ÇUKUROVA ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

YÜKSEK LİSANS TEZİ

Pelin İYİ

GENETİK ALGORİTMA UYGULANARAK VE BİLGİ KRİTERLERİ KULLANILARAK ÇOKLU REGRESYONDA MODEL SEÇİMİ

İSTATİSTİK ANABİLİM DALI

ADANA, 2006

ÇUKUROVA ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

GENETİK ALGORİTMA UYGULANARAK VE BİLGİ KRİTERLERİ KULLANILARAK ÇOKLU REGRESYONDA MODEL SEÇİMİ

Pelin İYİ

YÜKSEK LİSANS TEZİ
İSTATİSTİK ANABİLİM DALI

Bu tez 11 / 09 / 2006 tarihinde aşağıdaki jüri üyeleri tarafından oybirliği / oyçokluğu ile kabul edilmiştir.

İmza:	İmza:	İmza:
Prof.Dr. Hamza EROL	**Doç.Dr. Selahattin KAÇIRANLAR**	**Yard.Doç.Dr. Ahmet TEMİZYÜREK**
DANIŞMAN	ÜYE	ÜYE

Bu tez Enstitümüz İstatistik Anabilim Dalında hazırlanmıştır.
Kod No:

Prof.Dr. Aziz ERTUNÇ
Enstitü Müdürü
İmza ve Mühür

Bu çalışma Ç.Ü. Bilimsel Araştırma Projeleri Birimi tarafından desteklenmiştir.
Proje No:FEF2004YL59

Not: Bu tezde kullanılan özgün ve başka kaynaktan yapılan bildirişlerin, çizelge, şekil ve fotoğrafların kaynak gösterilmeden kullanımı, 5846 sayılı fikir ve Sanat Eserleri Kanunundaki hükümlere tabidir.

ÖZ

YÜKSEK LİSANS TEZİ

GENETİK ALGORİTMA UYGULANARAK VE BİLGİ KRİTERLERİ KULLANILARAK ÇOKLU REGRESYONDA MODEL SEÇİMİ

Pelin İYİ

ÇUKUROVA ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ
İSTATİSTİK ANABİLİM DALI

Danışman: Prof.Dr. Hamza EROL

Yıl: 2006, Sayfa: 125

Jüri: Prof.Dr. Hamza EROL

Doç.Dr. Selahattin KAÇIRANLAR

Yard.Doç.Dr. Ahmet TEMİZYÜREK

Çoklu lineer regresyon modelinde açıklayıcı değişken sayısı fazla olduğunda aday model sayısı da üstel olarak artmaktadır. Bu durumda geleneksel yöntemlerle, adımsal yöntemlerle ve istatistik paket programları kullanılarak model seçimi mümkün değildir.

Bu çalışmada açıklayıcı değişken sayısının fazla olması durumunda ortaya çıkan model seçimi problemi, genetik algoritma uygulanarak ve bilgi kriterleri kullanılarak incelenmiştir. Bu amaçla çalışmada önce, çoklu lineer regresyon modeli hakkında genel bilgiler verilmiş ve çoklu lineer regresyon modellerinin oluşturulması açıklanmıştır. Sonra, açıklayıcı değişken sayısının fazla olması durumunda çoklu regresyonda ortaya çıkan en iyi modelin seçimi problemi adımsal yöntemlerle incelenmiştir. Daha sonra da, çoklu lineer regresyon modeli için genetik algoritma ve bilgi kriterleri açıklanmıştır. Çoklu lineer regresyonda genetik algoritma uygulanarak ve bilgi kriterleri kullanılarak model seçimi incelenmiştir. Genetik algoritma için kod oluşturulması ele alınmıştır. Son olarak, sonuç ve öneriler tartışılmıştır.

Anahtar kelimeler: Bilgi kriteri, Çoklu lineer regresyon, Genetik algoritma, Model seçimi.

