / \mu_J)$$

2. ve 3. bağ fonksiyonları binomial dağılımı temel alırlar. Sosyal bilimlerde kullanılan değişkenlerin çoğu bu dağılımı gösterir. Örneğin; oy verme-vermeme, ölme-yaşama, katılma-katılmama, göç etme-etmeme, bir olayın ortaya çıkıp çıkmaması gibi durumların hepsi binomial dağılıma uyar. Lojit ve Probit Modeller çoğunlukla bu tip olaylar üzerinde modellenir (Liao, 1994).

Bağ fonksiyonunun ya da istatistiksel modelin seçimi, araştırmacıya verinin doğasını anlama imkanı veren teoriye ve verinin dağılımına bağlıdır. Özellikle, Y 'deki rasgele değişimin dağılımı (x değişkenleri tarafından sistematik olarak açıklanamayan) bağ fonksiyonuna ve GDM'nin türüne bağlıdır. GDM'de rasgele bileşenin dağılımı üstel aileden gelmektedir. Bunlar arasında normal, binomial, poisson ve bazı diğer dağılımlar bulunmaktadır. Bu dağılım normal olduğu zaman bağ fonksiyonu doğrusaldır. En küçük kareler regresyon uygulamalarının tümünde Y 'nin sürekliliğini ifade eden rasgele bileşeni normallik varsayımına sahiptir (Liao, 1994). Yukarıda adı geçen bağ fonksiyonları ve bu fonksiyonları kullanan modellere ilişkin tablo aşağıda verilmiştir.

Tablo 3.1 Genelleştirilmiş doğrusal modellerde yer alan modeller ve kullanılan bağ fonksiyonları (Uçar, 2004)

Rasgele Bileşen	**Kanonik Bağ**	**Kanonik Bağ Adı**	**Sistematik Bileşen**	**Model**
Normal	$\eta = \mu$ $E(Y) = \mu$ iken	Özdeş	Sürekli	Doğrusal Regresyon
Binom	$\eta = \log[\mu / (1-\mu)]$ $E(Y) = \mu$ iken	Lojit	Karışık	Lojistik Regresyon
Binom	$\eta = \Phi^{-1}(\mu)$ $E(Y) = \mu$ iken	Probit	Karışık	Probit Regresyon
Poisson	$\eta = \log \mu$ $E(Y) = \mu$ iken	Logaritma	Karışık	Poisson Regresyon
Çok Düzeyli	$\eta_J = \log(\mu_j / \mu_J)$ $E(Y) = \mu$ iken	Genelleştirilmiş(Çok düzeyli) Lojit	Karışık	Çok Düzeyli Cevap

3.4. Regresyon ve Olasılık Modelleri

Yapılan birçok istatistiksel araştırmada iki veya daha fazla değişken arasındaki ilişkinin incelenmesi söz konusudur. Aralarında sebep sonuç ilişkisi bulunan iki veya daha fazla değişken arasındaki ilişkiyi kullanarak konuyla ilgili tahminler ya da kestirimler yapabilmek için en çok kullanılan analizlerden biri regresyon analizidir. Bu analiz tekniğinde değişkenler arasındaki ilişkiyi açıklamak için matematiksel bir model kullanılır. Kullanılan bu model regresyon modeli olarak adlandırılır.

Regresyon modelleri, çoğu zaman bir değişkenin ya da değişken kümesinin bağımlı değişken üzerindeki rasgele etkilerini keşfetmeyi amaçlar. Regresyon modelleri bağımlı değişken değerlerini tahmin etmek için de kullanılabilir. Son olarak regresyon modelleri, bir bağımlı değişken ile açıklayıcı değişkenler arasında tanımlayıcı bir bağ kurmayı amaçlar (Powers, Xie, 1999).

Araştırmalarda oldukça büyük bir ham veriyle ilgili, büyük bir değişim olmaksızın, önemli bilgiyi temsil eden bir özetleme yapılmak istenebilir. Özetleme ya da veri küçültme; örnekler, frekans tablolarını, grup ortalama ve varyanslarını içerir. Birçok istatistik tekniğinin olduğu gibi regresyon da bir veri küçültme tekniğidir. Regresyon analizinde hedef; bağımsız değişkenlerin bir fonksiyonu olan bağımlı değişkeni, gözlenen değerler dizisine olabildiğince yakın şekilde tahmin

etmektir. Regresyon modelinden elde edilen tahmini değerler ile gözlenen değerler tamamen aynı olmayabilir (Powers, Xie, 1999).

Karakteristik olarak regresyon bir gözlemi iki parçaya böler;

Gözlem = Yapısal + Olasılıksal

Gözlem kısmı bağımlı değişkenin eldeki gerçek değeridir. Yapısal kısım bağımlı ve açıklayıcı değişkenler arasındaki ilişkiyi temsil eder. Olasılıksal kısım ise yapısal kısım ile açıklanamayan rasgele bileşendir. (Powers, Xie, 1999)

Doğrusal regresyon uygulamalarının çoğunda amaç; bir bağımlı değişken Y_i ile açıklayıcı değişkenler vektörü x_i arasındaki doğrusal ilişkiyi araştırmaktır. Koşullu beklenen değer fonksiyonu ise doğrusal regresyon analizinin temel bileşenidir. Temel düşünce Y_i bağımlı değişkeni modellemek ve x_i koşulunda beklenen değerini tahmin etmektir. Eşitlik (3.23) verilen x_i için Y_i'nin koşullu beklenen değer fonksiyonunu temsil eder (Winkelmann, Boes, 2006).

$$E(Y_i \mid x_i) = \mu(x_i; \beta) \tag{3.23}$$

Koşullu beklenen değer veriden tahmin edilen bilinmeyen β parametre değerlerine bağlıdır. Doğrusal regresyon için özel olarak (3.24) eşitliği geçerli olur.

$$\mu(x_i; \beta) = x_i^{'}\beta \tag{3.24}$$

Doğrusal regresyon yapısından farklı olarak olasılık modellerinde koşullu beklenen değer fonksiyonu yerine koşullu olasılık fonksiyonunun modeli kurulur. Koşullu olasılık fonksiyonunun tercih edilmesinin nedenlerinden biri, nitel bir değişkenin beklenen değerinin tanımlanmasının çoğunlukla kolay olmamasıdır (özellikle sıralı ve çok düzeyli bağımlı değişkenler için). Başka bir neden, bağımlı değişken bir tercih durumunu gösteriyorsa olasılık temelli yaklaşımın daha fazla bilgi içermesi ve olasılıklar biliniyorken beklenen değerin tanımlanabilmesidir. Koşullu olasılık fonksiyonu aşağıdaki gibi tanımlanır (Winkelmann, Boes, 2006).

$$P(Y_i \mid x_i) = f(Y_i \mid x_i; \theta) \tag{3.25}$$

Eğer bağımlı değişken iki düzeyliyse, yani 0 ve 1 gibi iki değer alıyorsa, koşullu olasılık fonksiyonu $P(Y_i = 1 \mid x_i)$ şeklinde tanımlanır. Bu durumda koşullu beklenen

değer fonksiyonu ile koşullu olasılık fonksiyonu birbirine eşit olur ve aşağıdaki gibi gösterilir.

$$E(Y_i \mid x_i) = 0 \times P(Y_i = 0 \mid x_i) + 1 \times P(Y_i = 1 \mid x_i) = P(Y_i = 1 \mid x_i) \qquad (3.26)$$

Doğrusal regresyonda ve koşullu beklenen değer fonksiyonu uygulamalarında, en küçük kareler (EKK) yöntemi ile artıkların kareleri toplamı minimize edilmektedir. Artıklar, gözlenen değerler ile tahmin edilen değerler arasındaki farkı ifade etmektedir ve $\hat{u}_i = Y_i - \hat{Y}_i$ şeklinde gösterilir. Burada $\hat{Y}_i$, $E(Y_i \mid x_i)$'nin bir tahmin edicisidir. En küçük kareler yönteminde $\sum \hat{u}_i^2$ değerini minimum yapan $\hat{\beta}$ katsayıları elde edilir. Eğer koşullu beklenen değer fonksiyonu doğrusal ise $E(Y_i \mid x_i) = x_i^{'}\beta$ $\hat{Y}_i = x_i^{'}\hat{\beta}$ ve $\hat{u}_i = Y_i - x_i^{'}\hat{\beta}$ şeklinde yazılabilir (Winkelmann, Boes, 2006).

Koşullu olasılık fonksiyonunun kullanıldığı uygulamalarda, koşullu beklenen değer genellikle tanımsızdır. Bu yüzden, regresyon hatası yoktur ve artık temelli yöntemler geçersizdir. Böyle durumlarda EKK yöntemi yerine yaygın olarak kullanılan bir yöntem olan En Çok Olabilirlik (EÇO) yöntemi kullanılabilir. Bu yöntemde, eldeki probleme ait bir olabilirlik fonksiyonu söz konusudur ve bu fonksiyon bilinmeyen parametrelere bağlı olarak maksimize edilir (Winkelmann, Boes, 2006).

EKK yönteminde bağımlı değişkenin gerçek değerleri ile tahmin değerleri arasındaki farkların karelerinin toplamını minimum yapan parametre değerleri elde edilmektedir. EÇO yönteminde ise ana kitle ve bu ana kitleden çekilen örnek arasındaki benzerlik ilişkisinden yararlanılarak bu örneğin elde edilme olasılığını maksimum yapan parametre değerleri tahmin edilmektedir (Sönmez, 2006; Akın, 2002).

3.5. En Çok Olabilirlik Yöntemi

Eğer rasgele bir örnekte, bir dağılımdan bağımsız olarak $x_1, x_2, ..., x_n$ değerleri elde ediliyorsa, bilinmeyen bir θ parametresine bağlı olarak bir $f(x_1, x_2, ..., x_n \mid \theta)$ fonksiyondan bahsedilir ve θ'nın verilen bir değeri için elde edilen verilerin olasılıkları ifade edilebilir. Bu ifade, örneğin olabilirliği (ya da verideki olabilirlik)

olarak adlandırılır. Örneğin olabilirliği, örnek verisinin gerçekte verilen θ'dan elde edilmiş olma olasılığı olarak düşünülebilir. θ bilinmediğinden veriden tahmin edilmesi gerekir. θ'nın tahmini $\hat{\theta}$, örneğin olabilirliğini maksimum yapan değer olarak seçilir. Bilinmeyen parametrelerin değerlerinin tahminini bulma sürecine En Çok Olabilirlik yöntemi denir (Powers ve Xie, 1999).

i gözlemi için koşullu olasılık fonksiyonu $f(y_i \mid x_i;\theta)$ ile gösterilirse, n büyüklüğündeki rasgele örneklem için, gözlenen örneğin bileşik olasılık fonksiyonu tüm gözlemlerin olasılıkları üzerinden aşağıdaki şekilde türetilebilir.

$$f(y \mid x;\theta) = \prod_{i=1}^{n} f(y_i \mid x_i;\theta) \tag{3.27}$$

Bu fonksiyon bilinmeyen θ parametre vektörünün bir fonksiyonu ve bir olabilirlik fonksiyonudur. Daha genel bir ifadeyle yukardaki eşitliğe orantılı olarak, olabilirlik fonksiyonu aşağıdaki gibi tanımlanabilir:

$$L(\theta; y, x) = c\prod_{i=1}^{n} f(y_i \mid x_i;\theta) \tag{3.28}$$

Burada $c > 0$ oransal sabittir. Varsayımların sağlanması halinde, EÇO tahmin edicisi $\hat{\theta}$ vardır, tektir, tutarlıdır ve asimptotik olarak etkindir. (Winkelmann, Boes 2006).

Olasılık modellerinde bağımlı değişken ile açıklayıcı değişkenler arasındaki ilişkinin doğrusal olmamasından ve EKK varsayımlarının ihlal edilmesinden dolayı, model parametrelerinin EKK yöntemi ile tahmin edilmesi hatalı bir yaklaşımdır. Bu nedenle bu modellerde parametre tahminleri elde etmek için EÇO yöntemi tercih edilir (Uçar, 2004).

$P_i = P(Y_i = 1 \mid x_i)$ ve $P(Y_i = 0 \mid x_i) = 1 - P_i$ şeklinde tanımlanmış olsun. Buradaki P_i veya $P(Y_i \mid x_i)$, k adet β katsayılarının değerlerine bağlıdır. Amaç β'ları tahmin etmek olduğundan bu bağlılık olabilirlik fonksiyonuyla daha açıkça $L(Y \mid x, \beta) \equiv P(Y \mid x)$ olarak tanımlanır. EÇO fonksiyonu; k adet β katsayısı tahminleri $\hat{\beta}$'lar için, gözlenen Y değerlerinin elde edilme olabilirliğini mümkün olduğu kadar en büyük yapan değerleri seçmeyi amaçlar. β'nın her "deneme değeri" bir $L(Y \mid x, \beta)$ değeri oluşturacaktır. EÇO fonksiyonu ise bu değerler içerisinde

maksimumu verecek $\hat{\beta}$'yı seçer. Yani, $L(Y \mid x, \hat{\beta}) = \max_{\beta} L(Y \mid x, \beta)$ şeklindedir (Aldrich ve Nelson, 1984).

Gözlenen iki düzeyli Y_i değerlerinin 0 ya da 1 olma olasılığı $P(Y_i \mid x_i) = P_i^{Y_i}(1-P_i)^{1-Y_i}$ olarak verilir. Y değerlerinin gözlemlendiği N büyüklüğündeki örneklemin verilen tüm x_i değerleri koşulundaki olasılığı, N adet olasılığın çarpımına eşittir. Çünkü gözlemler bağımsızdır. x tüm x_i değerlerini göstermek üzere bu ifade eşitlik (3.29)'de verilmiştir.

$$P(Y \mid x) == \prod_{i=1}^{N} P_i^{Y_i}(1-P_i)^{1-Y_i} \tag{3.29}$$

İki düzeyli cevap değişkeni için olabilirlik ifadesi genel olarak aşağıdaki gibi verilir:

$$L = \prod_i F(x_i\beta)^{Y_i}\left[1-F(x_i\beta)\right]^{1-Y_i} \tag{3.30}$$

Logaritma monotonik bir fonksiyon olduğundan L yerine $\log L$ ile çalışmak arasında bir fark olmaz. Burada kullanılan *log* ifadesi e tabanındaki doğal logaritmayı ifade etmektedir. Bu sayede çarpımlarla işlem yapmak yerine toplamları kullanmak işlem kolaylığı sağlayacaktır. Olabilirlik fonksiyonunun logaritması eşitlik (3.31)'de verilmektedir.

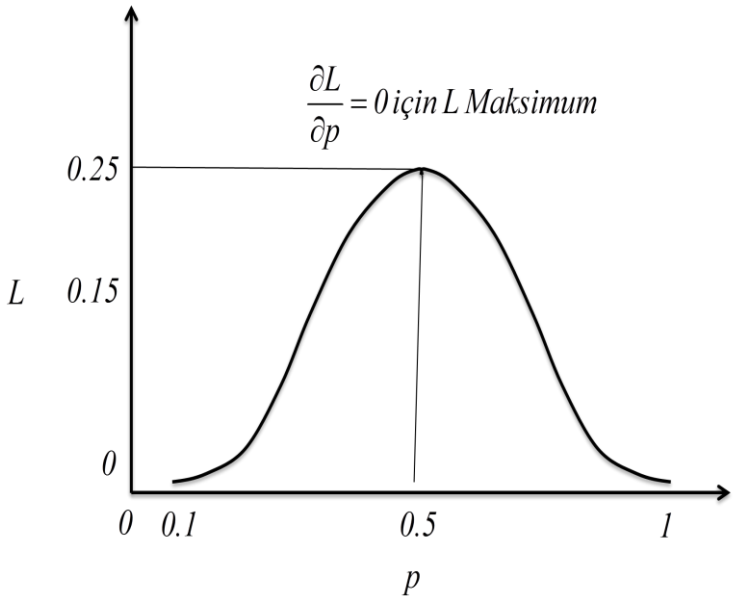
$$\log L = \sum_i \left\{ Y_i \log F(x_i'\beta) + (1-Y_i)\log\left[1-F(x_i'\beta)\right] \right\} \tag{3.31}$$

Aşağıda EÇO yöntemine ilişkin bazı örneklere yer verilmiştir.

Örnek 1: (Binomial Oran)

Hileli bir para n kez atılıyor olsun. x rasgele değişkeninin beklenen değeri, bir denemede paranın yazı gelme olasılığı p olmak üzere $E(x) = np$'dir. Örneğin olabilirliği, p tek bir denemedeki başarı olasılığı, L n denemede x adet başarının olasılığı olmak üzere, binomial bir rasgele değişkenin olasılık fonksiyonudur ve eşitlik (3.32)' deki gibi ifade edilir.

$$L(p) = L = \binom{n}{x} p^x (1-p)^{n-x} \tag{3.32}$$



Şekil 3.4 En Çok Olabilirlik fonksiyonu

Şekil 3.4'te olabilirlik fonksiyonun maksimumu $\hat{p}$ değeri, L'yi maksimum yapan p değeridir. Öncelikle, L'nin değerinin maksimum olduğu noktayı bulmak için, onun değişim oranının, p'ye bağlı olarak sıfır olduğu noktayı bulmak gerekir. Geometrik olarak değişim oranının sıfır olması, L eğrisinin eğiminin sıfır olması anlamına gelir. Bu değerin bulunması için fonksiyonun birinci türevi gereklidir. Yani fonksiyonun eğimi ya da diğer bir ifade ile değişim oranı bulunmalıdır.

L'nin maksimuma ulaştığı değeri bulmak için L'nin p'ye göre birinci türevini bulunur. Ardından p'nin tahmini için bu ifade sıfıra eşitlenir ve $\hat{p}$ değeri bulunur. Dikkat edilmesi gereken bir nokta L'nin maksimum mu yoksa minimum değerinin mi elde edildiğidir. Çünkü, L'nin minimum değerinde de eğim sıfır olacaktır. Bu nedenle, L'nin birinci türevi alınıp sıfıra eşitlendiğinde maksimum değerinin elde edilmiş olması için L'nin ikinci türevinin negatif olması gerekir. Bunun anlamı; L'nin eğiminin $\hat{p}$ komşuluğunda azalıyor olduğudur.

Eşitlik (3.32)'de $\binom{n}{x} = c$ olarak alınırsa, eşitlik (3.33)'deki ifade elde edilir. Bu ifadede c, p'ye bağlı olmayan bir değerdir.

$$L(p) = L = \binom{n}{x} p^x (1-p)^{n-x} = cp^x (1-p)^{n-x} \qquad (3.33)$$

Eşitlik (3.33)'ün logaritması alınırsa eşitlik (3.34) elde edilir.

$$\log L(p) = \log c + x \log p + (n-x)\log(1-p) \tag{3.34}$$

Logaritma monotonik bir dönüşüm olduğu için yerine $\log L$ ile çalışmak arasında bir fark yoktur. Bu nedenle işlem kolaylığı için logaritma ile çalışılır. Burada $0 < L \leq 1$ aralığındayken $-\infty \leq \log L \leq 0$ aralığındadır. $\log L$'nin birinci türevi alındığında, $log\, c$ p'ye bağlı olmayan bir değer olduğundan türevi sıfıra eşit olur ve eşitlik (3.35) elde edilir.

$$\frac{\partial \log L(p)}{\partial p} = \frac{x}{p} - \frac{n-x}{1-p} \tag{3.35}$$

Eşitliğin sağ tarafı sıfıra eşitlenerek p için EÇO tahmin edicisi $\hat{p} = x/n$ bulunur.

Örnek 2: (Normal Ortalama ve Varyans)

Ortalaması μ ve varyansı σ^2 olan normal dağılımlı kitleden n adet değişken gözlemlendiğinde, tek bir gözleme ilişkin olabilirlik aşağıdaki gibi tanımlanır.

$$L(\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2} \right) \tag{3.36}$$

Gözlemler bağımsız olduğundan, örneğin olabilirliği çarpım şeklinde yazılır.

$$L(\mu,\sigma^2) = \prod_{i=1}^{n} L(\mu,\sigma^2) = \frac{c}{\sigma^n} \exp\left\{ -\sum_{i=1}^{n} \left[(x_i-\mu)^2 / 2\sigma^2 \right] \right\} \tag{3.37}$$

Burada c, μ ya da σ'ya bağlı olmayan bir bileşendir. $\hat{\mu}$'yı elde etmek için $\log L$'deki diğer parametre σ^2 sabitken μ'ye göre değişim oranı bulunur. Bu değişim de birinci kısmi türeve eşit olacaktır. Sonrasında bulunan ifade sıfıra eşitlenerek $\hat{\mu} = \sum_{i=1}^{n} x_i / n$ elde edilir. μ'nün EÇO tahmini, örneğin ortalamasıdır. σ^2'nin tahmini için ise μ sabitken σ^2'ye göre kısmi türev alınır ve sıfıra eşitlenir. Elde edilen $\hat{\sigma}^2 = \sum_{i=1}^{n} (x_i-\mu)^2 / n$ değeridir ve bu değer örneğin varyansının $(n-1)/n$ ile çarpımına eşittir (Powers ve Xie, 1999).

3.5.1. En Çok Olabilirlik tahmin edicisinin genel ilkeleri

1- Eğer $\log L_i$ aynı dağılımdan gelen n bağımsız gözlemden birinin olabilirliği ise örneğin olabilirliği $\log L = \sum_{i=1}^{n} \log L_i$ 'dir.

2- $\theta = (\theta_1, \theta_2, ..., \theta_n)'$ $K \times 1$ boyutlu verinin dağılımı hakkında bilgi veren (ortalama, varyans, model etkileri, regresyon eğimleri, vs.) parametreler vektörünü ifade etmek üzere θ'nın EÇO tahmin edicisi aşağıdaki adımlar izlenerek elde edilir.

a) θ'ya göre $\log L$'nin kısmi türevlerinin vektörü eşitlik (3.38) kullanılarak bulunur.

$$U(\theta) = \frac{\partial \log L(\theta)}{\partial \theta} \tag{3.38}$$

U_k, θ_k'ya göre her bir kısmi türevi ifade etmek üzere, eşitlik (3.39) kullanılarak elde edilen K tane eşitlik 0'a eşitlenerek θ_k'lar elde edilir. Böylece K bilinmeyenli K adet eşitlik elde edilmiş olur $U(\theta)$, $\log L$'nin birinci türev değerlerinden oluşan $K \times 1$'lik bir vektördür ve skor vektör ya da S*kor Fonksiyon* olarak adlandırılır U_k'lar skor vektörünün elemanlarıdır.

$$U_k = \frac{\partial \log L(\theta)}{\partial \theta_k}, (k = 0, ..., K) \tag{3.39}$$

b) Sonraki adımda, elde edilen birinci kısmi türevlerin maksimumu tanımladığından emin olunmalıdır. Bu nedenle, θ'ya göre $\log L$'nin ikinci türevinin negatif olup olmadığı kontrol edilir. Sözü geçen negatif tanımlı matris eşitlik (3.40)'taki gibi elde edilir.

$$H(\theta) = \frac{\partial^2 \log L(\theta)}{\partial \theta \partial \theta'} \tag{3.40}$$

İkinci türevler matrisi $H(\theta)$ *Hessian Matris* olarak adlandırılır ve negatif tanımlı olmalıdır. Bu matrisin elemanları h_{kl}'lerden oluşur.

$$h_{kl} = \frac{\partial^2 \log L(\theta)}{\partial\theta_k \partial\theta_l}, (k = 0,...,K), (l = 0,...,K) \quad (3.41)$$

h_{kl} elemanlarından oluşan bu matrisin negatifi, *Bilgi Matrisi* olarak adlandırılır ve $I(\theta)$ şeklinde gösterilir. Bilgi matrisi $-H(\theta)$ şeklinde elde edilebileceği gibi farklı yaklaşımı da mevcuttur. Örneğin; eğer $\log L_i$ örnek log olabilirliğine i gözleminin katkısını gösteriyorsa birinci türevlerin nxK boyutlu matrisi (ya da birey skor fonksiyonları) gradyan vektör $g(\theta)$ olarak adlandırılır. Tahmini bilgi matrisi eşitlik (3.42) ile tanımlanır.

$$\hat{I}(\theta) = g(\theta)'g(\theta) \quad (3.42)$$

c) Bilgi matrisinin tersinin köşegen elemanlarının karekökü EÇO tahmincisinin standart hatalarını verir. $I^{-1}(\theta)$ $\hat{\theta}$'da değer aldığında, $\hat{\theta}$'nın asimptotik varyans kovaryans matrisini verir. $I^{-1}(\theta)$'in köşegen elemanları $\text{var}(\hat{\theta})$'dır. EÇO tahminleri asimptotik olarak normal dağıldığından bu bilgi θ etrafında güven aralıkları oluşturmakta kullanılabilir.

Örnek 3: (İki Düzeyli Lojit Model)

Gruplanmış ya da birey düzeyli binominal cevap değişkenli model için m_i denemedeki (örn; gözlenen hücre frekansları) başarı sayısını temsil etmek üzere, log olabilirlik eşitlik (3.43) şeklinde yazılabilir.

$$\log L(\beta) = \sum_i \left\{ \log \binom{m_i}{Y_i} + Y_i \log F(x_i'\beta) + (m_i - Y_i) \log\left[1 - F(x_i'\beta)\right] \right\} \quad (3.43)$$

Burada $\binom{m_i}{Y_i}$ çarpanı bilinmeyen parametreleri içermediğinden tahmin etme sürecinde ihmal edilebilir. Birey düzeyli veri durumunda, bu terim her zaman 1'e eşit olacağından iki düzeyli Lojit Model aşağıdaki şekilde sadeleştirilir ve eşitlik (3.45) elde edilir.

$$\log L = \sum_i \left\{ Y_i \log F(x_i'\beta) + (1 - Y_i) \log\left[1 - F(x_i'\beta)\right] \right\} \quad (3.45)$$

Λ lojistik dağılımın olasılık yoğunluk fonksiyonu Λ ile temsil edilmek üzere eşitlik (3.46)'daki gibi ifade edilir.

$$\Lambda_i = F(x_i'\beta) = \exp(x_i'\beta)/\{1+\exp(x_i'\beta)\} \tag{3.46}$$

Eşitlik (3.46) eşitlik (3.45)'te yerine yazılırsa aşağıdaki eşitlik elde edilir.

$$\log L(\beta) = \sum_i \{Y_i \log \Lambda_i + (1-Y_i)\log[1-\Lambda_i]\} \tag{3.47}$$

İki düzeyli Lojit Model için skor fonksiyonu eşitlikleri aşağıda verilmiştir.

$$u_k = \frac{\partial \log L(\beta)}{\partial \beta_k} = 0 \text{ , } (k = 0,...,K) \tag{3.48}$$

$$u_k = \sum_i (Y_i - \Lambda_i) x_{ik} \text{ , } (k = 0,...,K) \tag{3.49}$$

Matris gösterimi ile bu ifade $U = x'(Y-\Lambda)$ şeklinde yazılır. Lojit Model için ikinci türev matrisinin k . satır l. kolon elemanı h_{kl} eşitlik (3.50)'deki gibi yazılır.

$$h_{kl} = \frac{\partial^2 \log L(\beta)}{\partial \beta_k \beta_l} < 0 \text{ , } (k = 0,...,K), (l = 0,...,K) \tag{3.50}$$

Eşitlik (3.49)'un β_l 'ye göre türevi alınırsa eşitlik (3.50)'deki h_{kl} eşitlik (3.51)'deki gibi yazılabilir

$$h_{kl} = -\sum_i x_{ik} x_{il} \Lambda_i (1-\Lambda_i) \tag{3.51}$$

ya da matris gösterimiyle $H = -x'wx$ şeklindedir. Burada w matrisi, köşegen elemanları $\Lambda_i(1-\Lambda_i)$ olan köşegen bir matristir.

Lojit ve Probit tahminlerinin asimptotik varyans kovaryans matrisi Hessian (ya da beklenen Hessian) matrisinin negatifinin tersi ya da diğer bir ifadeyle bilgi matrisinin tersi ile tahmin edilir. Lojit Model için bilgi matrisi aşağıdaki gibi verilir:

$$-\left[\frac{\partial^2 \log L(\beta)}{\partial \beta \beta'}\right] = I(\beta) = \sum_i \Lambda_i (1-\Lambda_i) x_i x_i' \tag{3.52}$$

Λ_i bilinmeyen parametrelerin doğrusal olmayan bir fonksiyonu olduğundan EÇO tahminleri için iteratif yöntemler kullanılır. Olabilirlik eşitlikleri için parametre

tahminleri yerine konulur. Bu nedenle Λ_i yerine $\hat{\Lambda}_i$ kullanılır. İki düzeyli değişken üzerinden EKK regresyondan elde edilen tahmini değer başlangıç değeri olarak alındıktan sonra, yapılan en son iterasyon ile bir önceki iterasyondan elde edilen tahminler arasındaki fark önemsiz oluncaya kadar iterasyon işlemi tekrarlanır.

İteratif süreçte kullanılabilecek Newton-Raphson ya da Fisher skor iteratif yöntemleri benzer sonuçlar verir. Çünkü, beklenen ve gerçek Hessianlar aynıdır. Sonuç matrisi negatif tanımlıdır ve sonuç tahminleri global bir maksimum sağlarlar. Bu iki yöntem dışında kullanılabilecek başka yöntemler de mevcuttur.

Newton-Raphson ve skor metodu bir başlangıç değeri ile başlar. Bir sonraki iterasyon, bir önceki iterasyon değerleriyle elde edilir. $(t+1)$. iterasyon için tahminler eşitlik (3.53)'deki gibi yazılabilir:

$$\hat{\beta}^{(t+1)} = \hat{\beta}(t) - \left[H(t) \right]^{-1} U(\hat{\beta}(t)) \qquad (3.53)$$

Burada $\hat{\beta}$, $(K+1)\times 1$ boyutlu parametreler vektörü, H ve U sırasıyla $\log L$'nin β 'ya göre ikinci ve birinci kısmi türevlerini ifade etmektedir. Ayrıca bilgi matrisi kullanılarak bu eşitlik aşağıdaki gibi ifade edilebilir:

$$\hat{\beta}^{(t+1)} = \hat{\beta}(t) + \left[I(\hat{\beta}(t)) \right]^{-1} U(\hat{\beta}(t)) \qquad (3.54)$$

$\Delta\hat{\beta} = \left| \hat{\beta}(t) - \hat{\beta}^{(t-1)} \right|$ farkı yeterince küçük olduğu zaman tahmin edilen parametre vektörünün gerçek parametre vektörüne yakınsadığı söylenir ve iteratif süreç sonlandırılır. Bunun anlamı; log olabilirlik değerindeki oransal değişimin δ gibi çok küçük bir değerden daha küçük olması ya da skor vektörünün sıfıra yeterince yakın olmasıdır. Tahminlerin asimptotik varyans kovaryans matrisi, $I(\hat{\beta})^{(-1)}$ bilgi matrisinin (negatif Hessian) tersi olarak son iterasyondan elde edilir. EÇO tahminleri asimptotik olarak normal dağılır ve tahmini varyanslar bilgi matrisinin tersinin köşegen elemanlarına eşittir ($DiagI(\hat{\beta})^{-1}$). $\hat{\beta} / Diag\sqrt{I(\hat{\beta})^{-1}}$ ifadesi asimptotik standart normal dağılım (ya da z dağılımı) gösterir. Bu ifade asimptotik olarak bir z istatistiği olmasına rağmen, genellikle bir t istatistiği olarak

bilinir. Parametrelere ilişkin dağılım bilindiğinden parametrelerin anlamlılık testlerinin yapılması mümkündür (Powers ve Xie, 1999).

Probit ve Lojit Modeller için olabilirlik eşitlikleri aşağıdaki gibi özetlenebilirler.

Probit Model için olabilirlik eşitlikleri:

$$\sum \frac{\left[Y_i - \Phi(\sum \hat{\beta}_k x_{ik}) \varphi(\hat{\beta}_k x_{ik}) \right]}{\Phi(\sum \hat{\beta}_k x_{ik})[1-\Phi(\sum \hat{\beta}_k x_{ik})]} x_{ij} = 0 \quad j = 1,...,K \tag{3.61}$$

Burada $\varphi(.)$ standart normal dağılım için olasılık yoğunluk fonksiyonudur (yoğunluk fonksiyonu kdf'nin türevidir).

Lojit Model için eşitlikler:

$$\sum_{i=1}^{N} \left[Y_i - \frac{\exp(\sum_k \hat{\beta}_k x_{ik})}{1+\exp(\sum_k \hat{\beta}_k x_{ik})} \right] x_{ik} = 0 \quad k = 1,...,K \tag{3.62}$$

Olabilirlik eşitlikleri genel olarak aşağıdaki şekilde ifade edilebilir:

$$\sum_{i=1}^{N} \left[Y_i - P(Y_i = 1 | x_i, \beta) \right] A_i x_{ik} = 0 \text{ , } (k = 1,...,K) \tag{3.63}$$

Eşitlikteki A_i değeri Lojit için 1'e, Probit için $\varphi / \Phi(1-\Phi)$ eşit olacaktır. Eşitlik (3.63)'te parantez içindeki değer, Y_i gerçek değeriyle onun beklenen ya da tahmin edilen değeri arasındaki sapmayı ifade eder. Bu durumda EÇO, AEKK ve EKK yöntemlerinin hepsinin amacı; gerçek değerlerle beklenen değerler arasındaki ağırlıklandırılmış sapmalar toplamını 0 yapan β tahminlerini elde etmektir. Aralarındaki farklılık ise beklenen değer tanımlarından (Doğrusal EKK'da $\sum_k \beta_k x_{ik}$, doğrusal olmayan olasılık modellerinde $P(Y_i = 1 | x_i, \beta)$) ve ağırlık tanımlarından (Doğrusal EKK'da, x_{ik}, AEKK'da $w_i^2 x_{ik}$ ve doğrusal olmayan olasılık modellerinde $A_i x_{ik}$) kaynaklanmaktadır (Aldrich ve Nelson,1984; Maddala 1983).

3.5.2. En Çok Olabilirlik sonuçlarının anlamı ve yorumu

Bütün yöntemlerde alışılmış olarak parametre tahminleri $\hat{\beta}_k$ ve ilgili standart hataları S_k olarak gösterilir. S_{jk}, β_j ve β_k arasındaki kovaryans tahminini temsil eder. Varyans kovaryans matrisi simetrik bir matris olduğundan $S_{jk} = S_{kj}$'dir ve dolayısıyla alt ya da üst üçgen matrisi olarak gösterilmektedir. Bu matrisin köşegen elemanlarından S_{kk}, β_k katsayısının varyansıdır ve karekökü ilgili katsayının standart hatası S_k'yı verir. İki katsayı arasındaki korelasyon $r_{jk} = S_{jk} / S_j S_k$ formülüyle hesaplanır. Korelasyonları yorumlamak kovaryansları yorumlamaktan daha kolay olacağından analizlerde korelasyonlar tercih edilir.

Probit ve Lojit analizlerde elde edilen katsayı tahminleri asimptotik olarak yansız ve etkindirler. Tahmin edilen standart hatalar, tahmini katsayıların olası değişkenliklerinin genel ölçümünü sağlar. β_k katsayılarının sıfıra eşitliğinin testi için bilinen regresyon analizinde *t* istatistiği kullanılır. Doğrusal regresyondaki *t* istatistiği lojistik regresyonda Wald istatistiği olarak da adlandırılır. Bu istatistik eşitlik (3.55)'deki gibi tanımlanır:

$$\hat{t}_k = \hat{\beta}_k / S_k \qquad (3.55)$$

(Daha genel olarak $t_k = (\beta_k - \beta_k^*) / S_k$ dır. Burada β_k^*, β_k'nın herhangi bir değere eşitliğinin testinde kullanılır. Yani sıfır olması şart değildir.) t_k değerine bakılarak katsayıların sıfıra eşit olduğu hipotezi reddedilir veya edilemez. t_k değeri *N-K* serbestlik dereceli ve α anlamlılık düzeyli *t* tablo değeriyle karşılaştırılır (eğer veri tekrarlı ise *N* yerine toplam N_i olur). Katsayı tahminleri asimptotik olarak normal dağılmasına rağmen, *student's t* dağılımı yaygın olarak kullanılır. Hipotez testlerinde *t* istatistiğini kullanmak, z testini kullanmaktan daha mantıklı (ölçülü, sade vs.) olur. *t* dağılımı büyük serbestlik derecelerinde normal dağılıma yakınsar ya da daha doğrusu *t* dağılımı katsayı tahminlerinin dağılımı gibi asimptotik olarak normaldir. Parametre tahminlerine ilişkin güven aralığı da aşağıdaki gibi ifade edilir.

$$\hat{\beta}_k \pm t_{(N-k;\alpha)} S_k \qquad (3.56)$$

3.5.3. Gruplanmış veriler için En Çok Olabilirlik yöntemi

Gruplanmış verileri kullanarak model parametrelerinin tahminin elde edilmesi için genel eşitlik (3.43) kullanılır. (3.43) eşitliğinin logaritması alınırsa binomial modeller için eşitlik (3.57) elde edilir.

$$\log L(\beta) = \sum_i \{Y_i \log F_i + (m_i - Y_i)\log(1 - F_i)\} \tag{3.57}$$

Y_i, m_i denemedeki (örn; gözlenen hücre frekansları) başarı sayısını temsil eder. β 'ya göre $\log L$ 'nin türevi eşitlik (3.58)'te yer almaktadır.

$$U_k = \frac{\partial \log L(\beta)}{\partial \beta_k} = \sum_i Y_i x_{ik} - \sum_i m_i x_{ik} F_i \text{ , } (k = 1,...,K) \tag{3.58}$$

Eşitlik (3.58), eşitlik (3.59)'in matrisler yardımıyla gösterimidir.

$$U = x'(Y - r) \tag{3.59}$$

Burada Y başarı sıklıklarının kolon vektörü, x bağımsız değişkenler matrisi, r $r_i = m_i F_i$ elemanlarından oluşan kolon vektörüdür.

İkinci türev matrisinin $k.$ satır $l.$ kolon elemanı aşağıdaki gibidir.

$$h_{kl} = \frac{\partial^2 \log L(\beta)}{\partial \beta_k \beta_l} = -\sum_i x_{ik} x_{il} m_i F_i (1 - F_i) \text{ , } (k = 1,...,K) \tag{3.60}$$

$H = -x'wx$ şeklinde de ifade edilebilir. Burada w köşegen matristir ve elemanları $m_i F_i (1 - F_i)$ 'dir (Powers ve Xie, 1999).

4. İKİ DÜZEYLİ BAĞIMLI DEĞİŞKENLERDE EN ÇOK KULLANILAN MODELLER

Bağımlı değişken iki düzeyli olduğu durumda en yaygın olarak kullanılan modeller: Doğrusal Olasılık, Lojit ve Probit Modellerdir. Doğrusal Olasılık Modeli kolay uygulanabilmesi açısından tercih edilmesine rağmen, gerçekleşmesi istenen varsayımların sağlanamamasından ötürü birçok istenmeyen durumla karşılaşılır. Bu durumlar aşağıda ayrıntılı olarak ele alınmıştır. Ardından, Doğrusal Olasılık Modeli nde karşılaşılan problemlere çözüm olarak geliştirilen Lojit ve Probit Modeller anlatılmıştır. Bu iki model farklı olmalarına rağmen oldukça yakın tahminler elde ederler. İki modelin tahminlerinin arasındaki farkın, farklı koşullarda değerlendirilebilmesi adına sonraki bölümde bir simülasyon çalışması yapılmıştır. İki modeli karşılaştırmada kullanılan ölçütlerin çoğunun Doğrusal Olasılık Modeli için hesaplanamaması ya da anlamsız olmasından ötürü karşılaştırma sadece Probit ve Lojit Modeller arasında gerçekleştirilmiştir.

4.1. Doğrusal Olasılık Modeli

Doğrusal regresyonda bağımlı değişkenin sürekli olma ve normal dağılıma sahip olma şartı vardır. Ancak bazı durumlarda bağımlı değişken Y_i'nin 0 ve 1 gibi sadece iki değer alması, süreklilik varsayımı ihlaline sebep olur.
Y_i'nin iki değer alması durumunda bağımlı değişkenin beklenen değeri eşitlik (4.1) gibi olacaktır.

$$E(Y_i) = 1 \times P(Y_i = 1) + 0 \times P(Y_i = 0) = P(Y_i = 1) \tag{4.1}$$

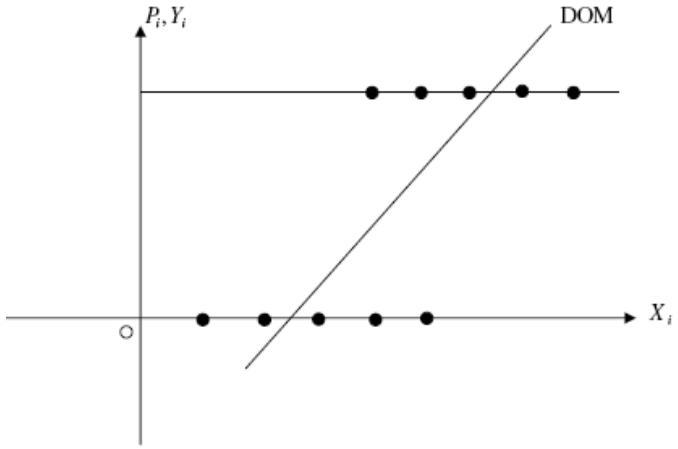
Buradan hareketle, bağımlı değişkenin beklenen değerinin, açıklayıcı değişkenler ile katsayıların çarpımlarının toplamına eşitliği yazılabilir.

$$E(Y_i) = P(Y_i = 1) = \sum_k \beta_k x_{ik} \tag{4.2}$$

Elde edilen bu eşitliğin sol tarafı bir olasılık ifadesi olduğundan ancak 0 ile 1 arasında değerler alabilecektir. Bu nedenle, bağımlı değişkeni 0 ya da 1 değerlerini alan regresyon modeline Doğrusal Olasılık Modeli (DOM) adı verilir. Diğer bir

ifadeyle, bağımlı değişkenin gölge değişken olduğu ve açıklayıcı değişkenin doğrusal bir fonksiyonu olarak gösterildiği regresyon modeli DOM'dur.

DOM, parametre tahmin değerlerine bağlı olarak açıklayıcı değişkenin çeşitli değerleri karşısında ilgilenilen olayın ortaya çıkma olasılığını verirken; modeldeki herhangi bir katsayı, diğer değişkenler sabitken, ilgili olduğu açıklayıcı değişkendeki bir birimlik değişmeyle olayın gerçekleşme olasılığı üzerindeki etkisini ifade etmektedir.



Şekil 4.1 Doğrusal Olasılık Modeli

Eğer Y_i bağımlı değişkeni sadece iki değer alabiliyorsa, herhangi x_k açıklayıcı değişkeninin değerleri için, hata terimi de sadece iki değer alabilecektir. Eğer $Y_i = 0$ ise $0 = \sum_k \beta_k x_{ik} + u_i$ ya da $u_i = -\sum_k \beta_k x_{ik}$, eğer $Y_i = 1$ ise $1 = \sum_k \beta_k x_{ik} + u_i$ ya da $u_i = 1 - \sum_k \beta_k x_{ik}$ olur. Hata teriminin beklenen değerinin sıfır olduğu varsayımının sağlandığı eşitlik (4.4)'teki gibi gösterilebilir.

$$\begin{aligned} E(u_i) &= P(Y_i = 0)(-\sum \beta_k x_{ik}) + P(Y_i = 1)(1 - \sum \beta_k x_{ik}) \\ &= P(Y_i = 0)(-\sum \beta_k x_{ik}) + P(Y_i = 1) - P(Y_i = 1)(-\sum \beta_k x_{ik}) \\ &= ((-\sum \beta_k x_{ik})(P(Y_i = 0) + P(Y_i = 1)) + P(Y_i = 1) \\ &= (-\sum \beta_k x_{ik}).1 + P(Y_i = 1) \\ &= P(Y_i = 1) - \sum \beta_k x_{ik} \\ &= 0 \end{aligned} \tag{4.3}$$

Hata teriminin beklenen değerinin 0'a eşit olma varsayımı sağlandığından en küçük kareler tahmin edicisi, β_k için yansız bir tahmin edici olacaktır. Ancak bu durum her zaman geçerli ve tek başına yeterli bir özellik değildir. DOM'dan elde edilen sonuçların anlamlı olabilmesi için EKK varsayımlarını sağlaması gerekmektedir. Bu varsayımların ihlal edilmesi bazı sorunlara sebep olmaktadır. Bu sorunlara aşağıda yer verilmiştir.

4.1.1. Doğrusal Olasılık Modeli'nde karşılaşılan sorunlar

DOM'da karşılaşılan sorunlar beş başlık altında aşağıda açıklanmıştır.

1. Sabit olmayan varyans sorunu

Varsayımlardan biri hata teriminin varyansının sabit olmasıdır. Ancak hata terimi u_i'nin varyansı açıklayıcı değişken değerlerine göre değişecektir. Bu yüzden, katsayıların EKK tahminleri etkin ve standart hataları yansız olmayacaktır. Sonuç olarak test istatistikleri doğru sonuçlar vermeyecektir. Aşağıdaki eşitliklerde $P(Y_i = 1) = \sum \beta_k x_{ik} = P_i$ olmak üzere, hatanın beklenen değeri ve varyansının elde edilmesine ilişkin eşitlikler verilmiştir.

$$E(u_i) = P(Y_i = 0)(-\sum \beta_k x_{ik}) + P(Y_i = 1)(1 - \sum \beta_k x_{ik}) = 0 \tag{4.4}$$

$$\begin{aligned} V(u_i) &= E(u_i^2) - \left[E(u_i)\right]^2 \\ V(u_i) &= E(u_i^2) \\ V(u_i) &= (1 - \sum \beta_k x_{ik})^2 (\sum \beta_k x_{ik}) + (\sum \beta_k x_{ik})^2 (1 - \sum \beta_k x_{ik}) \\ &= (\sum \beta_k x_{ik})(1 - \sum \beta_k x_{ik}) \\ V(u_i) &= P_i(1 - P_i) \\ &= [1 - E(Y_i \mid x_i)]E(Y_i \mid x_i) \end{aligned} \tag{4.5}$$

Eşitlik (4.5) 'te görüldüğü gibi, hatanın varyans değeri, P_i olasılık değerine bağlıdır ve olasılıklar da açıklayıcı değişkenlerin değerlerine bağlı olduklarından, hata varyansı her bir gözlem için değişim gösterecektir. Bunun anlamı, hata varyansının sabit bir değer almamasıdır. Hatanın varyansı, P_i'nin 0 ve 1'e yakın olduğu durumlarda küçük, 0.5'e yakın olduğu durumlarda ise büyük değerler alır.

2. Normal dağılmama sorunu

Açıklayıcı değişken sadece 0 ve 1 değerlerinden birini alabileceğinden, hata terimi de gözlenen bağımlı değişken değerleri ile beklenen değer arasındaki farkı ifade etmesi açısından aynı şekilde iki düzeyli olacaktır. Bu durumda hata teriminin sürekli olması beklenemez. Bu nedenle hata terimi normal dağılım göstermez. EKK tahminlerinde yansızlık için normallik varsayımı gerektiğinden, yansız olmama sorunu ortaya çıkar.

Eğer bağımlı değişken P_i olasılığıyla 1 değerini alıyorsa hata terimi eşitlik (4.6) ile ifade edilir.

$$\begin{aligned} u_i &= 1 - E(Y_i) \\ u_i &= 1 - \sum \beta_k x_{ik} \\ u_i &= 1 - P_i \end{aligned} \qquad (4.6)$$

Eğer bağımlı değişken $1-P_i$ olasılığıyla 0 değerini almışsa hata terimi $-P_i$'ye eşit olur. Böylece hata teriminin sadece iki değer aldığı görülmektedir.

$$\begin{aligned} u_i &= 0 - E(Y_i) \\ u_i &= 0 - \sum \beta_k x_{ik} \\ u_i &= -P_i \end{aligned} \qquad (4.7)$$

3. Olasılıkların [0,1] aralığında olma varsayımının yerine gelmeyişi

Yukarıda bahsedilen hata teriminin beklenen değerinin 0'a eşit olma varsayımı, açıklayıcı değişkenlerin sınırlı aralıkları için anlamlı olabilmekte, yansızlığı getirebilmektedir. Ancak açıklayıcı değişkenlerin değer aralıklarının geniş olması, olayın olma olasılığı P_i'nin olması gereken [0,1] aralığı dışında değerler almasına sebep olmaktadır. Bu aralıklar dışındaki bir olasılığı yorumlamak anlamsızdır. Bu sorunu ortadan kaldırmak için P_i'ye bir takım dönüşümler uygulamak gerekir. Elde edilen tahmini değerlerin [0,1] arasında olup olmadığına bakıldıktan sonra, bu sınırları aşan değerler mevcutsa 0'dan küçük olanlara 0, 1'den büyük olan değerlere ise 1 değeri verilir. Bu durumda tahmin süreci yansız tahmin ediciler verebilmesine rağmen, modelin kestirim amacıyla kullanılması halinde elde edilen tahminler yanlı olacaktır. P_i'nin aldığı değerler eşitlik (4.8)'de verilen fonksiyonla ifade edilebilir.

$$P_i = \begin{cases} \beta_0 + \beta_1 x & 0 < \beta_0 + \beta_1 x < 1 \\ 1 & \beta_0 + \beta_1 x \geq 1 \\ 0 & \beta_0 + \beta_1 x \leq 0 \end{cases} \quad (4.8)$$

4. Fonksiyonel yapı sorunu

DOM doğrusal olduğundan, x_k'daki bir birimlik değişim, geri kalan tüm değişkenler sabitken, ilgilenilen olayın gerçekleşme olasılığı üzerine β_k katsayısı kadar sabit bir etkiye sahiptir. Bu ifade eşitlik (4.9)'daki gibi verilebilir.

$$\frac{\partial P_i}{\partial x_i} = \beta_i$$

Bu etki açıklayıcı değişkenin aldığı değerlere ve başlangıç değerine bağlı olmayacaktır. Ancak bu durum çoğu uygulama için gerçekçi ve doğru değildir. Katsayıların bu yorumu, bağımlı değişken ortalamasının açıklayıcı değişkenler ile doğrusal bir yapıda modellendiği doğrusal EKK regresyonda anlamlı iken, bağımlı değişken ortalamasının bir olasılık olarak modellenmesi durumunda anlamlı olmayacaktır.