ABSTRACT

MSc THESIS

MODEL SELECTION IN MULTIPLE REGRESSION BY APPLYING GENETIC ALGORİTHM AND BY USING INFORMATION CRITERIA

Pelin İYİ

DEPARTMENT OF STATISTICS
INSTITUTE OF NATURAL AND APPLIED SCIENCES
UNIVERSITY OF ÇUKUROVA

Supervisor: Prof.Dr. Hamza EROL

Year: 2006, Pages: 125

Jury: Prof.Dr. Hamza EROL

Assoc.Prof.Dr. Selahattin KAÇIRANLAR

Asist.Prof.Dr. Ahmet TEMİZYÜREK

The number of models increases exponentially when the explanatory variables increases in a multiple linear regression model. In this case, model selection is impossible by using traditional procedures, stepwise methods and even existing statistical softwares.

In this study, the model selection problem in a multiple linear regression model when there are more explanatory variables or regressors is considered by applying genetic algorithm and by using information criterias. For this purpose first, general information about multiple linear regression model are given and building multiple linear regression model is explained. Then, the best model selection problem in a multiple linear regression model when there are more explanatory variables is examined by stepwise methods. After than, genetic algorithm and information criterias for multiple linear regression model are emphasized, following model selection in multiple regression by applying genetic algorithm and by using information criterias is explained. Finally, results and discussions are given.

Key words: Information criterion, Multiple linear regression, Genetic algorithm, Model selection.

TEŞEKKÜR

Bu tezin hazırlanmasında, bilgi ve birikimlerinden daima faydalandığım ve yardımlarını hiçbir zaman esirgemeyen danışmanım, Prof.Dr. Hamza EROL'a; İstatistik bölümü öğretim elemanlarına ve maddi ve manevi desteklerini hiçbir zaman esirgemeyerek her zaman yanımda olan aileme teşekkür ederim.

İÇİNDEKİLER **SAYFA NO**

ÖZ …….....I
ABSTRACT……....II
TEŞEKKÜR……...III
İÇİNDEKİLER……...IV
TABLOLAR DİZİNİ…….VI
ŞEKİLLER DİZİNİ…….…VIII
1. GİRİŞ……...1
1.1. Çoklu Lineer Regresyon Modeli……...1
1.2. En İyi Çoklu Lineer Regresyon Modelinin Seçilmesi…….……..1
1.3. En İyi Çoklu Lineer Regresyon Modelinin Seçiminde Uygulanacak Kriterler – Klasik Yöntem…….……1
1.4. En İyi Çoklu Lineer Regresyon Modelinin Seçiminde Uygulanacak Kriterler – Adımsal Yöntemler……..2
1.5. Çoklu Lineer Regresyon Modellerinde Genetik Algoritmanın Uygulanması…….2
1.6. Çoklu Lineer Regresyon Modellerinde Genetik Algoritma Uygulanırken Bilgi Kriterlerinin Kullanılması…….…….3
2. ÖNCEKİ ÇALIŞMALAR…….…..4
2.1. Çoklu Lineer Regresyon Modeli İle İlgili Çalışmalar…….…….4
2.2. En İyi Çoklu Lineer Regresyon Modelinin Seçilmesi İle İlgili Çalışmalar – Klasik Yöntem ve Adımsal Yöntemler…….…….6
2.3. En İyi Çoklu Lineer Regresyon Modelinin Seçilmesi İle İlgili Çalışmalar – Genetik Algoritma……....10
3. ÇOKLU LİNEER REGRESYON MODELİ…….16
3.1. Çoklu Lineer Regresyon Modeli Hakkında Genel Bilgiler…….16
3.2. Çoklu Lineer Regresyon Modelindeki Parametrelerinin Tahmin Edilmesi…….19
3.2.1. Regresyon Katsayılarının En Küçük Kareler Yöntemiyle Tahmin Edilmesi……..19