Örneğin; *x* açıklayıcı değişkeni kişinin gelirini ve Y ise kişinin otomobil satın alma kararını ifade eden iki düzeyli cevap değişkeni olarak tanımlansın.

$$Y_i = \begin{cases} 1 & \text{otomobil satın alacak} \\ 0 & \text{diğer durum} \end{cases}$$

İlgili Doğrusal Olasılık Modeli aşağıdaki gibi ifade edilir.

$$P(Y_i = 1) = P_i = \beta_0 + \beta_1 x_1$$

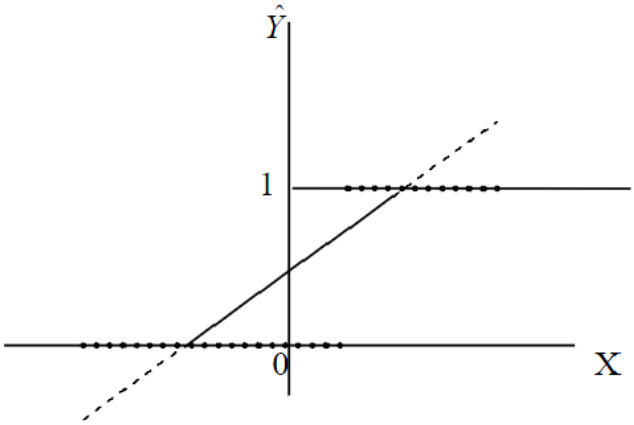
Bu modele ilişkin tahmin edilen β_1 katsayısının pozitif olması durumunda, gelirdeki 1 birimlik artış kişinin otomobil satın alma olasılığında en fazla β_1 kadar bir artışa sebep olacaktır. Bu model gelir ile otomobil sahibi olmak arasında doğrusal bir ilişkinin varlığını kabul etmiş olur. Ancak bu pratikte gerçekçi değildir. Çünkü, kişinin başlangıçtaki servetinin ne düzeyde olduğu dikkate alınmaksızın yapılan yorum yanlıştır. Başlangıçta düşük servet düzeyine sahip bireyin gelirindeki 1 birimlik artış otomobil alması olasılığını çok yükseltmeyeceği gibi, yüksek servet düzeyine sahip bireyin de gelirindeki 1 birimlik artış da otomobil satın alma olasılığını çok arttırmayacaktır. En önemli değişimler gelirin uç noktalarında değil

orta noktalarında meydana gelecektir. Gelir ile otomobil sahibi olma arasındaki ilişki doğrusal değil S biçimli olarak adlandırılan bir eğri ile ifade edilmelidir (Uçar, 2004).

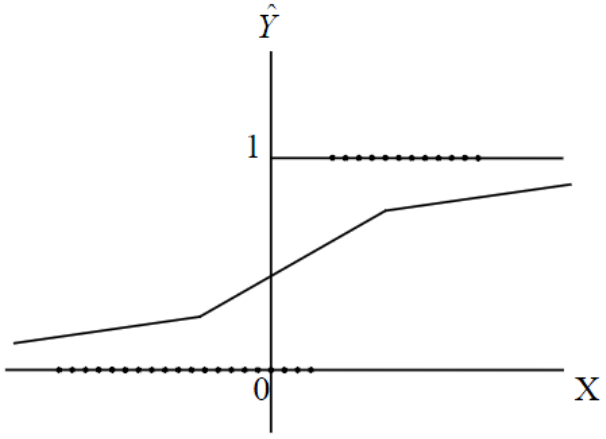
İki düzeyli yanıt değişkenli modeller olayın olasılığı ve açıklayıcı değişkenler arasında S biçimli ilişkiye sahiptir ve bu durum DOM'da fonksiyonel yapıdaki problemi işaret eder.

5. Uyum iyiliği ölçüsü olarak R^2'nin kuşkulu değeri

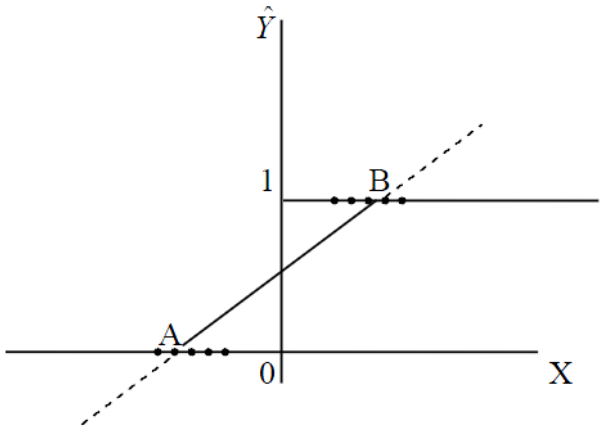
Bilinen belirlilik katsayısı R^2 değeri, iki düzeyli cevap değişkenli modellerde kullanmak için çok uygun bir ölçüt değildir. Bunu görmek için Şekil 4.2, 4.3 ve 4.4 incelenebilir. Belirli bir x'e karşılık gelen Y değişkeni 0 ya da 1 değerini alır. Bu nedenle Y değerleri, ya x ekseni üzerinde ya da 1'in hizasındaki doğru üzerinde yer alır. Dolayısıyla ister sınırlanmamış DOM, ister 0-1 mantıksal sınırları dışına çıkmayacak biçimde tahmin edilen sınırlanmış DOM olsun, genellikle hiçbir DOM bu şekilde değerler almış veriye iyi bir uyum gösteremez. Sonuç olarak hesaplanan R^2 değeri böyle modellerde 1'den çok küçük çıkma eğilimi gösterir. R^2 uygulamaların çoğunda, 0.2 ile 0.6 arasında yer alır. Bu tür modellerde R^2, ancak serpilme noktalarının A ile B'ye çok yakın dağılması durumunda (şekil 4.4) yüksek olacaktır (örneğin 0.8'den büyük bir değer). Çünkü A ile B noktalarını birleştiren doğrunun bu noktaları temsil etmesi kolay olur. O zaman Y_i tahminleri 0'a ya da 1'e yakın olacaktır (Gujarati, 1999).



Şekil 4.2 Sınırlanmamış Doğrusal Olasılık Modeli



Şekil 4.3 Sınırlanmış Doğrusal Olasılık Modeli



Şekil 4.4 A ile B çok yakınken Doğrusal Olasılık Modeli

Bu nedenlerle Aldrich ve Nelson (1984), nitel bağımlı değişkenli modellerde, belirlilik katsayısının bir özetleyici istatistik olarak kullanımından kaçınılması gerektiğini ileri sürer.

4.1.2. Tekrarlanmamış veriler ile Doğrusal Olasılık Modeli için ağırlıklı en küçük kareler tahmin tekniği

Goldberg (1964), DOM'daki etkin parametre tahminlerinin elde edilmesini engelleyen değişen varyans sorununu çözmek için iki-adım ağırlıklandırılmış tahmin ediciyi önermiştir. İlk adımda bilinen en küçük kareler yöntemi uygulanarak regresyon analizi gerçekleştirilir. Böylece yansız $\hat{\beta}_k$ tahminleri elde edilir. Bu

tahminlerden her gözlem için ağırlıklar hesaplanır ve eşitliğin her iki tarafı ile çarpılır. (Aldrich ve Nelson, 1984)

İlk adımda, her bir gözlem için elde edilen ağırlıklar aşağıdaki gibi hesaplanır:

$$W_i = \left[\frac{1}{\left[\left(\sum \hat{\beta}_k x_{ik} \right) \left(1 - \sum \hat{\beta}_k x_{ik} \right) \right]^{1/2}} \right] \tag{4.10}$$

İkinci adımda, Doğrusal Olasılık Modeli nin her iki tarafı W_i ile çarpılır.

$$\left(W_i Y_i \right) = \sum (W_i \hat{\beta}_k x_{ik}) + (W_i u_i) \tag{4.11}$$

Elde edilen $(W_i u_i)$ sabit varyansa sahip olur. Böylece devamında elde edilen $\hat{\beta}_k$ tahmin edicisi yansız olmakla beraber ayrıca, mümkün olan en küçük örneklem varyansına sahip olmuş olur. $\hat{\beta}_k$ .tahminleri hipotez testleri için de doğru sonuçlar verecektir. Son olarak u_i 'nin iki değer alması ve dolayısıyla normal dağılıma sahip olmaması sorunuyla karşı karşıya kalınır. Buna rağmen büyük örneklemlerde $\hat{\beta}_k$ normal dağılıma yakınsama gösterecektir ve hipotez testleri için kullanılabilecektir (Aldrich ve Nelson,1984).

İkinci aşamadan sonra elde edilen tahmini parametreler belirlilik katsayısı R^2 'yi hesaplamak için uygun değildir. R^2 modelin bağımlı değişkende meydana gelen değişimi açıklama oranına dair bilgi vermektedir. R^2 'yi ağırlıklandırılmış verilerle hesaplamak, gerçek bağımlı değişkenin varyansında meydana gelen değişim yerine ağırlıklandırılmış bağımlı değişkenin varyansında meydana gelen değişim hakkında bilgi vermiş olacaktır. Bu nedenle R^2, dönüştürülmemiş verilerden tekrar elde edilmelidir. R^2 'ye ilişkin eşitlik (4.12)'te verilmektedir (Aldrich ve Nelson, 1984).

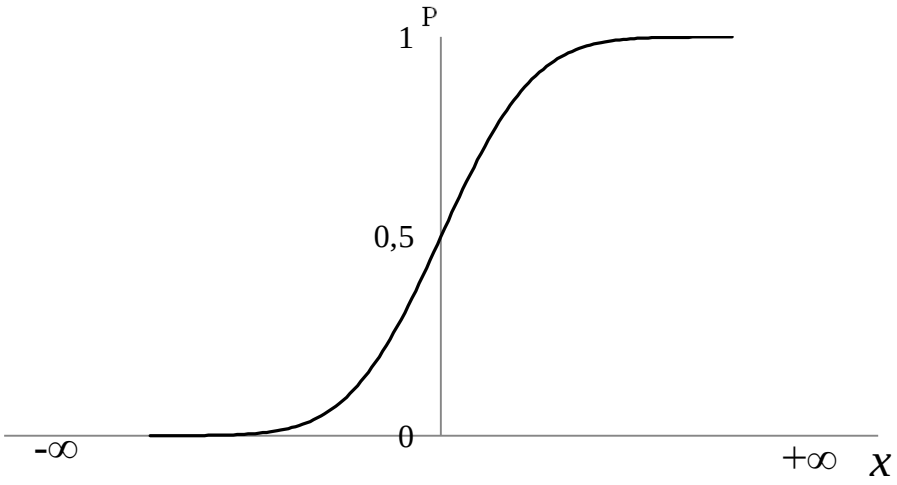
$$R^2 = \frac{\left[1 - \sum_{i=1}^{N} (Y_i - \sum_{k=1}^{K} \hat{\beta}_k x_{ik}) \right]^2}{\sum_{i=1}^{N} (Y_i - \bar{Y})^2} \tag{4.12}$$

Eşitlik (4.12)'de $\bar{Y}$ Y 'nin örneklem ortalamasını göstermektedir. Bu eşitlikle hesaplanan R^2 değeri EKK ile elde edilen R^2 'den daha küçük çıkmaktadır. Bu nedenle Aldrich ve Nelson (1984), gizli bağımlı değişkenli modellerde belirlilik

katsayısı değerinin bir özet istatistiği olarak kullanılmasından kaçınılması gerektiğini ileri sürmüştür.

4.2. İki Düzeyli Lojit Model

DOM'da karşılaşılan sorunları ortadan kaldırmak amacıyla geliştirilen modellerden biri Lojit Modeldir. Bu model özellikle $0 \le E(Y_i \mid x_i) \le 1$ varsayımının yerine gelmeme ve açıklayıcı değişkenlerin marjinal etkilerinin sabit kalması sorunlarından kurtulmak için dağılım fonksiyonları yardımıyla kurulan modellerden biridir. Lojit Model için kullanılan bu kdf, lojistik dağılıma aittir.



Şekil 4.5. Lojit Model

Bağımlı değişken *Y*'nin binomial olduğu yani 0, 1 gibi iki değer aldığı varsayılırsa, hata teriminin lojistik dağılımlı olduğu da varsayılabilir. Böylece Lojit Model veriye uygulanabilir ve bağ fonksiyonu Lojit ($\eta = \log[\mu/(1-\mu)]$) olur.

İki düzeyli *Y* ve bir açıklayıcı değişken *x* için olasılık ifadesi $\pi(x) = P(Y=1|x) = 1 - P(Y=0|x)$ olmak üzere, Lojistik regresyon modeli

$$\pi(x) = \frac{\exp(\alpha + \beta x)}{1 + \exp(\alpha + \beta x)} \qquad (4.13)$$

olarak ifade edilir. $\pi(x)$'in $1-\pi(x)$'e oranı odds oranı olarak adlandırılır. Bu oranın logaritması ise Lojit ismini alır ve aşağıdaki doğrusal ilişkiye sahiptir (Agresti, 2002).

$$lojit[P_i] = \log \frac{P_i}{1-P_i} = \sum_k \beta_k x_{ik} \qquad (4.14)$$

Eşitlik (4.14)'de P_i, i. gözleme ait olasılık değerini, x_{ik} açıklayıcı değişkeni, β açıklayıcı değişkene ait katsayıları ifade eder. Eşitliğin sol tarafındaki odds oranı dikkate alındığında, DOM'da meydana gelen olasılığın [0,1] aralığı dışında değerler alma sorunuyla karşılaşılmaz. Çünkü bu oranda üst sınırın 1, alt sınırın 0 olma kısıtı ortadan kaldırılmış olur. P_i 0'dan 1'e giderken $lojit[P_i]$ $-\infty$'dan $+\infty$'a gitmektedir. Olasılıklar [0,1] arasında yer alırken Lojitler için böyle bir sınırlama yoktur.

$lojit[P_i]$ ile açıklayıcı değişkenler x_{ik}'lar arasında doğrusal bir ilişki varken $\pi(x)$ ile açıklayıcı değişkenler arasındaki ilişki doğrusal değildir. DOM'da ise olasılık değeri ile açıklayıcı değişkenler arasındaki ilişki doğrusal olmaktadır ve bu durum daha önce bahsedilen sorunları ortaya çıkarmaktadır. Bu farklılık sebebiyle, Lojit Modelde bu sorunlarla karşılaşılmaz.

Eşitlik (4.15)'te kdf'nin yani *F*'nin yerine lojistik dağılım gösteren Λ gibi belirli bir kdf'nin koyulması durumunda *i*. gözlem için bağımlı değişkenin 1 olarak ifade edilen düzeyinin gözlenmesi olasılığı Lojit Modelin bir gösterimine dönüşecektir (Liao, 1994)

$$P(Y_i = 1) = 1 - F(-\sum_{k=1}^{K} \beta_k x_{ik}) = \Lambda(\sum_{k=1}^{K} \beta_k x_{ik}) = \frac{e^{\sum \beta_k x_{ik}}}{1 + e^{\sum \beta_k x_{ik}}} \qquad (4.15)$$

Bağımlı değişkenin 0 gözlenmesi olasılığı eşitlik (4.16) ile verilmiştir.

$$P(Y_i = 0) = F(-\sum_{k=1}^{K} \beta_k x_{ik}) = \Lambda(-\sum_{k=1}^{K} \beta_k x_{ik}) = \frac{e^{-\sum \beta_k x_{ik}}}{1 + e^{-\sum \beta_k x_{ik}}} = \frac{1}{1 + e^{\sum \beta_k x_{ik}}} \qquad (4.16)$$

Lojistik fonksiyonu ifadesi süreklidir ve 0'dan 1'e kadar tüm değerleri alabilir. $Z_i = \sum \beta_k x_{ik}$ ifadesi kullanılırsa, Z_i değeri $-\infty$'a yakın olduğunda $P(Y_i = 1)$ olasılığı 0'a yaklaşacak, monoton olarak artacaktır. Z_i $+\infty$'a yaklaştığında ise $P(Y_i = 1)$ olasılığı 1'e yaklaşacaktır. Böylece $Z_i = 0$ noktası etrafında simetrik olan S şeklinde bir eğri oluşacaktır. Bu dönüşüm sayesinde Z_i'ye ve dolayısıyla x_{ik} ve β_k'lara hiçbir kısıt getirmeksizin olasılık değerinin [0,1] aralığında olma durumu

gerçekleşmiş olur. Olasılık üzerinde yapılan bu dönüşüme Lojit teriminden dolayı, Lojit dönüşüm adı verilir.

4.3. İki Düzeyli Probit Model

DOM'da karşılaşılan sorunlardan kaçınmak amacıyla, alternatif olarak önerilen diğer bir model Probit Modeldir. Lojit Modelle arasındaki fark kullandıkları dağılım fonksiyonlarından kaynaklanmaktadır. Lojit Modelde lojistik dağılım kullanılırken, Probit Modelde normal dağılım fonksiyonu kullanılmaktadır. Normal dağılım kullanıldığından normit olarak da isimlendirilmesine karşın, daha çok Probit olarak anılmaktadır.

Probit Model de Lojit Model gibi GDM ailesinin bir üyesidir. GDM'de hatanın dağılımına ilişkin varsayım, kullanılacak modelin belirlenmesini sağlar. Modele ilişkin hata dağılımının standart normal dağılım olarak belirlenmesi ile Probit bağ fonksiyonunu ($\eta = \Phi^{-1}(\mu) = \sum_{k=1}^{K} \beta_k x_{ik} = Z$) kullanan olasılık modeli elde edilmiş olur. Kullanılan bağ fonksiyonundan dolayı bu olasılık model Probit Model adını almaktadır (Aldrich ve Nelson, 1984).

x değişkeni μ ortalama ve σ^2 varyansla normal dağılım gösteriyorsa, x değişkenine ait sırasıyla olasılık yoğunluk ve dağılım fonksiyonları aşağıda verilmiştir.

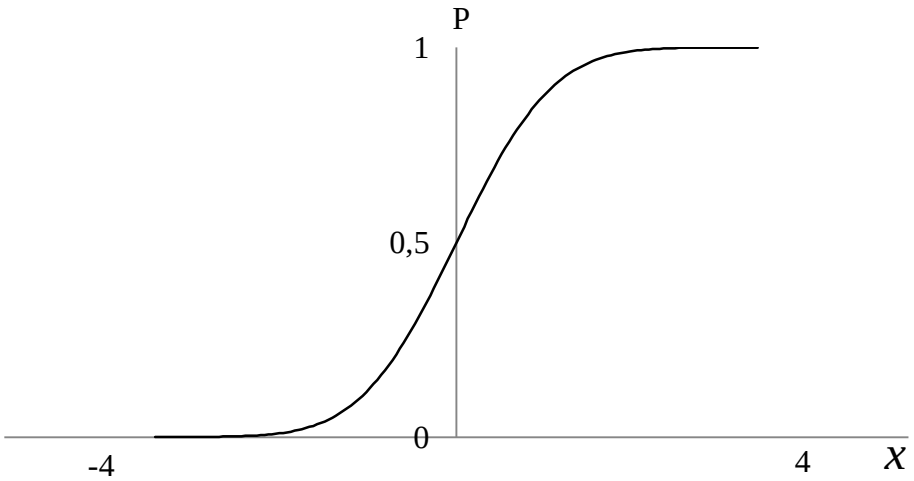
$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}, \quad -\infty < x < \infty \tag{4.17}$$

$$F(x) = \int_{-\infty}^{x_0} \frac{1}{\sqrt{2\pi}\sigma} e^{-(t-\mu)^2/2\sigma^2} dt \tag{4.18}$$

Probit Model de Lojit Model gibi gizli değişken yaklaşımını kullanır ($Y_i^* = \sum \beta_k x_{ik} + u_i$). Gizli değişkene ait eşitlik, Y_i^*'ın işaretini değiştirmeksizin keyfi pozitif bir sabit sayıyla çarpıldığında meydana gelen değişim nitel bağımlı değişkeni Y_i'nin gözlem değeri üzerinde herhangi bir etki yaratmayacaktır. Çünkü Y_i'nin alacağı değer Y_i^*'ın büyüklüğüyle değil işaretiyle ilgilidir. Böylece hem Probit, hem

Lojit analiz, keyfi ölçümlerin dışındakileri belirlemek için normalleştirmeyi kullanırlar. Probitte uygun normalleştirme 1'dir, yani u_i'nin standart sapmasını 1 olarak belirler. Probit ifadesi, eşitlik (4.18)'deki ortalamanın 0, varyansın 1 olmasıyla, standart normal rasgele bir değişken için dağılım fonksiyonu aşağıdaki gibi ifade edilebilir.

$$\Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} \exp(-u^2/2)du \tag{4.19}$$



Şekil 4.6 Probit Model

Eşitlik (4.20)'de *F*'nin yerine standart normal dağılım gösteren Φ gibi belirli bir kdf'nin koyulması durumunda *i*. gözlem için bağımlı değişkenin 1 olarak ifade edilen düzeyinin gözlenmesi olasılığı Probit Modelin bir gösterimine dönüşecektir.

$$P(Y_i = 1) = 1 - F(-\sum_{k=1}^{K} \beta_k x_{ik}) = F(\sum_{k=1}^{K} \beta_k x_{ik}) = \Phi(\sum_{k=1}^{K} \beta_k x_{ik}) \tag{4.20}$$

Probit Modelin açık ifadesi ise eşitlik (4.21)'de verilmektedir.

$$P(Y = 1 \mid x) = \Phi(\sum_{k=1}^{K} \beta_k x_{ik}) = \int_{-\infty}^{\sum_{k=1}^{K} \beta_k x_{ik}} \frac{1}{\sqrt{2\pi}} \exp(-u^2/2)du \tag{4.21}$$

4.4. İki Düzeyli Lojit ve Probit Modellerin Varsayımları

Bağımlı değişken Y'nin iki düzeyli olduğu ve 0, 1 gibi iki değer aldığı varsayılırsa, P parametresi $Y=1$ olması olasılığını temsil etmek üzere, P parametresinin değeriyle ilgilenilir ve $P=P(Y=1)$ şeklinde ifade edilir.

Y'nin gözlemlenebilen K adet X_k, $(k=1,...,K)$ açıklayıcı değişkenine bağlı olduğu ve bu açıklayıcı değişkenlerin P'de gerçekleşen tüm değişkenliği açıklayabildiği varsayılır. Bu ilişki $P=P(Y=1|x_1,...,x_k)$ ya da daha kısaca $P=P(Y|x)$ şeklinde gösterilir. Burada x, K adet açıklayıcı değişkenler kümesini temsil etmektedir. Bu varsayım standart regresyon modeline benzer ve açıklayıcı değişkenler ortalama ya da Y'nin beklenen değerindeki değişkenliği açıklar.
Doğrusal regresyondaki diğer bir varsayım Y ve x'lerin doğrusal ilişkiye sahip olduğudur. Burada ise Y'nin 1 olarak gözlenmesi ve x değişkenleri ile arasındaki ilişki hakkında farklı bir varsayım yapılır. Lojit Model için eşitlik (4.15) 'te ve Probit Model için bu ilişki eşitlik (4.20)'de verilmiştir.

Doğrusal regresyonda verinin N büyüklüğündeki rasgele bir örneklemden oluşturulduğu varsayılır ve i ile gözlemler indekslenir $(i=1,...,N)$. Bu varsayım Y üzerindeki gözlemlerin istatistiksel olarak birbirinden bağımsızlığını gerektirir. Doğrusal regresyondaki sabit varyans varsayımına benzer bir varsayıma burada ihtiyaç duyulmaz çünkü yukarıdaki (4.15) ve (4.20) eşitlikleri dolaylı yoldan bunu sağlar.

Doğrusal regresyondaki gibi burada da açıklayıcı değişkenler, rasgele değişkenler (anket sorularının cevapları gibi) olabilirler. Burada önemli olan şey değişkenler arasında çoklu bağlantı sorunun olmamasıdır. Bu varsayım örneklem büyüklüğünün parametre sayısından büyük $(N>K)$ olmasını gerektirir. x_k'lar gözlemlere karşı bir miktar değişim göstermelidirler ve iki veya daha fazla x_k tam mükemmel bir ilişkiye sahip olmamalıdır. Bu varsayımlar, klasik regresyonla tamamen aynıdır. Eğer tam değil de yakın bir doğrusal bağımlılık olursa bile ölçümsel belirsizlik problemleri, tutarsız tahminler ve büyük örneklem hataları meydana gelebilir. Probit ve Lojit Modeller, çoklu bağlantı problemlerinde klasik regresyonda olduğu gibi yetersiz kalırlar. Lojit ve Probit Modele ilişkin varsayımlar aşağıdaki gibi özetlenebilir;

1. $Y_i \in \{0,1\}$, $i=1,...,N$

2. $P(Y_i = 1 | x_i) = \exp(\beta_k x_{ik}) / [1 + \exp(\beta_k x_{ik})]$: Lojit için lojistik yoğunluk fonksiyonudur.