İÇİNDEKİLER **SAYFA NO**

3.2.2. Çoklu Lineer Regresyon Modelinde Matris Gösteriminin Kullanılması ve Regresyon Katsayılarının En Küçük Kareler Yöntemiyle Tahmin Edilmesi…….21
3.2.3. En Küçük Kareler Yönteminin Geometrik Yorumu…….29
3.2.4. En Küçük Kareler Tahmin Edicilerinin Özellikleri……..30
3.2.5. σ^2'nin Tahmini……32
3.2.6. Regresyon Katsayılarının En Çok Olabilirlik Yöntemiyle Tahmin Edilmesi……..33
3.3. Çoklu Lineer Regresyon Modelinde Hipotez Testi……35
3.3.1. Regresyonun Önemliliğinin Test Edilmesi……..35
3.3.2. Her bir Regresyon Katsayısı için Hipotezin Test Edilmesi…..39
3.3.3. Regresyon Katsayılarının Bir alt Kümesi İçin Hipotezin Test Edilmesi……..40
3.3.4. X Matrisinde Sütunların Ortogonal Olması Özel Durumu…..44
3.3.5. $\boldsymbol{T\beta} = 0$ Genel Lineer Hipotezlerinin Test Edilmesi……47
3.4. Çoklu Regresyonda Güven Aralıkları…….48
3.4.1. Regresyon Katsayıları İçin Güven Aralıkları……...48
3.4.2. Ortalama Yanıt İçin Güven Aralığı……..50
4. ÇOKLU LİNEER REGRESYON MODELLERİNİN OLUŞTURULMASI VE EN İYİ MODELİN SEÇİLMESİ……...52
4.1. Çoklu Lineer Regresyonda En İyi Modelin Seçilmesi……...52
4.2. Yanlış Modelin Belirlenmesinin Sonuçları……56
4.3. Regresörlerin Bir Alt Kümesini Seçmek İçin Kriterler……..61
4.3.1. Çoklu Belirleyicilik Katsayısı……..61
4.3.2. Düzeltilmiş Çoklu Belirleyicilik Katsayısı……..63
4.3.3. Hata Kareler Ortalaması……..64
4.3.4. Mallows'un C_p İstatistiği……65

İÇİNDEKİLER **SAYFA NO**

4.4. Regresyon ve Model Değerlendirme Ölçütlerinin Kullanımı……68
4.5. Değişken Seçimi İçin Hesaplama Teknikleri……69
4.5.1 Olası Bütün Regresyonlar……70
4.5.2. Adımsal Regresyon Yöntemleri……..78
4.5.2.1. İleriye Doğru Seçim Yöntemi……..78
4.5.2.2. Geriye Doğru Ayıklama Yöntemi……81
4.5.2.3. Adımsal Regresyon Yöntemi……...83
5. ÇOKLU LİNEER REGRESYON MODELİ İÇİN BİR GENETİK ALGORİTMA……..86
5.1. Genetik Algoritmalar Hakkında Genel Bilgiler……..86
5.2. Çoklu Lineer Regresyon Modelinde En İyi Modelin Oluşturulmasında Genetik Algoritmanın Kullanılması ve Bilgi Karmaşıklık Kriteri…….91
5.3. ICOMP Bilgi Karmaşıklık Kriteri Kullanılarak Genetik Algoritmanın Uygulanması……94
5.3.1. Karmaşıklık Kriteri ve Bir Sistemin Karmaşıklığı……...96
5.4. Çoklu Lineer Regresyon Modeli İçin Bilgi Kriteri…….98
5.4.1. Karmaşıklık Ölçülerine Dayalı ICOMP Değeri……..100
5.5. Çoklu Lineer Regresyon Modeli İçin Bir Genetik Algoritma……...101
5.5.1. Çoklu Lineer Regresyon Modelleri İçin Bir Genetik Kodlama Şeması……....102
5.5.2. Çoklu Lineer Regresyon Modeli İçin Genetik Algoritmada Kullanılacak Başlangıç Popülasyonunun Oluşturulması…....103
5.5.3. Herhangi Bir Çoklu Lineer Regresyon Modelinin Performansının Değerlendirilmesi için Bir Uyum Fonksiyonu……..104
5.5.4. Oluşturulan Çoklu Lineer Regresyon Modellerini Seçmek için Bir Mekanizma……....104

İÇİNDEKİLER **SAYFA NO**

5.5.5. Yeni Nesil Çoklu Lineer Regresyon Modellerini Üretmek İçin Erişkin Modellerin Eşleştirilmesini Yapmak Amacıyla Bir Yeniden Üretim İşlemi……..105
5.5.5.1. Tek Nokta Çaprazlama……107
5.5.5.2. İki Nokta Çaprazlama……..107
5.5.5.3. Düzgün Çaprazlama……108
5.5.6. Yeni Nesil Modellerin Birleşimini Değiştirmek İçin Değişmenin Etkisi……109
6. SONUÇ VE ÖNERİLER……118
KAYNAKLAR……119
ÖZGEÇMİŞ……..125

ÇİZELGE DİZİNİ **SAYFA NO**

Tablo 3.1. Çoklu lineer regresyon modeli için veriler. 19

Tablo 3.2. İçecek teslim/dağıtım verisi (Montgomery ve ark., 2001). 25

Tablo 3.3. İçecek teslim/dağıtım verisindeki yanıt değişken y ile açıklayıcı değişkenler x_1 ve x_2 arasındaki ilişki için oluşturulan matris grafiği. 26

Tablo 3.4. İçecek teslim/dağıtım verisindeki y, x_1 ve x_2 için tanımlayıcı istatistiklerin analiz sonuçlarının bilgisayar çıktısı. 27

Tablo 3.5. İçecek teslim/dağıtım verisindeki x_1 ve x_2 açıklayıcı değişkenlerinin y yanıt değişkendeki toplam değişimi açıklama oranı için analiz sonuçlarının bilgisayar çıktısı. 27

Tablo 3.6. İçecek teslim/dağıtım verisi için regresyonun önemliliği testinin varyans analizi tablosu. 28

Tablo 3.7. İçecek teslim/dağıtım verisi için oluşturulan regresyonun modelindeki parametre tahmin değerleri. 28

Tablo 3.8. Çoklu lineer regresyon modelinde regresyonun önemliliğini test etmek için kullanılan varyans analizi tablosu. 36

Tablo 3.9. İçecek teslim/dağıtım verisi için oluşturulan çoklu lineer regresyon modelinde regresyonun önemliliğini test etmek için kullanılan varyans analizi tablosu. 37

Tablo 4.1. Örnek 2.1. için Hald Çimento verisi (Montgomery ve ark., 2001). 69

Tablo 4.2. Örnek 2.1 için bütün aday regresyon modellerinin özeti (Montgomery ve ark., 2001). 70

ÇİZELGE DİZİNİ **SAYFA NO**

Tablo 4.3. Örnek 2.1 için bütün aday regresyon modellerindeki parametrelerin en küçük kareler yöntemiyle elde edilen tahminleri (Montgomery ve ark., 2001). 71

Tablo 4.4. Örnek 2.1'deki Hald çimento verisi için basit korelesyon matrisi (Montgomery ve ark., 2001). 73

Tablo 4.5. Hald çimento verisi için iki modelin karşılaştırılması (Montgomery ve ark., 2001). 76

Tablo 5.1. Beş açıklayıcı değişken bulunan ve sabit terim içeren çoklu lineer regresyon modeli için ikili string gösterimi. 102

Tablo 5.2. Vücut yağı verileri için bütün olası modeller arasından en küçük $ICOMP(IFIM)$ değerlerine göre seçilmiş on beş en iyi model (Bozdogan, 2004). 111

Tablo 5.3. Vücut yağı verileri için Matlab programında hazırlanan GA programının çalıştırılmasında kullanılan parametreler (Bozdogan, 2004). 112

Tablo 5.4. Genetik Algoritmanın 100 kez çalıştırılmasından sonra vücut yağı veri kümesi için ilk 10 sıradaki en iyi açıklayıcı değişkenlerin alt kümesi (Bozdogan, 2004). 112

Tablo 5.5. En iyi alt küme modelinin uyumunun özeti (Bozdogan, 2004). 113

Tablo 5.6. En iyi alt küme Genetik Algoritma modelinin parametre tahminleri (Bozdogan, 2004). 113

TABLO DİZİNİ **SAYFA NO**

Tablo 3.1. Çoklu lineer regresyon modeli için veriler. 19

Tablo 3.2. İçecek teslim/dağıtım verisi (Montgomery ve ark., 2001). 26

Tablo 3.3. İçecek teslim/dağıtım verisindeki yanıt değişken y ile açıklayıcı değişkenler x_1 ve x_2 arasındaki ilişki için analiz sonuçlarının bilgisayar çıktısı. 27