$P(Y_i = 1 | x_i) = \Phi(\sum \beta_k x_{ik})$: Probit için normal yoğunluk fonksiyonudur.

3. $Y_1, Y_2, ..., Y_N$ istatistiksel olarak bağımsızdırlar.

4. x_{ik}'lar arasında tam ya da yakın doğrusal bağımlılık yoktur (Aldrich ve Nelson, 1984)

4.5. İki Düzeyli Probit ve Lojit Modellerin Yorumu

Regresyon denkleminde β_k katsayısı, k. açıklayıcı değişkenin Y'nin üzerindeki ortalama etkisinin değerini ölçer. İki düzeyli Y değişkenin ortalama değeri, $Y = 1$ olma olasılığına eşit olur. Böylece Doğrusal Olasılık Modeli nde β_k, x_k'daki bir birimlik değişimin $P(Y = 1)$ olasılığı üzerindeki etkisini ölçer ve bu etki x_k'nın tüm değerleri için aynıdır, çünkü model doğrusaldır.

Probit ve Lojit Modellerde ise, x_k'lar ve $P(Y = 1)$ arasındaki doğrusal olmayan ilişki, $P(Y = 1)$ üzerindeki x_k'nın değişiminden kaynaklanan etkinin yorumlanmasını zorlaştırmaktadır. Doğrusal durumda etkiler daha açık bir şekilde görülürken, doğrusal olmayan durum için aynı şey söz konusu değildir.

Lojit ve Probit fonksiyonları kullanılarak bir modele ilişkin katsayılar elde edildiğinde, aynı açıklayıcı değişken için bulunan katsayılar birbirinden farklı olacaktır ve Z ($Z = \sum \beta_k x_k$) değerleri de elde edilen bu katsayı değerlerine bağlı olacağından, bulunan bu değerler her iki analizde farklılık gösterecektir. İki analizden elde edilen olasılık değerleri farklı olmakla birlikte, oldukça yakın değerlere sahiptirler. $Z = 0$ olduğunda ise, hesaplanan $P(Y = 1)$ olasılığı her ikisinde de 0.5 değerini alacaktır. Bunun anlamı, bu noktada her iki eğrinin simetrik olduğudur.

Z_i değerleri açıklayıcı değişkenlerin doğrusal fonksiyonudur ve bu nedenle karşılık gelen β_k'nın büyüklüğüne ve işaretine bağlı olarak her x_{ik} değeri ile değişir. $P(Y_i = 1)$ olasılık değerleri, Z_i değerlerine bağlı olarak değişmektedir ve Z_i'nin

negatif ve mutlak değerce en büyük noktasında sıfır değerini alır. Z_i'nin pozitif en büyük noktasında ise, olasılık bir değerini alacaktır. Ancak, Z_i değerlerine göre olasılığın değişim oranının sabit olmadığına dikkat edilmesi gerekir. Z_i mutlak değerce büyük ve negatif olduğunda, $P(Y_i = 1)$ olasılığı yavaşça artar. Z_i değerleri sıfıra yaklaşırken $P(Y_i = 1)$'deki değişim oranı yükselir ve Z_i değeri pozitif ve büyük olduğunda $P(Y_i = 1)$ değerinin artışı yine yavaşlar. β_k'nın etkilerin yönünü belirlediği görülür. Ancak etkilerin büyüklüğü Z_i'nin büyüklüğüne bağlıdır ve dolayısıyla tüm x_{ik}'ların büyüklüğüne bağlı olacaktır (Aldrich ve Nelson,1984).

Marjinal etkiler sabit olmadığından, x_k'nın $P(Y = 1)$ olasılığı üzerine etkisini değerlendirmek doğrusal regresyon modeline göre daha zordur. Bunu yapmanın bir yolu ilgilenilen açıklayıcı değişkenleri seçip $P(Y = 1)$ olasılığını hesaplamak ve sonrasında ilgilenilen x_k değerini az miktarda değiştirerek olasılık değerini tekrar hesaplamaktır. Hesaplanan bu değerlerden sonra değişim oranını bulmak gerekir. Yani $dP(Y = 1)/dx_k$ hesaplanır. Burada, $dP(Y = 1)$ iki olasılık değeri arasındaki farkı, dx_k ise ilgili x_k üzerindeki değişim miktarını ifade eder. dx_k çok küçük olduğu zaman bu değişim oranı x_k'ya göre $P(Y = 1)$'in türevini verir. Bu durumda, sırasıyla Probit ve Lojit için değişim oranları aşağıdaki gibi elde edilir:

$$\frac{dP(Y=1)}{dx_k} = \frac{1}{\sqrt{2\pi}} \exp(-Z^2/2)\beta_k \equiv \Phi(Z)\beta_k \qquad (4.22)$$

$$\frac{dP(Y=1)}{dx_k} = \frac{\exp(Z)}{1+\exp(Z)} \cdot \frac{1}{1+\exp(Z)} \beta_k = P(Y=1)[1-P(Y=1)]\beta_k \qquad (4.23)$$

Her iki modelde de β_k terimi çarpımsal bir faktördür ve etkinin işaretini belirler. Çünkü, eşitliklerdeki diğer çarpım halindeki faktörlerin pozitif olması gerekir. Ama x_k'nın $P(Y = 1)$ olasılığı üzerindeki etkisi her iki durumda da, Z'nin doğrusal olmayan fonksiyonu tarafından azaltılmıştır. Bu azalma Z sıfır olduğunda minimum olur ve Z büyüdükçe pozitif ya da negatif olarak artar.

Sonuç olarak, x_k'daki değişimin $Y = 1$ olarak gözlenmesi olasılığına etkisinin, tamamen olmasa da, β_k ile ilgili olduğu açıkça görülür. β_k'nın işareti etkinin

yönüne bağlıdır ve etki büyük olmaya meyilli ise β_k da büyük olur. Bu nedenle, nitelik bakımından β_k'nın yorumu doğrusal regresyon modelindekiyle benzerdir. Ama etkinin büyüklüğü açıklayıcı değişkenlerin değerleri ile değiştiği için etkiyi tanımlamak bu kadar basit değildir. Bu durum genellikle, karşılık gelen x değerlerine göre değişen dP/dx_k'nın değer aralıklarını göstermek için tablo ve grafik kullanma ihtiyacını ortaya çıkarır (Aldrich ve Nelson, 1984).

Doğrusal regresyon modelinde açıklayıcı değişkene ait tahmin edilen katsayı, o değişkenin değerindeki bir birimlik değişmeye karşılık bağımlı değişkenin ortalama değerinde yaratacağı etkiyi ölçer. DOM, Lojit, Probit Modelleri bir olayın gerçekleşme olasılığıyla ilgilendiklerinden, katsayıların yorumlanması aşamasında dikkatli olmak gerekir. DOM'da açıklayıcı değişkene ait katsayısı, o değişkende meydana gelen bir birim değişmeye karşılık, ilgilenilen olayın gerçekleşmesi olasılığındaki değişmeyi doğrudan ölçmektedir. Lojit Modelde katsayılarda meydana gelen değişimin olasılıktaki değişme oranı eşitlik (4.23)'da belirtilmiştir ve burada β_k *k*. açıklayıcı değişkenin katsayısıdır. Probitteki değişim oranı ise (eşitlik (4.22)) biraz daha karmaşıktır ve standart normal değişkenin yoğunluk fonksiyonunu içerir. Lojit ve Probit Modellerinde olasılıktaki değişimin hesaplanmasında bütün açıklayıcı değişkenler dikkate alınırken, DOM'da yalnızca *k*. açıklayıcı değişken dikkate alınır. Bu farkın, DOM'un geçmişte yaygın olarak kullanılmasına neden olduğu düşünülebir (Gujarati, 1999).

4.6. En Çok Olabilirlik, En Küçük Kareler ve Ağırlıklı En Küçük Kareler Tahminlerinin Karşılaştırılması

Bilinen doğrusal regresyon modelinin varsayımları sağlanıyorsa, EKK tahmin edicileri yansız ve etkin olurlar. Buna ek olarak hata terimi normal dağılıyorsa, tahminlerin de dağılımı biliniyor ve normal dağılım gösteriyor demektir. Böylece, istatistiksel testlerin yapılması mümkün olur. Eğer varsayımlardan herhangi biri gerçekleşmezse bu özelliklerin hepsi veya bir kısmı kaybedilebilir. Doğrusal Olasılık Modeli nde özellikle sabit varyans ve normallik varsayımlarının gerçekleştiği savunulamaz. Bu durumda EKK yöntemi yerine ağırlıklı en küçük kareler (AEKK) yönteminin kullanılması daha uygun olacaktır. Bu sayede ağırlıklandırılmış

tahminler yansızlık, etkinlik, normallik gibi istatistiksel özellikleri büyük örneklemlerde sağlar.

EÇO yöntemi çoğu tüm durumda EKK yöntemine alternatif olarak başvurulabilecek bir yöntemdir. Küçük örneklerde bu yöntem arzulanan tüm özellikleri (yansızlık, etkinlik, normallik) sağlamasa da, örneklem genişliğinin artmasıyla birlikte yaklaşık olarak bu özellikler sağlanır. Böylece EÇO tahmin edicisi yansızlık, etkinlik ve normalliğin asimptotik özelliklerini gösterir. Bazı özel durumlarda tam sonuçlar da elde edilebilmiştir. Probit ve Lojit Modeller bunlar arasında değildir.

DOM'un AEKK tahminleri ile Probit ya da Lojitin EÇO tahmin edicisi benzer istatistiksel özelliklere sahip tahminler ürettiklerinden, istatistiksel performansları modeller arasındaki seçimin temeli olarak kullanılamazlar. Ama her durum için istatistiksel özelliklerin sağlanması, model varsayımlarının doğrulanması durumunda gerçekleşir. Modellerin varsayımları arasındaki fark *Y*'nin beklenen değerinin yapısından kaynaklanır. Bu durum, modeller arasında ve tahmin süreçleri seçiminde anahtar rol oynar.

EÇO tahmin edicisinin dezavantajı, Probit ve Lojit Modeller için olabilirlik eşitliklerinin tahmin edilecek parametre açısından doğrusal olmamasıdır. Bunun anlamı, standart cebirsel işlemlerin yapılamayacak olmasıdır. Bu durumda iteratif yöntemlere başvurulması gerekir.

Probit ve Lojit Modeller üzerinde, en çok olabilirliğin istatistiksel özellikleri hakkında özetleme yapmak gerekirse; tahminler regresyon modelin EKK tahminleriyle hemen hemen aynı özelliklere sahiptir. İkisi arasındaki faklılıklar; en çok olabilirliğin bu modellerde doğrusal olmaması nedeniyle hesaplamada daha karmaşık matematiksel işlem kullanmayı gerektirmesi, özelliklerin asimptotik olmasıdır. Örneklem büyüklüğü arttıkça istatistiksel anlamda daha iyi olurlar. Bunun yanında bazı benzer özellikler göze çarpar. Bu özellikler yansızlık, etkinlik ve normalliktir. Son özelliği kullanmak için katsayı tahminlerinin kovaryans ve varyanslarının tahminine ihtiyaç duyulur. Uygun tahmin log-olabilirliğin ikinci türevinden elde edilen matrisin tersinin negatifidir ve iteratif maksimizasyon algoritmaları ve daha çok bilgisayar programlarıyla elde edilebilir (Aldrich ve Nelson, 1984).

4.7. Doğrusal Olasılık Modeli, Lojit ve Probit Modelin Karşılaştırılması

Kümülatif normal dağılım ve lojistik dağılım, kuyrukları hariç birbirlerine çok benzeyen dağılımlardır. Bu nedenle iki modele ait eşitliklerden elde edilen sonuçlar, örneklem çok büyük olmadıkça, çok farklı olmayacaktır. Örneklemin büyük olması kuyruklarda daha fazla gözlem olmasına sebep olacaktır ve bu durum farklılığı yaratabilir. Yinede iki yöntemden elde edilen β'lar doğrudan karşılaştırılabilir değildir (Maddala, 1983).

Lojit Modelden elde edilen β tahminlerinin, Probit Modelden elde edilen tahminlerle karşılaştırılabilir olması için, Lojitin varyansı $\pi^2/3$ olduğundan, $\sqrt{3}/\pi$ ile çarpılmalıdır (Maddala, 1983). Bu sayede sıfır ortalamaya sahip, birim varyanslı lojistik dağılıma ilişkin tahminler elde edilmiş olur.

Amemiya (1981), Lojit tahminlerinin $\sqrt{3}/\pi$ yerine 0.625 ile çarpılmasını önermektedir. Bu dönüşümün lojistik dağılım ve standart normal dağılım fonksiyonu arasında daha iyi bir yaklaşım ürettiğini iddia etmektedir.

Lojistik yoğunluk fonksiyonu, normal yoğunluk fonksiyonundan daha yavaş azalma gösterdiğinde Lojitten tahmin edilen katsayıların, Probitten elde edilen katsayılara oranı ($\beta_{lojit}/\beta_{probit} = 1/0.625 = 1.6$) 1.6'dan büyük olma eğilimi gösterir. Açıklayıcı değişkenlerin sürekli olması durumunda değişme oranlarını, marjinal etkileri bağımlı değişkenin beklenen değer fonksiyonunun türevi hesaplanarak bulmak mantıklı sonuçlar vermektedir. Ancak modelde gölge (yapay) değişkenlerin yer alması durumunda türevler veya marjinal etkiler anlamlı olmayabilir. Bu durumda, örneklem tahminleri kullanılarak β'nın x aralığı üzerinde $P(Y=1)$ olasılığı hesaplanarak gölge değişkenin etkisi analiz edilebilir.

Lojitte ve Probitte kullanılan dağılımların benzerliğinden dolayı büyük gözlemler olmadıkça iki model arasında istatistiksel olarak ayırım yapmak güçtür. Büyük örneklemlerde Lojit ve Probit Modeller tarafından elde edilen sonuçlar anlamlı olarak farklılaşır. Örneklem hacmi artarken iki model farklı marjinal etkiler meydana getirir.

Modelde tek açıklayıcı değişkenin yer alması durumunda kümülatif normal ve kümülatif lojistik dağılımlar arasında çok önemli bir fark yoktur. Dolayısıyla

verilerin yoğun olarak kuyruk kısımlarında yoğunlaştığı durumlar dışında iki modelden hangisinin kullanılacağı hususunda bir fark oluşmaz. Lojit Model, verilerin dağılımının kuyruk kısımlarında yoğunlaştığı durumlarda daha iyi sonuç verir.
Amemiya DOM ile Lojit Modellerin katsayıları arasındaki ilişkiyi ise aşağıdaki gibi ifade etmiştir (Amemiya, 1981).

$\beta_{DOM} \cong 0.25\beta_{lojit}$ sabit terim bunun dışındadır.

$\beta_{DOM} \cong 0.25\beta_{lojit} + 0.5$ sabit terim için.

DOM katsayıları ile Probitten elde edilen katsayılar arasında ilişkide ise 0.4 çarpanı kullanılır.

$\beta_{DOM} \cong 0.4\beta_{probit}$ sabit terim bunun dışındadır.

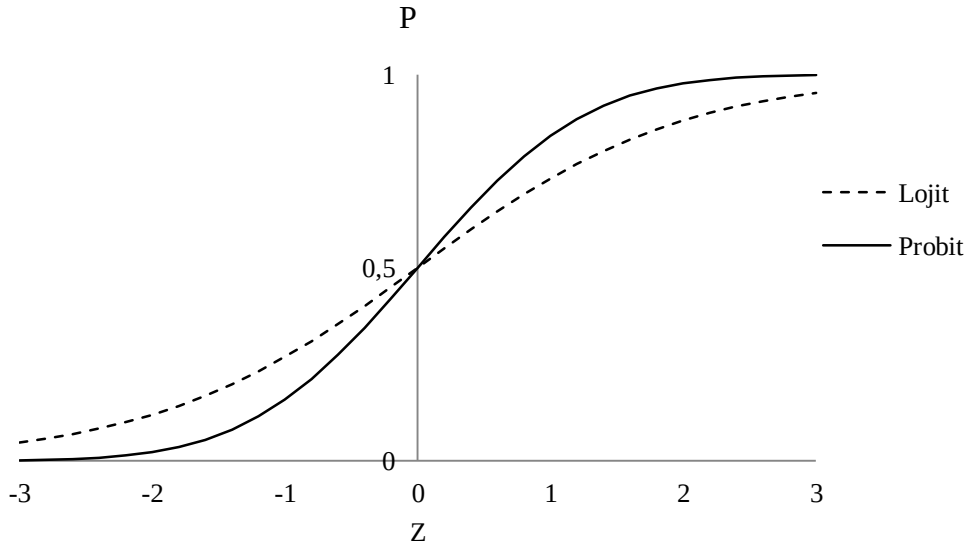
$\beta_{DOM} \cong 0.4\beta_{probit} + 0.5$ sabit terim için.

Lojit ve Probit Modelleri karşılaştırmak için, tahmin edilen olasılıklardan sapmaların kareleri toplamı hesaplanabilir, doğru sınıflandırma yüzdeleri karşılaştırılabilir, bağımsız değişkene göre olasılıkların türevlerine bakılabilir (Maddala, 1983).

Modellerin karşılaştırılması için hesaplanan türev değerleri DOM'da sabitken, Lojit ve Probitte açıklayıcı değişkenin farklı seviyelerine göre değişmektedir. Tablo 4.1'de Lojit ve Probit Modelden elde edilen değerlerin oldukça yakın olduğu ve noktasında eşit olduğu görülmektedir. Hesaplama açısından daha kolay olması nedeniyle genellikle Probit yerine Lojit Model tercih edilmektedir (Altıntaş, 2006).

Tablo 4.1 Lojit Model ve Probit Modelin olasılıklarla karşılaştırılması

Z	Normal $P(Z)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{Z} e^{-t^2/2}\,dt$	Lojistik $P(Z)=\frac{1}{1+e^{-Z}}$	Z	Normal $P(Z)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{Z} e^{-t^2/2}\,dt$	Lojistik $P(Z)=\frac{1}{1+e^{-Z}}$
-3	0.00135	0.04743	0.2	0.57936	0.54983
-2.8	0.00255	0.05732	0.4	0.65542	0.59869
-2.6	0.00466	0.06914	0.6	0.72575	0.64566
-2.4	0.00819	0.08317	0.8	0.78814	0.68997
-2.2	0.01390	0.09975	1	0.84134	0.73106
-2	0.02275	0.11920	1.2	0.88493	0.76852
-1.8	0.03593	0.14185	1.4	0.91924	0.80218
-1.6	0.05479	0.16798	1.6	0.94520	0.83202
-1.4	0.08075	0.19782	1.8	0.96407	0.85815
-1.2	0.11507	0.23147	2	0.97725	0.88080
-1	0.15865	0.26894	2.2	0.98610	0.90025
-0.8	0.21185	0.31003	2.4	0.99180	0.91683
-0.6	0.27425	0.35434	2.6	0.99534	0.93086
-0.4	0.34458	0.40131	2.8	0.99744	0.94268
-0.2	0.42074	0.45016	3	0.99865	0.95258
0	0.50000	0.50000			



Şekil 4.7 Lojit Model ve Probit Model dağılımları

4.8. İki Düzeyli Bağımlı Değişken için Uyum İyiliği Ölçütleri

Uyum iyiliği ölçüsü, kullanılan modelin verilere ne kadar uyum sağladığının ifadesidir. Eldeki verilere dayanarak elde edilen tahmini model ya da eğri gerçek gözlem değerlerine ne kadar iyi uyum sağladığı ya da onları ne kadar iyi temsil ettiği

uyum iyiliği ölçüleri kullanılarak bulunmaya çalışılır. Bağımlı değişkendeki değişimin, açıklayıcı değişkenler ile ne derece açıklanabildiği uyum iyiliği ölçütleri ile ortaya konulur. Açıklayıcı değişkenler, bağımlı değişkeni ne kadar iyi açıklıyorsa, ölçüt değerleri de o derece iyi çıkacaktır.

Bir modelin uyum iyiliği, modelin varyansının küçük olmasına, sistematik eğilim göstermemesine ve modelin değişkenliği ile ilgili kullanılan varsayımları yerine getirmesine bağlıdır. Uyumun eksikliği, bu üç karakteristik özelliğin bir veya birkaçının ihlalinden meydana gelir (Hosmer vd.,1997).

Doğrusal regresyon modelinde uyum iyiliği ölçüsü olarak R^2 ya da diğer bir adıyla belirlilik katsayısı kullanılmaktadır. R^2'nin yaygın kullanımının yanında, zaman zaman yanlış kullanımı sebebiyle birçok istatistikçi tarafından sevilmeyen bir ölçüdür. Bu konu üzerine açıklayıcı ve uyarıcı söylemler mevcuttur. Sabit katsayının yer almadığı modeller, dönüştürülmüş değişkenler ve ağırlıklı en küçük kareler için R^2'nin kullanımında hatalarla karşılaşılmaktadır. R^2'nin kullanımından vazgeçilmesi yakın zamanda olası gözükmemektedir ve R^2'nin daha iyi şekilde kullanımının geliştirilmesi önemli olacaktır (Anderson-Sprecher, 1994)

Doğrusal regresyonda kullanılan R^2, HKT hata kareler toplamını ve GKT genel kareler toplamını göstermek üzere aşağıdaki şekilde ifade edilmektedir (Magee, 1990).

$$R^2 = 1 - \frac{HKT}{GKT} = 1 - \frac{\sum (Y - \hat{Y})^2}{\sum (Y - \bar{Y})^2} \qquad (4.24)$$

Burada $\hat{Y}$, $Y = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k$ şeklindeki doğrusal modelden elde edilen bağımlı değişkene ait tahmini değerleri, $\bar{Y}$ bağımlı değişkenin ortalama değerini ifade etmektedir.

Yukarıdaki eşitlik için $Y = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k$ modeli, hata kareler toplamı $\sum (Y - \hat{Y})^2 = HKT_{(tam)}$ olmak üzere, tam model olarak ifade edilebilir. $\beta_1 = \beta_2 = ... = \beta_k = 0$ olduğunda $Y = \beta_0$ sınırlandırılmış model elde edilir ve hata kareler toplamı $\sum (Y - \bar{Y})^2 = HKT_{(Sınırlı)}$ şeklinde tanımlanır. Bu ifadelerle beraber

genelleştirilmiş belirlilik katsayısı, $GR^2 = 1-[HKT_{(tam)} / HKT_{(sınırlı)}]$ şeklinde ifade edilir. Bu gösterimin iki avantajı vardır. Birincisi, R^2'nin bir model karşılaştırmasının ifadesi olmasıdır. Elde edilen R^2 değerleri tam modelin ne kadar iyi olduğunun kesin bir göstergesi değil, tam modelin sınırlı modeli ne kadar içerdiğinin bir ifadesidir. Sonuç olarak, büyük R^2 değerlerinin model yeterliliğini gösterdiği konusundaki yaygın inanış kolayca çürütülebilir. Diğer bir avantajı, R^2 'nin iç içe geçmiş modeller arasında yapılan karşılaştırmanın bir göstergesi olmasıdır. Bir model diğer modelin alt modeli değilse değişimdeki azalma oranı kavramı güvenilir değildir ve GR^2 ifadesinin kullanılması bu gibi hataları engeller. Özel olarak, orijinden geçen model için R^2 yorumlanırken hata yapılabilmektedir. $Y = \beta_0$ modeli $Y = \beta_1 X$ modelinin bir daralması olmadığı halde öyleymiş gibi dikkate alınması hataya sebep olur. (Anderson-Sprecher 1994).

Hangle ve Mitchell (1992) R^2'nin bazı araştırmacılar tarafından çok faydalı bulunmadığını, bazıları tarafındansa önemli derecede kullanışlı bulunduğunu ifade etmiştir. Bazen yanlış kullanılmasına karşın R^2, model performansının ölçülmesini sağlar. Örneğin, R^2 regresyon ile bağımlı değişkende meydana gelen değişimin, yüzde açıklama oranı olarak kullanılır. Ayrıca R^2 sabit katsayı hariç diğer katsayıların sıfırdan farklı olup olmadığının hipotez testi için F test istatistiğine benzer olarak kullanılabilir. R^2 terimi ile F istatistiği ifadesi N gözlem sayısını, k sabit hariç, modelde yer alan katsayıların sayısını göstermek üzere aşağıdaki gibi verilebilir.

$$F = \frac{R^2 / k}{(1-R^2)(N-k)} \tag{4.25}$$

Ratkowsky (1990) R^2'nin doğrusal olmayan modellerde herhangi belli bir anlamı olmadığını belirtmiştir. Korelasyon katsayısı ya da HKT/GKT ifadesi temel alınan tanım kullanıldığında bu ifade kesinlikle doğrudur. Doğrusal olmayan bir regresyon modeli için en küçük kareler temel alındığında ve referans olarak kullanılan sezgisel mantıklı bir alt model var olsa bile GR^2 eşitliği yorumlanabilir.

Örneğin; $Y = \alpha / [1 + \exp(\beta + \gamma x)]$ modeli boş $Y = \alpha$ modeli ile karşılaştırılabilir. Bu durumda GR^2, doğrusal korelasyon ya da F testiyle ilişkili olsa da model karşılaştırması geçerlidir. Ayrıca her uygulama için alt model olarak bir boş modelin var olmayabileceğine dikkat etmek gerekir (Anderson-Sprecher, 1994).

Doğrusal olmayan regresyon modelleri için birçok araştırmacı bilinen R^2'yi yetersiz bulduklarından ve seçilen modelin fonksiyonel yapısının doğruluğu konusunda iyi bir fikir vermediğinden alternatif başka R^2'ler önermişlerdir. Özellikle karar modelleri olmak üzere, birçok iktisadi uygulamada aralık düzeyli bağımlı değişken gözlemlenebilir değilse gizli değişken yaklaşımından faydalanılır. Gizli değişkenlerin söz konusu olduğu durumlarda, gözlemlenebilir bir modele dönüşüm yapılır. Bu durumlarda kullanılan Lojit ve Probit gibi doğrusal olmayan modeller için farklı temellere dayanan R^2 ölçütleri önerilmiştir. Bu ölçütler yapay R^2'ler olarak adlandırılırlar.

Eğer ilgili aralık düzeyli bağımlı değişken elde edilebilirse, tahmin yöntemi olarak EKK tercih edilebilir. Ancak pratikte gizli değişken gözlemlenebilir değildir. Gizli değişkenin elde edilebilir olmaması yapay R^2'lerin anlamlı şekilde karşılaştırılmasını da engeller. Çünkü deneysel veriden gizli değişkene ilişkin R^2'nin elde edilmesi mümkün olmamaktadır (Hagle ve Mitchell, 1992).

Literatürde önerilen birçok uyum iyiliği ölçüsü olmasına karşın hangisinin kullanılacağı hususunda kesin bir yargı bulunmamaktadır. Buna karar vermek için, kullanılacak R^2'nin hangi amaçla kullanılacağı dikkate alınabilir. Amacın değişim ölçüsü ölçmek, hipotez testi yapmak ya da bağımlı değişkenin doğru bir şekilde sınıflandırılmasını belirlemek olması, kullanılacak R^2 seçimini etkileyecektir. Önerilen yapay R^2'leri temel aldıkları ölçütleri dikkate alarak da değerlendirmek mümkün olabilir. Bu ölçütler; olabilirlik, log-olabilirlik, olasılık tahminleri temelli ve tahmini olasılıkların varyansının ayrıştırılması temelli ölçütlerdir. Örneğin; araştırmacı açıklanan varyans ile ilgileniyorsa, McKelvey ve Zavoina (1975), Aldrich ve Nelson (1984) ya da Maddala (1983) tarafından önerilen yapay R^2'lerden birini seçebilir.

Yapay R^2'ler de dahil olmak üzere her hangi bir uyum iyiliği istatistiği ideal olarak aşağıdaki özellikleri taşımalıdır:

1. 0-1 aralığında olmalıdır.

2. Farklı örneklerle elde edilmiş modelleri karşılaştırmada anlamlı olmalıdır.

3. İstatistiksel hipotez testlerinin yapılabilmesi için, bilinen bir olasılık dağılımına uymalıdır (Tardiff, 1976).

Bilinen R^2'si için Dhrymes (1986) tarafından aşağıda verilen üç özellik yapay R^2'ler için de istenilen özelliklerdir:

a. Gerçek açıklayıcı değişkenlerin katsayılarının sıfır olmasının testinde kullanılan F istatistiği ile bire bir ilişkili olmalıdır.

b. Bağımlı değişkendeki değişimin açıklayıcı değişkenler tarafından açıklanmasının bir ölçüsüdür.

c. Örneğe ilişkin bağımlı değişkenin gerçek ve tahmini değerleri arasındaki basit korelasyon katsayısının karesidir.

Magee (1990) EKK ilişkisini temel almak üzere a. özelliği "uyumun önemliliği yaklaşımı" olarak adlandırılmıştır.

$$R^2 = kF / (kF + N - k + 1) \qquad (4.26)$$

Eşitlik (4.26) k, sabit katsayıyı içermemek üzere, açıklayıcı değişkenlerin sayısını, F hipotez testinde kullanılan F istatistiğini gösterir. N ise gözlem sayısını göstermektedir. F ile R^2 arasındaki ilişkiye benzer bir ilişki, aynı sıfır hipotezi için, F yerine olabilirlik oran ya da ki-kare istatistiği için de geçerlidir. Magee (1990) R^2 ölçülerinin birçoğunun, her hangi bir tahmin için, bu ilişki kullanılarak oluşturulabileceğini belirtmiştir. Örneğin, uygun F istatistiği hesaplanarak yukarıdaki formülde yerine konulursa iki düzeyli bir Probit Model için R^2 oluşturulabilir. Bunun anlamı, tahmin edilmiş katsayılar ve varyans kovaryans matrisi mevcut olduğunda bir R^2 ölçüsü elde edilebilir olmasıdır. Yaygın olarak kullanılan McFadden R^2'si bu yapıyla ilişkilendirilebilir (Veall ve Zimmermann, 1996).

Yapay R^2'lerin hiçbiri doğrusal regresyondan elde edilen R^2'nin sahip olduğu özelliklerin tümüne sahip değildir. Bilinen R^2'nin tüm regresyon katsayılarının sıfıra

eşit olma hipotezinin F istatistiği ile olan ilişkisi, bağımlı değişkenin gerçek değeri ile tahmini değerinin korelasyon katsayılarının karesi ve açıklanan değişimin toplam değişime oranı olarak yorumlanabilmesi özellikleri yapay R^2'lerde yoktur. Doğrusal regresyondan elde edilen R^2 açıklama derecelerinin bir ölçüsü olarak regresyon modellerinin karşılaştırılmasında elde edilebilir standart bir indekstir. Ancak farklı eşitliklerin karşılaştırılmasında veri tipi hakkındaki bilgi önemlidir. Örneğin, iktisadi veriler için zaman serileri çalışmalarında elde edilen R^2 değerlerinin en yüksek 0.7 ya da 0.8 ya da daha da fazla olması beklenirken, bireylerin üzerindeki seçim verileri için elde edilen R^2'nin 0.4 değeri yüksek bir değer olarak dikkate alınmaktadır (Veall ve Zimmermann, 1994).

Bir R^2 ölçüsü 0 ile 1 arasında değerler almalıdır. Ölçülerin hemen hemen hepsi bu özelliği taşırlar. Yine de, bazı durumlarda alt sınır değeri olarak 0'dan daha az değerlerin elde edildiği görülmüştür. Diğer bir kriter bir R^2 ölçüsünün açıklayıcı değişkenlerin eklenmesi durumunda artma eğilimi göstermesidir (Veall ve Zimmermann, 1996).

Bir yapay R^2'nin en önemli özelliği farklı çalışmaları tutarlı bir şekilde karşılaştırmada kullanılabiliyor olmasıdır (Aldrich ve Nelson, 1984). Ancak farklı çalışmalarda farklı yapay R^2'lerin kullanılmış olması bunu engellemektedir.

4.8.1. İki düzeyli bağımlı değişken durumunda kullanılan bazı yapay R^2 ölçütleri

Kareli Korelasyon Katsayısı

Uyum iyiliği için önerilen en populer ölçülerden biri Y ve F arasındaki kareli korelasyon katsayısıdır.

$$R_C^2 = \frac{|\operatorname{cov}(Y,F)|}{\operatorname{var}(Y)\operatorname{var}(F)} = \frac{\operatorname{var}(F)}{\operatorname{var}(Y)} \tag{4.27}$$

Morrison (1972) bu ölçünün üst sınırının 1'in altında olduğunu iddia ederken Golberger (1973) bu ölçünün üst sınırının 1 olduğunu göstermiştir. Eğer (F_i, i. gözlem için birikimli dağılım fonksiyonunun değeri olmak üzere) (Y_i, F_i) rasgele

seçilen bir çift olarak düşünülürse, $E(Y) = E(F)$ ve $E(YF) = E(F^2)$ olduğundan $\text{cov}(Y, F) = E(YF) - E(Y)E(F) = E(F^2) - (E(F))^2 = \text{var}(F)$ elde edilir (Veall ve Zimmerman, 1996).

Lave'in yapay R^2'si

Lave (1968) tarafından önerilen ölçüt aşağıdaki gibidir.

$$R_L^2 = 1 - \frac{\sum_{i=1}^{N}(Y_i - F_i)^2}{\sum_{i=1}^{N}(Y_i - \bar{Y})^2} = 1 - \frac{\sum(Gözlem - Gözlemin\,Olasılık\,Tahmini)^2}{\sum(Gözlem - Gözlemlerin\,Ortalaması)^2} \quad (4.28)$$

Bu ölçüt 0 ve 1 arasında değerler alır ve farklı uygulamalar için karşılaştırılabilir. Diğer taraftan istatistiksel testlerin yapılabilmesi amaçlı bu ölçünün henüz hangi dağılıma uyduğu gösterilmemiştir (Tardiff, 1976).

R_C^2 ve R_L^2'nin her ikisi de F_i yerine $\hat{F}_i$ konularak uygulanmaktadır. Eşitlik (4.27) ile eşitlik (4.28) tahmin edilen R_C^2 ve R_L^2 değerleri nicel bağımlı değişkenli modellerde aynı olmayabilirken, standart regresyon durumunda aynıdır. Aralarındaki fark küçük örneklemlerde ihmal edilebilecek kadar küçüktür.

Birçok yapay R^2'de eşitlik (4.29)'da yer alan olabilirlik oran test istatistiği (LRT) temel alınmaktadır. Aşağıdaki LRT^* ifadesi, modelden elde edilebilecek maksiumum log olabilirlik değerine ilişkin olabilirlik oran istatistiğidir. Olabilirlik oran istatistiği daha sonra ayrıntılı şekilde incelenecektir.

$$l = \sum (1 - Y_i)\log[1 - F(.)] + Y_i \log F(.) \quad (4.29)$$

$$LRT = 2(l_M - l_0) \quad (4.30)$$

$$LRT^* = 2(l_{Max} - l_0) \quad (4.31)$$

Burada l_M tahmin edilen modelin olabilirliğinin logaritmasını ve l_{Max} mümkün olan en büyük log olabilirlik değerini ifade etmektedir. l_0 ise modelde açıklayıcı değişkenlere ait katsayıların sıfır olduğu ve modelde sadece sabit katsayı varken elde edilen modelin olabilirliğinin logaritmasını ifade etmektedir.

Aldrich ve Nelson'un yapay R^2'si

Aldrich ve Nelson (1984) iki düzeyli bağımlı değişkenli modellerde uyum iyiliğini değerlendirmek için kullanılabilen aşağıdaki ölçütü önermiştir.

$$R^2_{AN} = LRT / (LRT + N) \tag{4.32}$$

Aldrich ve Nelson R^2'si için bir düzeltme

İki düzeyli bağımlı değişken için gözlenen bir ve sıfırların sayısı eşitse log olabilirlik başlangıç değeri l_0'a eşit olur. Bu durumun gerçekleştiğini varsayıldığı takdirde, eğer bir model bağımlı değişkendeki değişimi açıklamada katkı sağlamıyorsa l_M l_0'a eşit olur. Bu durumda R^2_{AN} eşitliğinde pay sıfıra eşit olacaktır. Eğer model bağımlı değişkendeki tüm değişimi açıklıyorsa, $l_M = 0$ olacaktır. Bu durumda Aldrich ve Nelson'un formülü aşağıdaki şekilde yazılır.

$$R^2_{AN} = \frac{-2l_0}{N - 2l_0} \tag{4.43}$$

N_1, iki düzeyli bağımlı değişkenin 1 değerini alan toplam gözlem sayısını, N_0, 0 değerini alan toplam gözlem sayısını ifade etmek üzere $l_0 = N_0(\log(N_0 / N) + N_1(\log(N_1 / N))$ eşitliği geçerlidir. Bu ifadede birlerin ve sıfırların sayısı eşit olduğu durumda, $l_0 = -0.6931 \times N$ olarak elde edilir. Bu değer eşitlik (4.43)'te yerine yazılırsa 0.5809 değeri elde edilir.

$$R^2_{AN} = \frac{-2 \times (-0.6931 \times N)}{N - 2 \times (-0.6931 \times N)} = \frac{1.3863 \times N}{N \times (1 + 1.3863)} = 0.5809$$

Bu değer Aldrich ve Nelson'un yapay R^2'si için maksimum değer olacaktır. Elde edilen maksimum değeri bir yapay R^2 için olması istenen 1 üst sınırından oldukça uzaktır. Üst sınırın 1 olabilmesi için, R^2_{AN}'nin 1.7214 (=1/0.5809) ile çarpılması gerekir. Bu değerin veri setindeki gözlem sayısından bağımsız olduğu görülür. Çünkü eşitlikteki N çarpanı sadeleşerek kaybolmuştur (Hagle ve Mitchell, 1992). Tablo 4.2'de örneklemde gözlenen 1 değerlerinin toplam gözlem sayısına oranlarına karşılık gelen düzeltme çarpanları verilmiştir.

Tablo 4.2 R_{AN}^2 için düzeltme çarpanı değerleri

N_1 / N	l_0 / N	Aldrich ve Nelson maksimum R_{AN}^2	Düzeltme Çarpanı
0.5	-0.6931	0.5809	1.72
0.51	-0.6929	0.5809	1.72
0.52	-0.6923	0.5807	1.72
0.53	-0.6913	0.5803	1.72
0.54	-0.6899	0.5798	1.72
0.55	-0.6881	0.5792	1.73
0.56	-0.6859	0.5784	1.73
0.57	-0.6833	0.5775	1.73
0.58	-0.6803	0.5764	1.73
0.59	-0.6769	0.5751	1.74
0.60	-0.6730	0.5737	1.74
0.61	-0.6687	0.5722	1.75
0.62	-0.6641	0.5705	1.75
0.63	-0.6590	0.5686	1.76
0.64	-0.6534	0.5665	1.77
0.65	-0.6474	0.5642	1.77
0.66	-0.6410	0.5618	1.78
0.67	-0.6342	0.5592	1.79
0.68	-0.6269	0.5563	1.80
0.69	-0.6191	0.5532	1.81
0.70	-0.6109	0.5499	1.82
0.71	-0.6022	0.5463	1.83
0.72	-0.5930	0.5425	1.84
0.73	-0.5833	0.5384	1.86
0.74	-0.5731	0.5340	1.87
0.75	-0.5623	0.5293	1.89
0.76	-0.5511	0.5243	1.91
0.77	-0.5393	0.5189	1.93
0.78	-0.5269	0.5131	1.95
0.79	-0.5140	0.5069	1.97
0.80	-0.5004	0.5002	2.00
0.81	-0.4862	0.4930	2.03
0.82	-0.4714	0.4853	2.06
0.83	-0.4559	0.4769	2.10
0.84	-0.4397	0.4679	2.14
0.85	-0.4227	0.4581	2.18
0.86	-0.4050	0.4475	2.23
0.87	-0.3864	0.4359	2.29
0.88	-0.3669	0.4232	2.36
0.89	-0.3465	0.4093	2.44
0.90	-0.3251	0.3940	2.54
0.91	-0.3025	0.3770	2.65
0.92	-0.2788	0.3580	2.79
0.93	-0.2536	0.3366	2.97
0.94	-0.2270	0.3122	3.20
0.95	-0.1985	0.2842	3.52

Veall ve Zimmermann'ın yapay R^2'si

Veall ve Zimmermann (1994), R^2_{AN}'nin, üst sınırının 1'den çok daha az olduğuna dikkat çekmiştir. Eğer $N_0 / N = 0.5$ olursa, Aldrich ve Nelson ölçütünün maksimum olabilecek değeri 0.581 olur. $N_0 / N = 0.1$ (ya da 0.9) olması durumunda maksimum değeri 0.394'a düşer. Bu durum normalleştirmeyi gerektirebilir. Veall ve Zimmermann, Aldrich ve Nelson'nun ölçüsünü normalleştirerek, gözlenen bağımlı değişken kesikli olsa da, üst sınırı 1 olan R^2_{ANN} ölçüsünü geliştirmişlerdir.

$$R^2_{ANN} = \frac{LRT}{LRT + N} / \frac{LRT^*}{LRT^* + N} \tag{4.33}$$

McFadden'in yapay R^2'si

McFadden (1973)'nın önerdiği yapay R^2 ölçüsü aşağıdaki gibidir.

$$R^2_{MF} = \frac{LRT}{LRT^*} = \frac{(l_M - l_0)}{(l_{Max} - l_0)} = 1 - \frac{l_M}{l_0} \tag{4.34}$$

Eğer l_M l_0'a yakın bir değerse bu ölçü 0'a yaklaşır. Log olabilirlik değeri maksimum olduğunda ($l_M = 0$) ise McFadden'ın ölçüsü 1 değerini alır. Ayrıca, açıklayıcı değişkenlerin katsayılarının (sabit hariç) sıfır olduğu hipotezinin testi için kullanılan ki-kare istatistiğiyle ilişkilidir.

R^2_{MF} STATA (1995) paket programında da yer alan, muhtemelen en yaygın şekilde kullanılan yapay R^2'dir. Eşitlik (4.34)'te bu ölçüt için üç farklı ifadeye yer verilmiştir. İlk ifadede açıklayıcı değişkenlere bağlı olarak log olabilirlikten elde edilen artışın maksimum olabilecek artışa bölümü yer almaktadır. Son ifadede Lojit ve Probit Modellerde $l_{Max} = 0$ olması ile elde edilen eşitlik yer alır. Bu bağlamda bazen "olabilirlik oran indeksi" olarak da isimlendirilmiştir(Veall ve Zimmermann, 1996).

Merkle ve Zimmermann (1992) ve Cameron ve Windmeijer (1993) bu yapay R^2 için ilk iki ifadenin l_{Max} 0 ya da 1 olmak zorunda olmadığı durumlarda kullanılabileceğini belirtmişlerdir. Ayrıca bu araştırmacılar, bilinen R^2'de olduğu

gibi toplam değişimin açıklanan ve açıklanmayan değişimin toplamıyla ifade edilmesine benzer bir şekilde görülebileceğini ifade etmişlerdir. Sapmalara dayanan bu değişim tanımında, toplam değişim yerine $l_{Max} - l_0$, açıklanan kısım olarak $l_M - l_0$ ve açıklanamayan kısım için de $l_{Max} - l_M$ ifadesi düşünülebilir

Maddala'nın yapay R^2'si

Bir diğer benzer ölçüt aşağıda verilmiştir (Maddala, 1983)

$$R_M^2 = 1 - \exp\left(-\frac{LRT}{N}\right) \tag{4.35}$$

Modelin olabilirliği L_M, olasılıklara bağlı bir değer olduğundan en fazla 1 değerini alabilir. Bu durumda L_0 modelde sadece sabit varken elde edilen olabilirlik değeri olmak üzere, $R_{M_{max}}^2 = 1 - (L_0)^{2/N}$ olacaktır. R^2'nin maksimum değeri sıfır modelin olabilirliğine, yani sadece sabit katsayı olduğunda bulunan olabilirliğe bağlı olur. Burada doğrusal regresyondaki sıfır model için kullanılan $\hat{Y} = \bar{Y}$ eşitliğine benzer olarak olasılık değeri verideki birlerin yüzdesine eşit olur (N_1 / N). Burada N_1 1'lerin sayısı, N toplam veri sayısı olmak üzere $R_{M_{max}}^2$, aşağıdaki eşitlikle ifade elde edilir.

$$R_{M_{max}}^2 = 1 - ((N_1 / N)^{N_1}(1 - N_1 / N)^{N - N_1})^{2/N} \tag{4.36}$$

$\alpha = N / (2l_0)$ olarak tanımlanırsa R_{ANN}^2 ile R_M^2 arasındaki bağıntı eşitlik (4.37)'deki gibi yazılabilir:

$$R_{ANN}^2 = \frac{\alpha - 1}{\alpha - R_M^2}.R_M^2 \tag{4.37}$$

Böylece $R_M^2 = 1$ veya $R_M^2 = 0$ ise $R_M^2 = R_{ANN}^2$ ve $0 < R_M^2 < 1$ ise $R_{ANN}^2 > R_M^2$ olur. α artarsa iki uygunluk ölçüsü arasındaki fark küçülecektir (Sönmez, 2006).

Maddala'nın R^2'si, $N_1 / N = 0.5$ olduğunda $R_{M_{max}}^2 = 0.75$ olarak en büyük değerini alır. 1'lerin sayısı az olduğunda ise düşük değerler alacaktır. Bu nedenle bu ölçüt yerine eşitlik (4.58)'deki Nagelkerke R^2'si önerilmiştir.

$$\bar{R}_N^2 = \frac{R_M^2}{R_{M_{Max}}^2} \tag{4.38}$$

Cragg ve Uhler'in yapay R^2'si

Cragg ve Uhler (1970) ise Maddala'nın R^2'si için aşağıdaki şekilde bir normalleştirme önermiştir ve eşitlikte olabilirlik değeri $L_{Max} = 1$'alınmıştır.

$$R_{CU}^2 = \frac{1-\exp(-LRT/N)}{1-\exp(-LRT^*/N)} = \frac{1-(L_0/L_M)^{2/N}}{1-(L_0)^{2/N}} \tag{4.39}$$

Achen'in yapay R^2'si

Achen (1979) yukarıdaki yapay R^2 istatistiklerinin test istatistikleri olmaları bakımından eksik olduklarını ve modelde yer alan sabit hariç tüm katsayıların ya da bir alt kümesinin sıfırdan farklı olup olmadığını belirlemede başarılı olmadıklarını iddia etmektedir (Aldrich ve Nelson, 1984). Aldrich ve Nelson (1984) bu durumun Achen'in iddia ettiği kadar ciddi olmadığını, çünkü olabilirlik oranının bir χ^2 istatistiği olarak, regresyon analizindeki F testi ile eş olduğunu söyler. Achen (1976) asimptotik olarak χ^2 istatistiği ile aynı testi veren bir yapay R^2 önermiştir. Achen'in ölçüsü iki düzeyli bağımlı değişkene dayanan açıklanan varyansı tanımlaması bakımından avantaja sahiptir. İki düzeyli bağımlı bir değişkenin varyansı, p_i, $Y_i = 1$ olma olasılığını göstermek ve $q_i = (1-p_i)$ olmak üzere, $p_i q_i$ 'dir. Probit katsayılarını ve standart normal dağılım özelliklerin kullanarak Achen'in yapay R^2'si aşağıdaki gibi tanımlanır (Hangle ve Mitchell, 1992).

$$R_A^2 = \frac{\frac{1}{N}\sum_i \frac{(\hat{p}_i - \bar{Y})^2}{\hat{p}_i \hat{q}_i}}{1 + \frac{1}{N}\sum_i \frac{(\hat{p}_i - \bar{Y})^2}{\hat{p}_i \hat{q}_i}} \tag{4.40}$$

McKelvey ve Zavoina'nın yapay R^2'si

McKelvey ve Zavoina gerçek R^2'nin tahminini veren bir yapay R^2 önermişlerdir. McKelvey ve Zavoina'nın R^2'sinde pay kısmında yer alan ifade gizli

değişkenin beklenen değeri ile tahmin edilen değeri arasındaki farkın kareleri toplamıdır. $N\hat{\sigma}^2$ ifadesi açıklanamayan varyanstır ve bu nedenle ölçüt kareler toplamının açıklanan tahmininin açıklanan ve açıklanamayan kareler toplamının tahminine bölümü olarak ifade edilebilir. Dikkat edilmesi gereken nokta kesikli modellerde $\hat{\sigma}^2$'nin tahmin edilemediğidir. Bu nedenle normalleştirme yapılması gerekir. Örneğin bu değer iki düzeyli Lojitte $\pi^2/3$, Probitte 1'dir (Veall ve Zimmermann, 1996).

$$R^2_{MZ} = \frac{\sum_{i=1}^{N}(\hat{Y}_i^* - \bar{Y}^*)^2}{\sum_{i=1}^{N}(\hat{Y}_i^* - \bar{Y}^*)^2 + N\hat{\sigma}^2} \tag{4.41}$$

Aynı zamanda, tahmin edilen Probit ve Lojit katsayıları, gizli bağımlı değişken için tahmin edilen değerlerin varyansı bulunarak, açıklanan varyansı hesaplamada kullanılabilir. Bu $\text{var}(\hat{Y}_i)$ ile gösterilir. Açıklanamayan varyans birim varyansa sahipse, toplam varyans sonrasında açıklanan varyansa bir eklenmesiyle elde edilen değere bölünür ve McKelvey ve Zavoina yapay R^2'si aşağıdaki gibi ifade edilir (Hangle ve Mitchell, 1992). Aynı zamanda bu değer $\hat{\sigma}^2 = 1$ olarak alındığından Probit Model için kullanılan ölçüte eş değerdir.

$$R^2_{MZ} = \frac{\text{var}(\hat{Y}_i)}{1 + \text{var}(\hat{Y}_i)} \tag{4.42}$$

R^2_{AN} gibi olabilirlikleri temel alan yapay R^2'lerin değerlerinde, gerçek R^2'de olduğu gibi, modele anlamsız açıklayıcı değişkenlerin eklenmesiyle azalma görülmez. Bunun sebebi bu ölçülerin log-olabilirlik değerlerini kullanıyor olmalarıdır. Anlamsız değişkenler modele eklendiğinde l_M değeri artmaz. l_M'nin tersine, tahmin edilen $y_i = 1$ olma olasılığı (örneğin Achen'in ölçüsü olasılıkları temel alır) ve $\sum \beta_k x_{ik}$ (örneğin McKelvey-Zavoina'nın ölçüsü bu ifadeyi temel alır) modele değişken eklenmesine bağlı olarak artacak ya da azalacaktır (Hagle ve Mitchell 1992).

Pseudo $\boldsymbol{R^2}$

Bir diğer ölçüt olarak aşağıdaki eşitlik kullanılabilir.

$$R^2_{PSD} = \frac{e^{\frac{2}{N}l_M} - e^{\frac{2}{N}l_0}}{1 - e^{\frac{2}{N}l_0}} \tag{4.44}$$

Literatürde farklı R^2 ölçütlerinin karşılaştırılması ile ilgili çeşitli çalışmalar yapılmıştır. Yapılan bu çalışmalarda Veall ve Zimmerman (1994), ölçütlerden elde edilen değerler bakımından örneklem büyüklüğünün 200 ya da 1000 olması arasında çok az bir fark olduğu sonucuna varmıştır. Hagle ve Mitchell (1992) ise 100 örneklem büyüklüğünü kullanarak R^2_{MZ}'nin örnek varyansının R^2_{ANN}'nin örnek varyansından biraz daha büyük olduğunu bulmuştur.

Windmeijer (1995) R^2_{MZ} ve R^2_{MF} gizli modeldeki sabit değerinde meydana gelen değişime duyarsız olan ölçüler olduğunu bulmuş ve R^2_{MZ} skorlarını gerçek ve tahmini olasılıkların kareli korelasyonuna yakınlığı kriterini kullanarak en iyi ölçüt olarak değerlendirmiştir.

Veall ve Zimmerman, R^2_{AN}, R^2_{MF}'leri ile korelasyon katsayısı ölçütlerini diğer yapay R^2'lerden daha küçük olma eğilimli olduklarını belirtmiştirler. Gerçek R^2'ye en yakın ölçütün R^2_{MZ}'si ve normalleştirilmiş Aldrich ve Nelson R^2'si olduğunu ve gerçek R^2 ile yapay R^2'ler arasında kübik bir ilişkinin uygun olabileceğini ifade etmişlerdir.

Hagle ve Mitchell (1992), tahmin yönteminin hataların olasılık dağılımıyla eşleşmemesi durumunu da dikkate almışlardır ve gerçek hatalar normal olduğu halde iki düzeyli Probit yerine iki düzeyli Lojit Model uygulanırsa, yapay R^2 performanslarında hemen hemen hiçbir farklılık olmayacağını ifade etmişlerdir. Diğer taraftan, gerçek hatalar çarpık ya da iki modlu olduğu durumda Lojit ya da Probit analizin uygulanmasının R^2 ölçülterinin büyük olmasına sebep olacağını, R^2_{MF}'in tercih edilmesi yönünde sonuçlar elde edildiğini belirtmişlerdir

Windmeijer (1995) ayrıca açıklayıcı değişkenleri modele birer birer ekleyerek ölçütleri hesaplamış ve değişkenleri ekledikçe R^2_{MZ}'nin gerçek R^2'ye daha yakın olduğunu bulmuştur

4.8.2. Uyum iyiliği için kullanılabilecek diğer ölçütler

Olabilirlik Oran İstatistiği ve Sapma Ölçütü

Yapay R^2'lerden farklı olarak, olabilirlik oran istatistiği model üzerindeki gelişmenin istatistiksel anlamlılığının bir ölçüsüdür. Genellikle burada sözü geçen model bağımlı değişkeni sadece sabit ile açıklayan boş modeldir. Ayrıca log olabilirlik oran değeri aynı veri seti için kullanılan modellerin karşılaştırılmasında kullanılabilir (Aldrich ve Nelson 1984). Olabilirlik oran istatistiği eşitlik (4.30)'da ifade edilmiştir.

Log olabilirlik oranı, R^2'lerin aksine, modelin iki düzeyli bağımlı değişkendeki değişimi açıklama kabiliyetini doğrudan vermez. Aslında iki düzeyli bağımlı değişkenin dağılımı daha çarpık olduğunda değişimi daha iyi açıklıyor olmasına rağmen olabilirlik oran kriterine göre istatistiksel anlamda anlamsız gözüküyor olabilir. Çünkü böyle bir dağılım için log olabilirlik artışı çok küçüktür (Hagle ve Mitchell 1992).

Olabilirlik oran istatistiği belirli bir olasılık dağılımı ile ilişkilendirilebilir. Hipotez modelin log olabilirliğinden boş modelin log olabilirliğinin çıkarılıp 2 ile çarpılmasından elde edilen ifade asimptotik olarak ki-kare dağılımı gösterir (Nerlove and Press, 1973; McFadden, 1973). Belirlenen anlamlılık düzeyinde olabilirlik oranı test istatistiği k (sabit terimi de içermeyen parametre sayısı) serbestlik derecesine sahip ki-kare tablo değerinden büyük olduğunda, sabit terim hariç tüm katsayıların sıfır olması hipotezinin reddine karar verilir. Bu istatistik doğrusal regresyonda elde edilen katsayıların anlamlılığı testinde kullanılan k ve $N-k+1$ serbestlik dereceli F istatistiğine karşılık gelir.

Tatlıdil (1996), tahmin edilen modele ilişkin olabilirliği ve doygun modelin olabilirliğinin kullanıldığı olabilirlik oran istatistiğini sapma ölçütü olarak dikkate almıştır. Sapmaların doğrusal regresyondaki artık kareler toplamına karşılık geldiğini ifade etmiştir. Sapma ölçütü tahmin edilen modele ilişkin olabilirliği ve doygun

modelin olabilirliğini dikkate alırken, olabilirlik oran istatistiği tahmin edilen modele ilişkin olabilirliği ve boş modelin olabilirliğini kullanır. Kurulan modelin olabilirliğini test etmek ve modele gerçek açıklayıcı değişkenleri belirlemek için de sapma (G^2) yaklaşımı kullanılabilir. Tam doygun model, önemli olduğu düşünülen değişkenleri içeren model olmak üzere sapma ölçütü aşağıda verilmiştir.

$$G^2 = -2\log\left(\frac{\textit{Tahmin edilen modelin olabilirliği}}{\textit{Doygun modelin olabilirliği}}\right) \tag{4.45}$$

Bu istatistik, bir değişkenin modele dahil edilip edilemeyeceğine karar vermek için kullanılabilir. Fakat, bu istatistiğin modelin kalitesini ölçmede kullanılmasının sakıncalı olduğu belirtilmektedir (Uçar, 2004).

Sapmalar ile tahmin edilen modelin olabilecek en iyi modelden ne kadar saptığı ölçülmüş olur. Herhangi iki modelin karşılaştırılması için kullanıldığında sapması küçük olan model tercih edilir.

Tekrarlanmış veriler için bu istatistik m_i açıklayıcı değişkenlerin *i.* bileşenin tekrar sayısı ve $\hat{Y}_i = m_i \hat{p}_i$ Y_i'nin tahmini değerini ya da model için beklenen başarı sayısını göstermek üzere, binomial modeller için sapma aşağıdaki şekilde yazılabilir:

$$G^2 = 2\sum_{i}^{n}\left[Y_i \log\frac{Y_i}{\hat{Y}_i} + (m_i - y_i)\log\left(\frac{m_i - Y_i}{m_i - \hat{Y}_i}\right)\right] \tag{4.46}$$

Bu ifade Y_i'nin tahmini değerleri ile gözlenen değerleri arasında nasıl bir sapma olduğunun karşılaştırılmasını gösterir. Bu istatistiğin doğru sonuçlar verebilmesi için tekrar sayısının oldukça büyük olması gerekmektedir.

Doğrusal modellerde sapma, artıkların kareler toplamıdır ve bilinen bir dağılıma sahiptir. Binomial veri için kullanılan modeller ise asimptotik özelliklere dayanırlar. Gruplanmış veriler için grup büyüklüğü kabul edilebilecek kadar büyük olduğunda, sapma yaklaşık olarak ki-kare dağılım gösterir. Malesef, grup büyüklüğünün ne kadar büyük olması gerektiğine dair kesin doğru bir yargı ya da kural yoktur. Verilen bir tabloda frekans değerleri 5'in altında olan hücre sayısı toplam hücre sayısının %20'sinden fazla olmamalıdır. Bu durumda yaklaşık olarak ki-kare dağılan sapma istatistiğinin serbestlik derecesi, hücre sayısı eksi modeldeki parametre sayısına

eşittir. Eğer hesaplanan G^2 değeri ki-kare tablo değerinden büyükse modelin veriye iyi uyum sağlamadığı söylenir. Genel olarak, eğer mevcut modelin sapması yaklaşık olarak serbestlik derecesine eşitse uyumun tatmin edici olduğu söylenebilir. ($G^2 \approx DF$ ya da $G^2 / DF \approx 1$). Sapma iyi bir uyum istatistiği olarak, her hücrede yeterli gözlem sayısı olduğu zaman kullanılmalıdır (Powers ve Xie, 1999).

İki düzeyli veride, doygun modelin olabilirliği 1'e eşittir ($L_{Max} = 1$ ve $l_{Max} = 0$). Böylece, sapma formülü; $G^2 = -2l_0$ olur. L_0 (ya da $-2l_0$) uyumun bir ölçüsü olarak kullanılmamalıdır. Birey düzeyli veriler için modellerin sapmaları, iç içe geçmiş modeller kümesinden en iyi uyumu sağlayan modeli seçmek için kullanılabilir (Powers, Xie, 1999).

Belirli bir açıklayıcı değişken için olabilirlik oran istatistiği, iki sapma arasındaki fark olarak da ifade edilebilir. Olabilirlik oran istatistiği değişkenin varlığında ve yokluğunda hesaplanır ve sapma değerleri elde edilir. Örneğin bir açıklayıcı değişkenin varlığındaki sapma G_1^2 ve olmadığı durumdaki sapma G_0^2 ise bunların arasındaki $G_1^2 - G_0^2$ farkı hesaplanır ve bu fark değişkenin modelde olduğu durumdaki serbestlik derecesi k_1 ile yer almadığındaki serbestlik derecesi k_2 arasındaki fark ile ki-kare dağılır.

Sapmalar ki-kare dağılımı göstermesede sapmalar arası fark ki-kare dağılımı gösterir (Uçar, 2004).

Doğru Tahminlerin Oranı

Burada, n_c, doğru tahminlerin sayısını, n_1, değeri 1 olan gözlemlerin sayısını, n_0 0 olarak gözlenen veri sayısını ve $n_m = \max(n_1, n_0)$ değerini göstermektedir (Sönmez, 2006).

$$PCP = \frac{n_c - n_m}{N - n_m} \tag{4.47}$$

Hata kareler toplamı

$P(Y_i = 1) = F(x_i'\beta) = F_i$ olmak üzere hata kareler toplamı aşağıdaki gibidir.

$$SSR = \sum_{i=1}^{N} (Y_i - \hat{F}_i)^2 \tag{4.48}$$

Ağırlıklı artık kareler toplamı

Ağırlıklı artık kareler toplamının asimptotik olarak χ^2 dağıldığı varsayılır. Ağırlıklı artık kareler toplamının asimptotik dağılımının normal dağılım gösterdiği ispatlanmıştır ve bu ölçüt eşitlik (4.53)'de ki gibi hesaplanır (Özarıcı, 1996).
Eşitlik (4.53) her bir gözleme ait Pearson artıklarının (eşitlik (4.49)) kareleri toplamına eşittir.

$$r_i = \frac{(Y_i - \hat{F}_i)}{\sqrt{\hat{F}_i(1-\hat{F}_i)}} \tag{4.49}$$

$$WSSR = \sum_{i=1}^{N} \frac{(Y_i - \hat{F}_i)^2}{\hat{F}_i(1-\hat{F})} \tag{4.50}$$

Buse'un R^2'si

Modelde sadece sabit katsayı olduğunda modelden elde edilen ağırlıklı hata kareler ortalaması $WSSR_0$ olmak üzere Buse'un R^2 aşağıdaki gibi tanımlanır.

$$R_B^2 = \frac{WSSR_0 - WSSR}{WSSR_0} \tag{4.51}$$

Akaike Bilgi Kriteri

Ayrıca olabilirliği kullanan bir diğer ölçüt olan Akaike'nin bilgi kriteri de uyum iyiliği ölçütü olarak kullanılabilir (Sönmez, 2006).

$$AIC = -2l_M / N + 2(k/N) \tag{4.52}$$

Doğru Sınıflama Oranı

Bu ölçüt gerçek gözlem değerleri ile tahmini değerlerinin doğru şekilde sınıflanmasının oranını verir.

Tablo 4.3 Sınıflandırma tablosu

	Tahmini Değerler		
Gözlenen Değerler		0	1
	0	*a*	*b*
	1	*c*	*d*

$$DSO = \frac{a+d}{a+b+c+d} \tag{4.53}$$

Yule's Q Ölçütü

Bu ölçüt araştırma sonuçlarına ilişkin oddslar farkının oddslar toplamına oranıdır.

$$Q_{Yule} = \frac{\frac{a}{b} - \frac{c}{d}}{\frac{a}{b} + \frac{c}{d}} = \frac{ad - bc}{ad + bc} \tag{4.54}$$

Q_{Yule} ölçütü, gerçek olasılıklarla tahmin edilen olasılıkların birbirlerine uygunluğunu göstermek için kullanılan bir uyum iyiliği ölçütüdür. -1 ile 1 arasında değer almaktadır ve Q_{Yule} değerinin 1'e yaklaşması gerçek değerler ile tahmini değerler arasında pozitif yönlü güçlü bir ilişki olduğunu belirtir diğer taraftan -1 olması negatif yönlü bir ilişkiyi ifade eder. 0 değerini alması ise gerçek değerler ve tahmin edilen değerler arasında ilişkinin yokluğunun göstergesi olarak kabul edilir. Kullanılan modelin tahmin etmedeki gücünün ifadesi olarak kullanıldığında, elde edilen değerin 1'e yakın değerler alması istenir.

B ***indeksi***

B indeksi tahminlerin doğruluğunu değerlendirmede kullanılabilir. *B* indeksi tahmini değer ve gerçek değerler arasındaki farkların karelerinin ortalamasın bir ölçüsüdür. B indeksi 0 ile 1 aralığında değerler alır. *B* indeksinin 1 değeri alması "mükemmel tahmin"e atfedilir. Örneklem büyüklükleri eşit iken, rastgele tahmin (bağımlı değişkenin gözlenen sınıfı ile tahmin edilen sınıfı arasında ilişki olmaması) durumunda, *B* indeksinin değeri 0.75 civarındadır.

$$B = 1 - \sum_i (P_i - Y_i)^2 / N \tag{4.55}$$

Q ***indeksi***

Q indeksi de *B* indeksine benzer br ölçüttür. *Q* indeksinin 1 değerini göstermesi "mükemmel tahmin"e, 0 ve daha küçük değer alması "rastgele tahmin"e atfedilir.

$$Q = \sum_i [1 + \log_2 (P_i^{Y_i} (1 - P_i)^{1-Y_i})] / N \quad (4.56)$$

5. UYGULAMA

İki düzeyli bağımlı değişkenli verilerin analizinde Lojit ve Probit Modeller yaygın olarak kullanılan modellerdir. Her iki modelden elde edilen tahmini olasılık değerleri oldukça yakındır. Aralarındaki farklılık; kullandıkları fonksiyonlardan ve hatanın dağılımına ilişkin kabul edilen varsayımlardan meydana gelmektedir. İki model de aynı verinin analizi için kullanılabilmelerine rağmen, literatürde hangisinin daha iyi bir model olduğu ya da hangi şartlarda birinin diğerine daha üstün olacağı konusunda teorik ya da uygulamaya dayalı kesin sonuçların yer aldığı bir çalışma bulunmamaktadır.

Bu çalışmada, farklı koşullar dikkate alınarak, belirli kriterlerle hangi modelin daha iyi sonuçlar verdiğine dair bilgi edinilmesi amacıyla bir simülasyon çalışması gerçekleştirilmiştir. Bu bölümde, öncelikle simülasyon çalışmasında kullanılan materyal ve yöntemler üzerinde durulmuş, sonrasında veri üretiminden başlayarak tüm simülasyon adımları ayrıntılı olarak anlatılmıştır. Son olarak, çalışmadan elde edilen sonuçlar yorumlanmış ve değerlendirilmiştir.

5.1. Materyal ve Yöntem

Çalışmada, iki düzeyli Lojit ve Probit Modellerin karşılaştırılması amacıyla, gizli değişken yaklaşımından faydalanılmıştır. Gizli değişkenin standart normal dağılıma sahip olduğu ve bu sürekli gizli değişkeni etkileyen üç açıklayıcı değişkenin söz konusu olduğu varsayılmıştır. Bu varsayımlar dikkate alınarak uygun veri üretimi ve sonrasında yapılan analizler için macrolar yardımıyla program kodları yazılmış ve MINITAB paket programında çalıştırılmıştır.

Veri üretimi üç farklı varyans kovaryans matrisi dikkate alınarak gerçekleştirilmiştir. Bu matrislerin belirlenmesi aşamasında, varyans kovaryans matrisinin pozitif tanımlı olması gerektiği dikkate alınarak, bağımlı değişken ile açıklayıcı değişkenler arasında farklı ilişki büyüklüklerini oluşturacak özel değerler verilmiştir. Varyans kovaryans matrisleri ilişki büyüklükleri dikkate alınarak “çok”, “orta” ve “yok” olarak sınıflandırılmıştır. Ayrıca, bu matrislerde açıklayıcı değişkenler arasındaki ilişkiler sıfır olarak alındığından, çoklu bağlantı sorunu ile karşılaşılmamıştır. Veri üretimi aşamasında, kullanılan varyans kovaryans matrisleri aşağıdaki gibidir:

$$\sum_{çok} = \begin{bmatrix} 1 & 0.4 & 0.5 & -0.7 \\ 0.4 & 1 & 0 & 0 \\ 0.5 & 0 & 1 & 0 \\ -0.7 & 0 & 0 & 1 \end{bmatrix}$$

$$\sum_{orta} = \begin{bmatrix} 1 & 0.4 & 0.2 & -0.3 \\ 0.4 & 1 & 0 & 0 \\ 0.2 & 0 & 1 & 0 \\ -0.3 & 0 & 0 & 1 \end{bmatrix}$$

$$\sum_{yok} = \begin{bmatrix} 1 & 0.01 & 0.1 & -0.1 \\ 0.01 & 1 & 0 & 0 \\ 0.1 & 0 & 1 & 0 \\ -0.1 & 0 & 0 & 1 \end{bmatrix}$$

Veri üretimleri her bir varyans kovaryans matirisi için, ortalama vektörü $\mu' = [0,0,0,0]$ olan dört değişkenli normal dağılımdan; biri bağımlı, üçü açıklayıcı değişken olmak üzere gerçekleştirilmiştir.

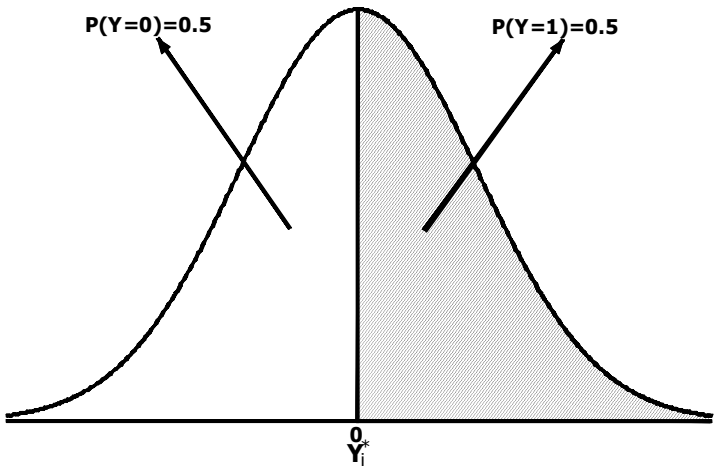
Örneklem büyüklüğünün kullanılan modellerin iyiliğine olan etkilerini ölçmek amacıyla, 5 farklı örneklem büyüklüğü dikkate alınmıştır. Bunlar yeterince küçük bir örneklem büyüklüğü olan 40, orta örneklem büyüklüğü 100 ve 200 ve son olarak 500 ve 1000 örneklem büyüklükleri için üretim gerçekleştirilmiştir. Yapılan ön çalışmalarda örneklem büyüklüğünün 1000'den büyük olması, modellerin iyiliği üzerinde fark yaratmadığından en çok 1000 örneklem büyüklüğü kullanılması uygun bulunmuştur.

Her bir varyans kovaryans matrisi ve örneklem büyüklüğü için gerçekleştirilen üretim 1000'er kez tekrarlanmıştır. Tekrar sayısı, kullanılan programın çalışma hızını etkilemesinden ve modellerin iteratif yöntemleri kullanmasından kaynaklanan hataların giderilmesinde yaşanılan zorluklardan dolayı daha fazla arttırılmamakla beraber, yeterli olduğu düşünülmüştür. Sözü geçen hatalar, parametre tahminlerinde karşılaşılan iteratif adımlarda elde edilen değerlerin yeterince yakınsayamamasından kaynaklanmıştır. Bu hatalı üretimler, analize dahil edilmeyip hatalı sonuçlar elde edilme durumu ortadan kaldırılmıştır.

Üretilen sürekli gizli değişkenin iki düzeyli olarak sınıflandırılması aşamasında ise iki farklı kesim noktası kullanılmıştır. Kullanılan iki farklı kesim noktası sayesinde, bağımlı değişkenin 1 olarak gözlenme olasılıklarındaki farklılık dikkate alınmış olur. İlk eşik noktası, 0 olarak belirlenmiştir. 0 noktası $P(Y=1)$ ve $P(Y=0)$ olasılıklarının 0.5 olmasını sağlayan standart normal dağılımın z değeridir. Bu ifade, Y_i^* $i.$ gizli değişen değerini ve Y_i sınıflandırılmış iki düzeyli bağımlı değişkenin $i.$ değerini göstermek üzere, aşağıdaki fonksiyon ile verilebilir.

$$Y_i = \begin{cases} 1 & Y_i^* > 0 \\ 0 & Y_i^* \leq 0 \end{cases}$$

Şekil 5.1' de ise, eşik noktası 0 olarak alınıp üretilen verilere uygulandığında elde edilen 1 ve 0 gözlenme olasılıklarını dağılım eğrisi yardımıyla göstermektedir. Şekilde eğrinin altında kalan taralı alan, $P(Y=1)$ olasılığını ifade etmektedir.

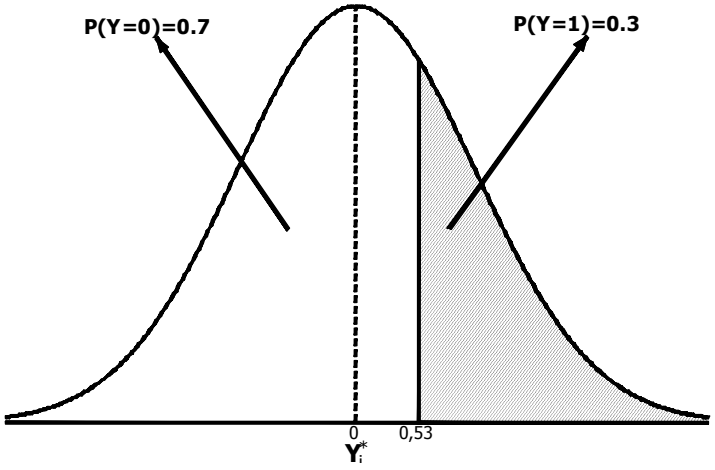


Şekil 5.1 Eşik noktası 0 iken dağılım eğrisi üzerinde olasılık değerleri

Diğer bir eşik noktası 0.53 olarak belirlenmiştir. Bu değer $P(Y=1)=0.3$ ve $P(Y=0)=0.7$ olasılıklarına karşılık gelen z değeridir ve benzer şekilde aşağıdaki fonksiyonla ifade edilir.

$$Y_i = \begin{cases} 1 & Y_i^* > 0.53 \\ 0 & Y_i^* \leq 0.53 \end{cases}$$

Şekil 5.2, eşik noktası 0.53 olarak alınıp üretilen verilere uygulandığında elde edilen 1 ve 0 gözlenme olasılıklarını göstermektedir. Şekilde eğrinin altında kalan taralı alan $P(Y=1)$ olasılığını ifade etmektedir.



Şekil 5.2 Eşik noktası 0.53 iken dağılım eğrisi üzerinde olasılık değerleri

Aşağıdaki tablo, veri üretim aşamasının hangi koşullar dikkate alınarak yapıldığının özetlemektedir. Çalışmada 30 farklı üretim gerçekleştirilmiş ve 1000'er tekrar olmak üzere toplamda 30000 veri üretilmiştir.

Tablo 5.1 Veri üretimi ve sınıflandırılması

<table>
<tr><td colspan="5">Veriler</td></tr>
<tr><td colspan="2">Kesim Noktası "0" Alınarak</td><td colspan="3">Kesim Noktası "0.53" Alınarak</td></tr>
<tr><td>Örneklem Büyüklükleri</td><td>Tekrar Sayısı</td><td>Varyans-Kovaryans Matrisi</td><td>Örneklem Büyüklükleri</td><td>Tekrar Sayısı</td></tr>
<tr><td rowspan="3">40, 100, 200, 500, 1000</td><td rowspan="3">1000</td><td>Çok</td><td rowspan="3">40, 100, 200, 500, 1000</td><td rowspan="3">1000</td></tr>
<tr><td>Orta</td></tr>
<tr><td>Yok</td></tr>
</table>

Veriler üretildikten ve bağımlı değişken sınıflandırıldıktan sonra, Lojit ve Probit Modeller kullanılarak parametre tahminleri elde edilmiştir. Sonrasında her iki modelin uyum iyiliklerini değerlendirmek amacıyla farklı ölçütler hesaplanmıştır. Dikkate alınan ölçütler; sapmalar, Pearson artıklar, doğru sınıflandırma oranı, Q indeksi, B indeksi, Akaike bilgi kriteri, Aldrich ve Nelson yapay R^2, Veall ve

Zimmerman yapay R^2, McFadden yapay R^2, Maddala yapay R^2, Achen yapay R^2, Pseudo R^2, Cragg ve Uhler yapay R^2, Lave yapay R^2, McKelvey ve Zavonia yapay R^2 Buse yapay R^2 ve Kareli korelasyon katsayısıdır.

Varyans kovaryans matrisi, örneklem büyüklüğü ve kesim noktası dikkate alınarak elde edilen modellere ait farklı ölçütlere ilişkin 1000'er adet değerin ortalaması alınarak iki model karşılaştırılmıştır. Bu karşılaştırmada her iki modelden elde edilen farklı ölçütlerin ortalama değerlerinde anlamlı bir farklılığın olup olmadığının testinde t istatistiği kullanılmıştır

Bağımlı değişkenin, gerçekte sürekli gözlenemeyen bir değişken olduğu varsayılmıştır. Normal dağılıma sahip bir bağımlı değişken ve bağımlı değişkeni etkileyen üç bağımsız değişken üretilmiştir. Üretim ve diğer analiz aşamaları için yazılan kodlar MINITAB paket programında çalıştırılmıştır. Üretilen sürekli bağımlı değişken iki farklı eşik değerine göre 0 ve 1 olmak üzere iki düzeyli olarak sınıflandırılmıştır. Sonrasında, iki düzeyli bağımlı değişkene Lojit ve Probit Modeller uygulanarak parametre ve olasılık tahminleri elde edilmiştir. Son aşamada ise, Lojit ve Probit Modellere ait sapmalar, artıklar, farklı yapay R^2 değerleri ve diğer kriterler dikkate alınrak iki model karşılaştırılmıştır. Modellere ait farklı koşullarda (örneklem büyüklüğü, varyans kovaryans matrisi, kesim noktası) elde edilen ölçüt değerlerinin birbirinden anlamlı olarak farklılığının testi için t testi kullanılmıştır. Ayrıca, gerçek sürekli bağımlı değişkenden elde edilen R^2 değeri, yapay R^2 değerleri ile karşılaştırılabilmesi için hesaplanmıştır.

5.2. Simülasyon Sonuçları

Kesim noktasının "0" olma durumu

Tablo 5.2'de kesim noktası 0 olarak alınarak, yani $P(Y=1)=P(Y=0)=0.5$ olacak şekilde bağımlı değişkenin sınıflandırılması ile elde edilen verilere Lojit ve Probit Modellerin uygulanmasından elde edilen uyum iyiliği ölçütleri yer almaktadır. Bu tabloda her bir ilişki düzeyi için ve tüm örneklem büyüklükleri için elde edilen sapmalar ve artık değerleri yer almaktadır. Tabloda N örneklem büyüklüğünü, sapma ve artıkların başında bulunan L ve P harfleri sırasıyla elde edilen değerlerin Lojit ve Probit Modellerden olduğunu ifade etmektedir. Tabloda her bir hücrede 1000 adet ölçüt değerinden elde edilen ortalama değer yer almaktadır. Bir modelin diğeriyle

karşılaştırılmasında sapma ya da artıklar kriter olarak ele alındığında, daha küçük değere sahip olan model iyi model olacaktır. Tablo 5.2'de daha iyi modele ilişkin değerler koyu renkle belirtilmiştir. Her iki modele ilişkin sapma değerleri dikkate alındığında, "çok" ve "orta" ilişki düzeyi için Probitten elde edilen değerler daha iyi gözükmesine rağmen, iki modelden elde edilen değerlerin birbirine çok yakın olduğu görülür. İki modele ilişkin sapma ölçüt değerleri dikkate alınarak, aralarındaki farkın anlamlı olup olmadığının belirlenmesi için yapılan t testine göre de iki model arasındaki farklılık anlamlı bulunmamıştır.

Tablo 5.2'de yer alan bir diğer ölçüt artıklardır. İlişki düzeyi "çok" ve örneklem büyüklükleri 40 ile 100 olduğunda Probit Modelin Lojit Modelden daha iyi sonuç verdiği ve aralarındaki farkın anlamlı olduğu test edilmiştir. Yine aynı ilişki düzeyinde örneklem büyüklüğü 500 ve 1000 için tersi sonuçlar söz konusu olmuştur. Bu örneklem büyüklüklerinde Lojit Model anlamlı derecede Probit Modelden daha iyidir. Örneklem büyüklüğü 200 olduğunda ise aralarındaki fark anlamlı bulunmamıştır. İlişki düzeyinin "orta" olması durumunda ise örneklem büyüklüğü 40 için Probit Model veriyi daha iyi modellemiştir. 500 ve 1000 örneklem büyüklüklerinde Lojit Model anlamlı derecede Probit Modelden farklılık göstermiştir. Bu örneklem büyüklüklerinde Lojit Modelin daha iyi tahminler elde ettiği söylenebilir.

Açıklayıcı değişkenler ile bağımlı değişken arasında anlamlı bir ilişki olmadığında kullanılan modelin doğru sonuçlar vermemesi beklenir ve dolayısıyla herhangi bir uyum iyiliği ölçütü bakımından kötü değerlere sahip olmalıdır. Bu durum modelin anlamlı olmayan ilişkiyi ifade edebilme yeteneği olarak düşünülebilir. Bu yüzden modelin uyumunun kötü olduğu bilindiğinde sapma ve artık değerlerine göre, büyük değerli olan modelin, değişkenler arasındaki ilişkinin anlamsızlığını göstermesi bakımından daha iyi olduğu düşünülür. İlişki düzeyi "yok" iken sapmalar, örneklem büyüklüğü 40 olduğu durum hariç aynı değerleri vermiştir. Bu açıdan bu ilişki düzeyinde de sapmalar açısından iki modele ilişkin anlamlı bir fark yoktur. Artıklar dikkate alındığında ise; 40, 100 ve 200 örneklem büyüklüğünde Lojit Model daha iyi değerler almış ama yinede farklılık anlamlı bulunmamıştır. Örneklem büyüklüğü 500 ve 1000 olduğunda ise değerlerde farklılığa rastlanmamıştır.

Tablo 5.2 Kesim noktası 0 olduğunda modellere ilişkin sapma ve artık değerleri

	N	***L_Sapma***	***P_Sapma***	***L_Artık***	***P_Artık***
ÇOK	40	15.611	15.456	17.859	**16.291**
	100	41.142	40.855	52.457	**48.910**
	200	87.223	86.792	115.820	114.240
	500	224.010	223.160	**304.670**	320.960
	1000	452.240	450.700	**619.170**	676.760
ORTA	40	43.135	43.055	37.820	**37.411**
	100	114.180	114.120	98.487	98.364
	200	232.270	232.220	198.490	198.840
	500	586.050	585.970	**497.600**	499.480
	1000	1177.800	1177.700	**995.080**	999.030
YOK	40	50.239	50.219	39.647	39.548
	100	132.950	132.950	99.953	99.934
	200	270.300	270.330	199.980	199.970
	500	682.620	682.620	499.990	499.990
	1000	1369.100	1369.100	1000.000	1000.000

Tablo 5.3'de DSO, Q indeksi, B indeksi, ve Akaike Bilgi Kriterine ilişkin değerler bulunmaktadır. Tabloda N örneklem büyüklüğünü L Lojit, P Probit Modeli ifade etmektedir. Akaike Bilgi Kriteri dışındaki ölçütler 0 ile 1 arasında değerler alabilen ve 1'e yakın değerlerin iyi uyumu ifade ettiği ölçütlerdir. Kategorik veri analizlerinde kullanılan en temel ve basit ölçülerden biri DSO'dur. Bu ölçüt olasılıklar dikkate alınarak bağımlı değişkene 1 ya da 0 atanmasıyla elde edilen değerlerin gerçek değerlerle karşılaştırılmasını temel aldığından, diğer ölçütlerden farklı olarak olasılıkları doğrudan dikkate almamaktadır. Her iki modele ilişkin DSO, B indeksi ve Akaike Bilgi Kriteri değerleri dikkate alındığında Lojit Modelin "çok" ve "orta" düzeyler için daha iyi değerler aldığı, ancak iki model arasında anlamlı bir farkın olmadığı görülmüştür. "Yok" ilişki düzeyinde ise modelin anlamsızlığının ifadesi küçük değerler ile sağlandığından Probit Modelin daha iyi değerler aldığı ya da iki modelden elde edilen değerlerin farksız olduğu görülür. Q indeksine ilişkin değerler incelendiğinde ise "çok" ve "orta" ilişki düzeyi için Probit Model, ilişkinin "yok" düzeyinde ise Lojit Modelin daha iyi olması yönünde farklılık göstermiştir. Ancak tablodaki tüm kriterler açısından iki model arasında anlamlı bir farklılık görülmemiştir.

Tablo 5.3 Uyum iyiliği için kullanılan bazı ölçütlerin kesim noktası 0 iken modellerden elde edilen değerleri

	N	***L_DSO***	***P_DSO***	***L_Q İndeksi***	***P_Q İndeksi***	***L_B İndeksi***	***P_B İndeksi***	***L_Akaike***	***P_Akaike***
ÇOK	40	0.90947	0.90607	0.71847	0.72128	0.93709	0.93645	0.54028	0.53639
	100	0.90521	0.90378	0.70322	0.70529	0.93456	0.93417	0.47142	0.46855
	200	0.90139	0.90053	0.68541	0.68696	0.93096	0.93076	0.46612	0.46396
	500	0.89905	0.89873	0.67682	0.67805	0.92923	0.92919	0.46003	0.45832
	1000	0.89844	0.89827	0.67378	0.67489	0.92862	0.92863	0.45824	0.45670
ORTA	40	0.71383	0.71158	0.22211	0.22356	0.81616	0.81582	1.22840	1.22640
	100	0.69498	0.69376	0.17636	0.17681	0.80467	0.80452	1.20180	1.20120
	200	0.68770	0.68747	0.16225	0.16245	0.80094	0.80086	1.19140	1.19110
	500	0.68498	0.68470	0.15451	0.15462	0.79884	0.79882	1.18410	1.18390
	1000	0.68272	0.68261	0.15038	0.15050	0.79763	0.79763	1.18380	1.18370
YOK	40	0.63819	0.63623	0.09401	0.09437	0.77991	0.77977	1.40600	1.40550
	100	0.59336	0.59275	0.04097	0.04100	0.76369	0.76366	1.38950	1.38950
	200	0.57159	0.57146	0.02500	0.02501	0.75847	0.75846	1.38160	1.38160
	500	0.55626	0.55613	0.01519	0.01519	0.75519	0.75519	1.37720	1.37720
	1000	0.55115	0.55115	0.01238	0.01238	0.75425	0.75425	1.37510	1.37510

Tablo 5.4'de kesim noktası 0 olarak alınıp bağımlı değişken sınıflandırıldığında, iki düzeyli modellerde uyum iyiliği için kullanılan bazı R^2 ölçütlerine ilişkin Probit ve Lojit Modellerden elde edilen değerlere yer verilmiştir. Bu ölçütler 0 ile 1 arasında değerler alan ve 1'e yakın değerlerin iyi uyumu gösterdiği ölçütlerdir. Tablo genel olarak incelendiğinde Lojit ve Probit Modellerden elde edilen değerlerin birbirlerine oldukça yakın olduğu görülmüştür. Sadece anlamlı farklılıklar McKelvey ve Zavonia'nın, Achen'nin, Buse'nin yapay R^2'lerinde ve kareli korelasyon katsayısında meydana gelmiştir. Aslında bu beklenen bir sonuçtur. Çünkü, diğer farklılık görülmeyen ölçütlerin çoğu, sapmaları temel alan ölçütlerdir. Dolayısıyla, sapmalarda anlamlı bir farklığın oluşmaması bu ölçütlerde de iki model arasında anlamlı bir farklılığın olmamasına sebep olacaktır. Achen'nin ve Buse'nin yapay R^2'sine göre "çok" ve "orta" ilişki düzeyi için tüm örneklem büyüklüklerinde Probit Model'in daha iyi olduğu görülmüştür. "Yok" ilişki düzeyinde ise Lojit Model daha iyidir. McKelvey ve Zavonia'nın yapay R^2'sine göre "çok" ilişki düzeyinde örneklem büyüklüğü 40 ve 100 için anlamlı bir fark olmamıştır. Diğer yandan, "çok" ilişki düzeyi için diğer örneklem büyüklüklerinde ve "orta" ilişki düzeyinde tüm örneklem büyüklüklerinde Probit Model anlamlı derecede daha iyi olarak bulunmuştur. "Yok" düzeyinde ise Lojit Model daha iyidir. Kareli korelasyon ölçütüne göre ise anlamlı bir farklılık sadece ilişki düzeyi "çok" iken 500 ve 1000 örneklem büyüklüklerinde Lojit Modelin daha iyi sonuçlar vermesi yönünde gerçekleşmiştir.

Tablo 5.4 Uyum iyiliği için kullanılan bazı yapay R^2 'lerin kesim noktası 0 iken modellerden elde edilen değerleri

	N	***Gerçek R^2***	***L_AldrichNelson***	***P_AldrichNelson***	***L_DüzeltilmişAN***	***P_DüzeltilmişAN***	***L_VeallZimmerman***	***P_VeallZimmerman***
ÇOK	40	0.90060	0.49026	0.49129	0.84325	0.84502	0.85038	0.85217
	100	0.90236	0.48965	0.49041	0.84220	0.84351	0.84543	0.84674
	200	0.90015	0.48521	0.48578	0.83456	0.83554	0.83648	0.83748
	500	0.90015	0.48325	0.48371	0.83119	0.83198	0.83237	0.83317
	1000	0.89957	0.48255	0.48297	0.82999	0.83071	0.83089	0.83160
ORTA	40	0.33250	0.21124	0.21231	0.36333	0.36517	0.36630	0.36817
	100	0.30984	0.18649	0.18686	0.32076	0.32140	0.32192	0.32255
	200	0.29908	0.17865	0.17884	0.30728	0.30760	0.30797	0.30829
	500	0.29534	0.17438	0.17449	0.29993	0.30012	0.30035	0.30053
	1000	0.29210	0.17144	0.17156	0.29488	0.29508	0.29520	0.29540
YOK	40	0.10026	0.08925	0.08961	0.15351	0.15413	0.15527	0.15590
	100	0.05203	0.04445	0.04448	0.07645	0.07651	0.07674	0.07680
	200	0.03703	0.02874	0.02875	0.04943	0.04945	0.04954	0.04956
	500	0.02595	0.01852	0.01852	0.03186	0.03186	0.03190	0.03191
	1000	0.02327	0.01585	0.01585	0.02726	0.02726	0.02729	0.02729

Tablo 5.4 devam

	N	*Gerçek* R^2	*L_Lave*	*P_Lave*	*L_McFadden*	*P_McFadden*	*L_Maddala*	*P_Maddala*
ÇOK	40	0.90060	0.74221	0.73958	0.71356	0.71642	0.61790	0.61941
	100	0.90236	0.73561	0.73402	0.70107	0.70315	0.61694	0.61805
	200	0.90015	0.72246	0.72164	0.68427	0.68583	0.61039	0.61124
	500	0.90015	0.71632	0.71618	0.67633	0.67757	0.60750	0.60817
	1000	0.89957	0.71420	0.71424	0.67354	0.67465	0.60646	0.60707
ORTA	40	0.33250	0.24684	0.24546	0.20855	0.21003	0.23916	0.24053
	100	0.30984	0.21130	0.21068	0.17075	0.17120	0.20662	0.20707
	200	0.29908	0.19990	0.19957	0.15930	0.15951	0.19628	0.19650
	500	0.29534	0.19384	0.19374	0.15334	0.15346	0.19074	0.19087
	1000	0.29210	0.18973	0.18972	0.14977	0.14989	0.18709	0.18723
YOK	40	0.10026	0.09598	0.09540	0.07632	0.07670	0.09527	0.09568
	100	0.05203	0.04603	0.04593	0.03457	0.03460	0.04592	0.04596
	200	0.03703	0.02935	0.02933	0.02170	0.02171	0.02934	0.02935
	500	0.02595	0.01876	0.01875	0.01372	0.01372	0.01875	0.01876
	1000	0.02327	0.01601	0.01601	0.01167	0.01167	0.01600	0.01600

Tablo 5.4 devam

	N	*Gerçek* R^2	*L_Achen*	*P_Achen*	*L_McKelveyZavonia*	*P_McKelveyZavonia*	*L_KareliKorelasyon*	*P_KareliKorelasyon*
ÇOK	40	0.90060	0.99770	**0.99985**	0.83929	0.84792	0.75049	0.75162
	100	0.90236	0.99929	**1.00000**	0.82240	0.82912	0.74232	0.74096
	200	0.90015	0.99972	**1.00000**	0.80991	**0.81586**	0.72830	0.72526
	500	0.90015	0.99989	**1.00000**	0.80261	**0.80784**	**0.72185**	0.71777
	1000	0.89957	0.99995	**1.00000**	0.80070	**0.80559**	**0.71945**	0.71496
ORTA	40	0.33250	0.42096	**0.48189**	0.28766	**0.31855**	0.25201	0.25391
	100	0.30984	0.33155	**0.37623**	0.22613	**0.25561**	0.21314	0.21301
	200	0.29908	0.30422	**0.33981**	0.21017	**0.23981**	0.20099	0.20040
	500	0.29534	0.28748	**0.31647**	0.19939	**0.22893**	0.19469	0.19390
	1000	0.29210	0.27915	**0.30478**	0.19449	**0.22412**	0.19064	0.18993
YOK	40	0.10026	**0.13238**	0.14479	**0.11358**	0.13907	0.09735	0.09805
	100	0.05203	**0.05294**	0.05381	**0.04820**	0.05942	0.04614	0.04620
	200	0.03703	**0.03203**	0.03220	**0.02732**	0.03452	0.02939	0.02940
	500	0.02595	**0.01974**	0.01978	**0.01561**	0.01983	0.01877	0.01877
	1000	0.02327	**0.01665**	0.01666	**0.01191**	0.01519	0.01601	0.01600

Tablo 5.4 devam

	N	*Gerçek* R^2	*L_Pseudo*	*P_Pseudo*	*L_Buse*	*P_Buse*	*L_CraggUhler*	*P_CraggUhler*
ÇOK	40	0.90060	0.57409	0.57764	0.60022	**0.69346**	0.83088	0.83291
	100	0.90236	0.55348	0.55600	0.53261	**0.63413**	0.82536	0.82685
	200	0.90015	0.53015	0.53200	0.48519	**0.57378**	0.81522	0.81635
	500	0.90015	0.51900	0.52045	0.45920	**0.52181**	0.81056	0.81147
	1000	0.89957	0.51518	0.51647	0.45077	**0.49613**	0.80889	0.80970
ORTA	40	0.33250	0.11853	0.11957	0.15377	**0.29623**	0.32149	0.32334
	100	0.30984	0.09110	0.09138	0.12289	**0.26452**	0.27634	0.27695
	200	0.29908	0.08325	0.08338	0.11834	**0.25847**	0.26213	0.26243
	500	0.29534	0.07931	0.07937	0.11735	**0.25613**	0.25448	0.25465
	1000	0.29210	0.07710	0.07717	0.11710	**0.25610**	0.24954	0.24973
YOK	40	0.10026	0.03929	0.03952	**0.11509**	0.25661	0.12856	0.12913
	100	0.05203	0.01667	0.01669	**0.10999**	0.25287	0.06143	0.06149
	200	0.03703	0.01028	0.01029	**0.11167**	0.25425	0.03918	0.03919
	500	0.02595	0.00643	0.00643	**0.11302**	0.25529	0.02502	0.02503
	1000	0.02327	0.00545	0.00545	**0.11274**	0.25537	0.02135	0.02134

Kesim noktasının "0.53" olma durumu

Kesim noktası 0.53, standart normal dağılım eğrisinin altında kalan alanın 0.3'e eşit olduğu z değeridir. Bu değer dikkate alınarak bağımlı gizli değişken, $P(Y=1)=0.3$ ve $P(Y=0)=0.7$ olacak şekilde iki düzeyli olarak sınıflandırılmış olur. İki düzeyli olarak sınıflandırılan bağımlı değişken ve açıklayıcı değişkenler Lojit ve Probit Modeller kullanarak modellendikten sonra, bu modellere ilişkin uyum iyiliği ölçütleri hesaplanmıştır. Bu ölçütlerden sapma ve artıklar kullanılarak elde edilen değerler Tablo 5.5'te yer almaktadır. Tabloda N örneklem büyüklüğünü, sapma ve artıkların başında bulunan L ve P harfleri sırasıyla elde edilen değerlerin Lojit ve Probit Modellerden olduğunu ifade etmektedir. Bu tablonun her bir hücresinde yer alan değer 1000'er adet ölçüt değerinin ortalamasıdır. Tablo incelendiğinde, genel olarak kesim noktası 0 iken elde edilen sonuçlar ile paralel sonuçların elde edildiği görülmüştür. Sapmalara göre anlamlı bir farklılık oluşmazken, artıklar dikkate alındığında ilişki düzeyi "çok" için 40, 100, 200 örneklem büyüklüklerinde Probit Model iyiyken, 500 ve 1000 olduğunda Lojit Model daha iyi olmuştur. "Orta" düzey ilişki için 40 örneklem büyüklüğünde Probit Model'in iyi olduğu yönünde anlamlı bir fark varken, 100 örneklem büyüklüğünde fark anlamsızdır. Büyük örneklemler için ise Lojit Model daha iyi sonuçlar vermiştir. "Yok" düzeyinde sadece 40 ve 1000 örneklem büyüklüğünde, ilişkisizliği ifade etmesi bakımından Lojit Modelin tercih edilebilir olması yönünde, anlamlı bir farklılık bulunmuştur.

Tablo 5.5 Kesim noktası 0.53 olduğunda modellere ilişkin sapma ve artık değerleri

	N	***L_Sapma***	***P_Sapma***	***L_Artık***	***P_Artık***
ÇOK	40	13.867	13.706	15.702	**14.327**
	100	35.098	34.867	46.683	**42.726**
	200	75.057	74.662	102.250	**99.450**
	500	194.770	193.950	**271.120**	283.320
	1000	391.920	390.330	**547.280**	598.820
ORTA	40	36.784	36.725	37.587	**36.173**
	100	99.573	99.467	96.354	96.099
	200	203.350	203.220	**195.440**	196.610
	500	514.590	514.370	**492.040**	497.420
	1000	1031.500	1031.200	**985.540**	998.240
YOK	40	44.10100	44.062	**39.019**	38.703
	100	116.18000	116.170	99.919	99.837
	200	237.62000	237.610	199.910	199.900
	500	599.21000	599.200	499.920	499.950
	1000	43.88700	43.871	**39.709**	39.239

Tablo 5.6'da yer alan DSO, Q indeksi, B indeksi ve Akaike bilgi kriteri ölçütlerine göre kesim noktası 0 olduğunda elde edilen sonuçlara paralel olmak üzere, iki model arasındaki farklar anlamlı olmamıştır.

Tablo 5.6 Uyum iyiliği için kullanılan bazı ölçütlerin kesim noktası 0.53 iken modellerden elde edilen değerleri

	N	***L_DSO***	***P_DSO***	***L_Q İndeksi***	***P_Q İndeksi***	***L_B İndeksi***	***P_B İndeksi***	***L_Akaike***	***P_Akaike***
ÇOK	40	0.91818	0.91495	0.74993	0.75283	0.94391	0.94338	0.49668	0.49265
	100	0.92162	0.91982	0.74682	0.74849	0.94463	0.94415	0.41098	0.40867
	200	0.91582	0.91510	0.72929	0.73071	0.94089	0.94067	0.40528	0.40331
	500	0.91286	0.91256	0.71900	0.72019	0.93869	0.93865	0.40154	0.39989
	1000	0.91212	0.91195	0.71729	0.71844	0.93829	0.93831	0.39792	0.39633
ORTA	40	0.77758	0.77390	0.33665	0.33771	0.84804	0.84723	1.06960	1.06810
	100	0.75296	0.75214	0.28173	0.28250	0.83446	0.83425	1.05570	1.05470
	200	0.74807	0.74745	0.26659	0.26706	0.83074	0.83065	1.04670	1.04610
	500	0.74507	0.74497	0.25761	0.25792	0.82852	0.82850	1.04120	1.04070
	1000	0.74429	0.74420	0.25590	0.25615	0.82810	0.82810	1.03750	1.03720
YOK	40	0.71852	0.71635	0.20469	0.20540	0.81300	0.81267	1.25250	1.25150
	100	0.70814	0.70775	0.16195	0.16201	0.80238	0.80230	1.22180	1.22170
	200	0.70279	0.70257	0.14296	0.14299	0.79700	0.79697	1.21810	1.21810
	500	0.70157	0.70151	0.13553	0.13554	0.79490	0.79490	1.21040	1.21040
	1000	0.72207	0.72007	0.20855	0.20885	0.81484	0.81443	1.24720	1.24680

Tablo 5.7'de kesim noktası 0.53 alınarak sınıflandırılan bağımlı değişkene uygulanan Probit ve Lojit Modele ilişkin bazı yapay R^2 ölçütlerinden elde edilen değerler yer almaktadır. Yine daha önceki tablolarda olduğu gibi, her bir hücrede yer alan değer 1000 adet değerin ortalamasıdır. Kesim noktası 0 olduğunda elde edilen sonuçlarla paralel sonuçlar elde edilmiştir. İki model arasındaki farklılık McKelvey ve Zavonia'nın, Achen'nin, Buse'nin yapay R^2'lerinde ve kareli korelasyon katsayısında meydana gelmiştir. Buse, Achen ile McKelvey ve Zavonia'nın yapay R^2'sine göre "çok" ve "orta" ilişki düzeyinde Probit daha iyi bir modeldir. Bu ölçütlerde McKelvey ve Zavonia'nın yapay R^2'si hariç tüm örneklem büyüklüklüklerinde iki model arasındaki fark anlamlıdır. Sadece McKelvey ve Zavonia'nın yapay R^2'si için "çok" ilişki düzeyinde 40 ve 100 örneklem büyüklüklerinde farklar yapılan t testine göre anlamlı bulunmamıştır. Kareli korelasyon kriterine göre iki modele ilişkin değerler arasındaki anlamlı farklılık sadece, "çok" ilişki düzeyinde 500 ve 1000 örneklem büyüklüklerinde gerçekleşmiştir.

Tablo 5.7 Uyum iyiliği için kullanılan bazı yapay R^2 'lerin kesim noktası 0.53 iken modellerden elde edilen değerleri

	N	***Gerçek R^2***	***L_AldrichNelson***	***P_AldrichNelson***	***L_DüzeltilmişAN***	***P_DüzeltilmişAN***	***L_VeallZimmerman***	***P_VeallZimmerman***
ÇOK	40	0.89818	0.45880	0.46000	0.83502	0.83720	0.84005	0.84226
	100	0.90133	0.45888	0.45957	0.83516	0.83642	0.84014	0.84140
	200	0.90052	0.45566	0.45625	0.82930	0.83038	0.83080	0.83188
	500	0.90017	0.45202	0.45252	0.82268	0.82359	0.82394	0.82486
	1000	0.90021	0.45233	0.45281	0.82324	0.82411	0.82360	0.82447
ORTA	40	0.34761	0.21104	0.21188	0.38409	0.38562	0.38802	0.38957
	100	0.30630	0.17265	0.17334	0.31422	0.31548	0.31551	0.31677
	200	0.29774	0.16280	0.16324	0.29630	0.29710	0.29681	0.29761
	500	0.29219	0.15662	0.15692	0.28505	0.28559	0.28542	0.28597
	1000	0.29200	0.15642	0.15667	0.28468	0.28514	0.28489	0.28535
YOK	40	0.10852	0.10127	0.10196	0.18431	0.18557	0.18573	0.18704
	100	0.05048	0.04087	0.04096	0.07439	0.07454	0.07497	0.07512
	200	0.03469	0.02563	0.02566	0.04664	0.04671	0.04675	0.04682
	500	0.02763	0.01914	0.01915	0.03484	0.03486	0.03485	0.03487
	1000	0.09297	0.08058	0.08091	0.14666	0.14726	0.14967	0.15029

Tablo 5.7 devam

	N	*Gerçek* R^2	*L_Lave*	*P_Lave*	*L_McFadden*	*P_McFadden*	*L_Maddala*	*P_Maddala*
ÇOK	40	0.89818	0.72689	0.72429	0.71192	0.71527	0.57215	0.57388
	100	0.90133	0.73122	0.72888	0.70879	0.71070	0.57196	0.57296
	200	0.90052	0.71638	0.71533	0.69121	0.69284	0.56715	0.56802
	500	0.90017	0.70597	0.70574	0.67952	0.68087	0.56176	0.56249
	1000	0.90021	0.70502	0.70512	0.67832	0.67963	0.56220	0.56290
ORTA	40	0.34761	0.26329	0.25918	0.23695	0.23819	0.23866	0.23971
	100	0.30630	0.20108	0.20003	0.17673	0.17762	0.18993	0.19076
	200	0.29774	0.18666	0.18620	0.16216	0.16269	0.17758	0.17810
	500	0.29219	0.17802	0.17790	0.15349	0.15384	0.16981	0.17017
	1000	0.29200	0.17758	0.17759	0.15274	0.15303	0.16944	0.16973
YOK	40	0.10852	0.11232	0.11065	0.09852	0.09939	0.10855	0.10936
	100	0.05048	0.04300	0.04256	0.03646	0.03653	0.04220	0.04229
	200	0.03469	0.02636	0.02625	0.02200	0.02203	0.02613	0.02617
	500	0.02763	0.01953	0.01949	0.01612	0.01613	0.01939	0.01940
	1000	0.09297	0.09008	0.08785	0.07932	0.07968	0.08577	0.08614

Tablo 5.7 devam

	N	*Gerçek* R^2	*L_Achen*	*P_Achen*	*L_McKelveyZavonia*	*P_McKelveyZavonia*	*L_KareliKorelasyon*	*P_KareliKorelasyon*
ÇOK	40	0.89818	0.99744	**0.99985**	0.79292	0.80330	0.73750	0.73976
	100	0.90133	0.99955	**1.00000**	0.76663	0.77289	0.73717	0.73474
	200	0.90052	0.99982	**1.00000**	0.73845	0.74438	0.72256	0.71899
	500	0.90017	0.99995	**1.00000**	0.72231	**0.72746**	**0.71210**	0.70751
	1000	0.90021	0.99998	**1.00000**	0.71983	**0.72489**	**0.71126**	0.70652
ORTA	40	0.34761	0.46436	**0.55538**	0.34260	**0.36754**	0.26645	0.26569
	100	0.30630	0.33197	**0.41848**	0.27748	**0.30570**	0.20472	0.20350
	200	0.29774	0.29444	**0.36920**	0.26171	**0.29004**	0.18944	0.18775
	500	0.29219	0.26933	**0.33060**	0.25091	**0.27930**	0.18037	0.17841
	1000	0.29200	0.26638	**0.32419**	0.24942	**0.27773**	0.17976	0.17770
YOK	40	0.10852	**0.16776**	0.19976	**0.24484**	0.27546	0.11509	0.11579
	100	0.05048	0.05074	0.05413	**0.21004**	0.24381	0.04330	0.04316
	200	0.03469	0.02890	0.02957	**0.19544**	0.23044	0.02653	0.02647
	500	0.02763	0.02069	0.02093	**0.18994**	0.22560	0.01958	0.01953
	1000	0.09297	**0.13164**	0.15472	**0.25284**	0.28356	0.09082	0.09057

Tablo 5.7 devam

	N	***Gerçek R^2***	***L_Pseudo***	***P_Pseudo***	***L_Buse***	***P_Buse***	***L_CraggUhler***	***P_CraggUhler***
ÇOK	40	0.89818	0.58858	0.59263	0.66173	**0.72704**	0.81756	0.82006
	100	0.90133	0.58202	0.58430	0.59888	**0.67449**	0.81725	0.81869
	200	0.90052	0.55729	0.55920	0.56136	**0.62209**	0.80657	0.80780
	500	0.90017	0.54215	0.54372	0.53506	**0.56927**	0.79865	0.79969
	1000	0.90021	0.54003	0.54155	0.53092	**0.54514**	0.79825	0.79925
ORTA	40	0.34761	0.14880	0.14972	0.18980	**0.30978**	0.34280	0.34432
	100	0.30630	0.10303	0.10363	0.17213	**0.26803**	0.27082	0.27201
	200	0.29774	0.09267	0.09301	0.16143	**0.25216**	0.25252	0.25327
	500	0.29219	0.08673	0.08695	0.15622	**0.24376**	0.24135	0.24185
	1000	0.29200	0.08608	0.08626	0.15520	**0.24152**	0.24067	0.24108
YOK	40	0.10852	0.05597	0.05654	**0.15983**	0.26438	0.15555	0.15675
	100	0.05048	0.01959	0.01963	**0.14105**	0.23857	0.06044	0.06056
	200	0.03469	0.01154	0.01155	**0.14234**	0.23978	0.03720	0.03725
	500	0.02763	0.00836	0.00836	**0.14279**	0.24029	0.02753	0.02754
	1000	0.09297	0.04556	0.04578	**0.14301**	0.24839	0.12473	0.12527

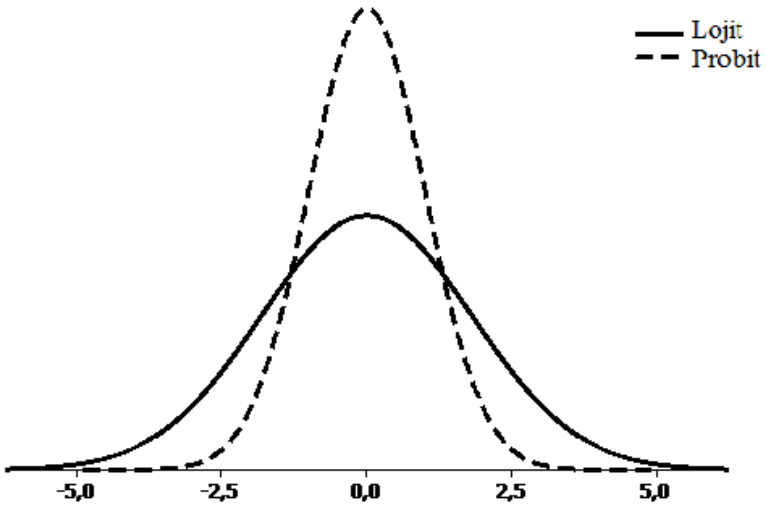
5.3 Sonuçlar ve Tartışma

Genel olarak sonuçlar incelendiğinde; kesim noktasının iki düzeyli Probit ya da Lojit Modelin daha iyi bir model olarak tercih edilmesi üzerinde etkisinin olmadığı söylenebilir. Yine güvenilen sapma ve artık ölçütlere göre ilişki düzeyinin model seçimi üzerinde önemli bir etkisinin olmadığı söylenebilir. Diğer bir dikkat çeken nokta farklı ölçütlere göre, model seçimi konusunda farklı sonuçların elde edilmesidir. Bu durum özellikle R^2 ölçütlerine olan güvenirliliği azaltmaktadır. Bölüm 4'te de belirtildiği gibi iki düzeyli bağımlı değişkenler için kullanılan ölçütlerin hangisinin daha doğru sonuç verdiği ya da hangisi dikkate alınarak yorum yapılmasının daha doğru olacağı konusunda kesin bir yargı bulunmamaktadır. Bu nedenle ölçütlerin güvenirliliği açısından doğrusal regresyon modelinden elde edilen gerçek R^2 değerlerine paralel doğrultuda değerler alan ölçütün daha güvenilir olduğu söylenebilir. Bu konu hakkında, Hagle ve Mitchell (1992) öncelikli olarak McKelvey ve Zavonia yapay R^2'sinin ve sonrasında düzeltilmiş Aldrich ve Nelson yapay R^2'sinin gerçek R^2'ye en yakın ölçütler olduğunu söylemiştir. Veall ve Zimmermann (1990) yaptığı çalışmada en iyi ölçüt olarak McKelvey ve Zavonia yapay R^2'sini önermiştir. Gerçek sürekli bağımlı değişkene ilişkin R^2 değerleri bu amaçla hesaplanmış ve tabloda belirtilmiştir. R^2'lere ilişkin dikkat çeken bir nokta da, gerçek R^2'de olduğu gibi örneklem büyüklüğü arttıkça yapay R^2 değerlerinin azalmasıdır. McKelvey ve Zavonia yapay R^2'si dikkate alınırsa, "çok" ve "orta" ilişki düzeyi için Probit Modelin Lojit Modelden daha iyi bir model olduğu, örneklem büyüklüğünün sadece ilişki düzeyi "çok" olduğunda iki model arasındaki anlamlı farklılığın oluşması bakımından etkili olduğu söylenebilir.

Yapay R^2'lerin doğruluğu konusunda yaşanabilecek sıkıntılardan kaçınmak amacıyla, modellerin uyumunu değerlendirmek için kullanılan artık değerlerine bakılabilir. Artık değerleri dikkate alındığında örneklem büyüklüğünün model seçiminde etkili olduğu görülür. İki model arasında anlamlı bir farklılık olduğu görülen düşük örneklem büyüklükleri için Probit Modelin daha düşük artık değerlere sahip olduğu ve dolayısıyla daha iyi bir model olduğu görülmüştür. Büyük örneklemlerde ise Lojit Model daha iyi bir model olmuştur.

Örneklem büyüklüğünün 200 ya da 1000 olması arasında artıklara göre model seçimi bakımından farklılık vardır. Artıklar dikkate alındığında, büyük örneklemlerde Lojit Model daha iyi sonuçlar vermiştir. İki modelin farklılığı dikkate alınarak, Lojit Modelin tercih edilebilmesi için 500 büyüklüğündeki bir örneklemin yeterli olacağı söylenebilir.

Probit Model varyansının 1, Lojit Modelin $\pi^2 / 3$ olması sebebiyle, Lojit Model daha basık bir dağılım eğrisine sahiptir. Şekil 5.3'te de görüldüğü gibi her iki model aynı eksen üzerinde olmalarına rağmen Lojit Modele ilişkin dağılım eğrisinin yayılımı daha fazla olduğundan, daha uzun kuyruklara (heavier tails) sahiptir. Bu durum, örneklem büyüklüğü arttıkça gözlemlerin kuyruklarda yer alma olasılığını arttıracağından, büyük örneklemlerde Lojit Modelin Probit Modelden daha iyi bir model olmasına sebep olur. Amemiya (1981) ve Maddala (1983), bu duruma ilişkin Probit ve Lojit Modelden elde edilen tahmini değerlerdeki farklılığın, büyük örneklemlerde meydana geldiğini ve bu durumun gözlemlerin yoğunluğunun kuyruklarda artmasından kaynaklandığını belirtmiştir. Bu çalışmadan elde edilen sonuçlar da bu yönde gerçekleşmiştir.



Şekil 5.3. Lojit ve Probit Modellere ilişkin dağılım eğrileri

Çalışmadan elde edilen diğer bir sonuç ise örneklem büyüklüğü 500 ve 1000 olduğunda ölçüt değerleri arasında meydana gelen değişimin oldukça az olduğudur. Veall ve Zimmerman (1994) yaptıkları çalışmada, bu konuya ilişkin, örneklem

büyüklüğünün 200 ya da 1000 olması arasında kullandıkları yapay R^2 değerleri bakımından çok az bir fark olduğu sonucuna ulaşmıştır. Çalışmadan elde edilen sonuçlara göre, genel olarak bu sonuç yapay R^2'ler için geçerlidir.

Yapılan bu tez çalışmasıyla beraber, literatürde iki düzeyli bağımlı değişkenli modellerin uyum iyiliği için kullanılan ölçütlerin doğru seçimi konusunda bazı sıkıntıların olduğu ortaya çıkmıştır. Bu sonuçtan yola çıkarak, bu konu hakkında mevcut kriterleri iyileştirmek ya da yeni kriterler önerme konusunda yeni çalışmaların yapılabileceği düşünülmektedir.

KAYNAKLAR

Achen, C.H., *Alternate Measures of Association in the Probit Model*, Unpublished manuscript, University of California, Berkeley, 1979.

Agresti, A., 1996. *An Introduction to Categorical Data Analysis*, John Wiley and Sons, New York, 290p.

Agresti, A., 2002. *Catagorical Data Analysis (Second Edition)*, Wiley, New Jersey, 717p.

Aldrich, J.H., Nelson, F.D., 1984. *Linear Probability, Logit, and Probit Models*, Sage Publications, London, 94p.

Altıntaş, T., *Yapay Bağımlı Değişkenli Tahmin Modelleri ve Bir Uygulama*, Yükseklisans Tezi, Ondokuz Mayıs Üniversitesi, Samsun, 2006.

Amemiya, T. 1981. Qualitative Response Models: A Survey. *JEL*, (19) : 1483-1536

Anderson-Sprecher, R. 1994. Model Comparisons and R2. *TAS*, 48, (2) : 113-117

Cameron, A.C. and Windmeijer, F.A.G., *R-Squared Measures for Count Data Regression Models with Applications to Health Care Utilization*, *Dept. of Economics Working Paper*, University of California at Davis, 93-24, 1993.

Cameron, A.C., Windmeijer, A.G. 1997. An R-squared Measure of Goodness of Fit for Some Common Nonlinear Regression Models. *JE*, (77) : 329-342

Chin, A.T.H. 2002. Impact of Frequent Flyer Programs on The Demand for Air Travel. *JAT*, 7, (2) : 53-86

Cox, D.R., Wermuth, N. 1992. A Comment on The Coefficient of Determination for Binary Responses. *TAS*, 46, (1) : 1-4

Dhrymes, P.J. 1986. Limited Dependent Variables. *Handbook of Econometrics*, Amsterdam : 1567-1631

Gan, C., et al. 2006. A logit analysis of electronic banking in New Zealand. *IJBM*, 24, (6) : 360-383

Gessner, G., et al. 1988. Estimating Models with Binary Dependent Variables: Some Theorical and Empirical Observations. *JBR*, 18, (1) : 49-65

Goldberger, A.S. 1973. Correlations Between Binary Choices and Probabilistic Predictions. *JASA*, (68) : 84

Gujarati, D.N., 1999. *Temel Ekonometri,* Literatür Yayıncılık, İstanbul, 849sy.

Hagle, T.M., Mitchell II, G.E. 1992. Goodness-of-Fit Measures for Probit and Logit. *AJPS*, 36, (3) : 762-784

Horowitz, J.L., Savin, N.E. 2001. Binary Response Models: Logits, Probits and Semiparametrics. *JEP,15, (4) : 43-56*

Hosmer, D.W., et al. 1997. A Comparison of Goodness-of-Fit Tests for The Logistic Regression Model. *SM*, (16) : 965-980

Hübler, Olaf. 1997. Pseudo Latent Models: Goodness of Fit Measures, Residuals, Estimation, Testing and Simulation. *Statistical Papers*, (38) : 271-285

Lee, L., Trost, R.P. 1982. Estimation of Some Limited Dependent Variable Models With Application to Housing Demand. *Professional Paper,* 333

Liao, T.M., 1994. *Interpreting Probability Models Logit, Probit, and Other Generalized Linear Models*, Sage Publications, London, 87p.

Long, J.S., 1997. *Regression Models for Categorical and Limited Dependent Variables*, Sage Publications, California, 297p.

Maddala, G.S., 1983. *Limited-Dependent and Qualitatives Variables in Econometrics*, Cambrige University Press, Cambrige, 397p.

Magee, L. 1990. R2 Measures Based on Wald and Likelihood Ratio Joint Significance Tests. *TAS*, 44, (3) : 250-253

Malhotra, N.K. 1983. A Comparison of the Predictive Validity of Procedures for Analyzing Binary Data. *JBES*, 1, (4) : 326-336

McCullagh, P., Nelder, J.A., 1989. *Generalized Linear Models (Second Edition),*Chapman and Hall, London, 526p.

McFadden, D.L., 1984. *Econometric analysıs of Qualitative Response Models,* Elsevier Science Publishers, BV, 1446p.

Merkle, L., Zimmermann, K.F, 1992. The Demographics of Labor Turnover: A Comparison of Ordinal Probit and Censored Count Data Models. *REL*, (58) : 283-307

Morimune, K. 1979. Comparisons of Normal and Logistic Models in the Bivariate Dichotomous Analysis. *Econometrica*, 47, (4) : 957-975

Morrison, D.G. 1972. Upper Bounds for Correlation Between Binary Outcomes and Probabilistic Predictions. *JASA*, (67) : 68-70

Nelder, A.J., Wedderburn, R.W.M. 1972. Generalized Linear Models. JRSS, 135, (3) : 370-384

Özarıcı, Ö., *Farklı Not sistemlerinde Öğrencini Başarılı Olma Olasılığının Probit Regresyon Analiziyle Değerlendirilmesi*, Yükseklisans Tezi, Osmangazi Üniversitesi, Eskişehir, 1996.

Pampel, F.C., 2000. *Logistic Regression: A Primer*, Sage Publications, California, 94p.

Powers, D.A., Xie, Y., 1999. *Statistical Methods for Categorical Data Analysis*, Academic Press, 309p.

Prentice. R.L. 1976. A Generalization of the Probit and Logit Methods for Dose Response Curves. *Biometrics*, 32, (4) : 761-768

Ragsdale, L. 2011. The Politics of Presidential Speechmaking, 1949-1980. *APSR*, 78, (4) : 971-984

Sönmez, Ö., *Nitel Tercih Modelleri, Çoklu Logit, Probit Modeller ve Bir Uygulama,* Yüksek Lisans Tezi, Yıldız Teknik Üniversitesi, İstanbul, 2006.

Tardiff, T.J. 1976. A Note on Goodness-of-Fit Statistics For Probit and Logit Models. *Transportation,* (5) : 377-388

Tatlıdil, H., 1996. *Çok Değişkenli İstatistiksel Analiz*. Akademi Matbaası, Ankara, 424sy.

Temple,J. 2004. Explaining The Private Health Insurance Coverage of Older Australians. *People and Place,* 12, (2) : 13-23

Uçar, Ö., *Nitel Verilerin Analizinde Lojit ve Probit Modeller,* Yükseklisans Tezi, Hacettepe Üniversitesi, Ankara, 2004.

Wilkermann, R., Boes, S., 2006. *Analyses of Microdata*, Springer, Berlin, 317p.

Veall, M.R., Zimmermann, K.F. 1994. Evaluating Pseudo-R^2's for Binary Probit Models. *Quality & Quantity,* 28, (2) : 151-164

Veall, M.R., Zimmermann, K.F. 1996. Pseudo-R2 Measures for Some Common Limited Dependent Variable Models. *Sunderforschungsbereich,* 386, (18) : 1-34

Wiersema, M.F., Bowen, P., 2009. The use of limited dependent variable techniques in strategy research: Issues and methods. *SMJ*

Windmeijer, F.A.G. 1995. Goodness of Fit Measures in Binary Choice Models. *Econometric Review,* (14) : 101-116

Winkelmann, R., Boes, S., 2006. *Analysis of Microdata*, Springer, Berlin, 317p.

Zelner. B.A., *Using Simulation to Interpret and Present Logit and Probit Results,* Working Paper, 2008.

ÖZGEÇMİŞ

Adı Soyadı **:** Selen ÇAKMAKYAPAN
Doğum Yeri **:** Ankara
Doğum Yılı **:** 1985
Medeni Hali **:** Bekar

Eğitim ve Akademik Durumu

Lise **:** (1999-2003) Büyükşehir Hüseyin Yıldız Anadolu Lisesi
Lisans **:** (2003-2007) Muğla Üniversitesi İstatistik Bölümü
(2005-2008) Muğla Üniversitesi Matematik Bölümü
Yabancı Dil **:** İngilizce

İş Tecrübesi

2010-2011 Muğla Üniversitesi İstatistik Bölümü Araştırma Görevlisi