Tablo 3.4. İçecek teslim/dağıtım verisindeki y, x_1 ve x_2 için tanımlayıcı istatistiklerin analiz sonuçlarının bilgisayar çıktısı. 27

Tablo 3.5. İçecek teslim/dağıtım verisindeki x_1 ve x_2 açıklayıcı değişkenlerinin y yanıt değişkendeki toplam değişimi açıklama oranı için analiz sonuçlarının bilgisayar çıktısı. 28

Tablo 3.6. İçecek teslim/dağıtım verisi için regresyonun önemliliği testinin varyans analizi tablosu. 28

Tablo 3.7. İçecek teslim/dağıtım verisi için oluşturulan regresyonun modelindeki parametre tahmin değerleri. 29

Tablo 3.8. Çoklu lineer regresyon modelinde regresyonun önemliliğini test etmek için kullanılan varyans analizi tablosu. 36

Tablo 3.9. İçecek teslim/dağıtım verisi için oluşturulan çoklu lineer regresyon modelinde regresyonun önemliliğini test etmek için kullanılan varyans analizi tablosu. 38

Tablo 4.1. Örnek 2.1. için Hald Çimento verisi (Montgomery ve ark., 2001). 70

Tablo 4.2. Örnek 2.1 için bütün aday regresyon modellerinin özeti (Montgomery ve ark., 2001). 71

TABLO DİZİNİ **SAYFA NO**

Tablo 4.3. Örnek 2.1 için bütün aday regresyon modellerindeki parametrelerin en küçük kareler yöntemiyle elde edilen tahminleri (Montgomery ve ark., 2001). 72

Tablo 4.4. Örnek 2.1'deki Hald çimento verisi için basit korelesyon matrisi (Montgomery ve ark., 2001). 74

Tablo 4.5. Hald çimento verisi için iki modelin karşılaştırılması (Montgomery ve ark., 2001). 77

Tablo 5.1. Beş açıklayıcı değişken bulunan ve sabit terim içeren çoklu lineer regresyon modeli için ikili string gösterimi. 103

Tablo 5.2. Vücut yağı verileri için bütün olası modeller arasından en küçük $ICOMP(IFIM)$ değerlerine göre seçilmiş on beş en iyi model (Bozdogan, 2003). 112

Tablo 5.3. Vücut yağı verileri için Matlab programında hazırlanan GA programının çalıştırılmasında kullanılan parametreler (Bozdogan, 2003). 113

Tablo 5.4. Genetik Algoritmanın 100 kez çalıştırılmasından sonra vücut yağı veri kümesi için ilk 10 sıradaki en iyi açıklayıcı değişkenlerin alt kümesi (Bozdogan, 2003). 113

Tablo 5.5. En iyi alt küme modelinin uyumunun özeti (Bozdogan, 2003). 114

Tablo 5.6. En iyi alt küme Genetik Algoritma modelinin parametre tahminleri (Bozdogan, 2003). 114

ŞEKİLLER DİZİNİ **SAYFA NO**

Şekil 3.1. İki boyutlu uzayda çoklu lineer regresyon modeli bir regresyon düzlemi belirtir (Montgomery ve ark., 2001). 17

Şekil 3.2. İçecek teslim/dağıtım verisindeki yanıt değişken y ile açıklayıcı değişkenler x_1 ve x_2 arasındaki ilişki için oluşturulan matris grafiği. 26

Şekil 3.3. En küçük kareler yönteminin bir geometrik yorumu (Montgomery ve ark., 2001). 29

Şekil 4.1. Modeldeki terim sayısı p değerlerine karşı çoklu belirleyicilik katsayısı R_p^2'nin grafiği (Montgomery ve ark., 2001). 62

Şekil 4.2. p değerlerine karşı $MS_E(p)$ değerlerinin grafiği (Montgomery ve ark., 2001). 64

Şekil 4.3. p değerlerine karşılık C_p değerlerinin grafiği (Montgomery ve ark., 2001). 67

Şekil 4.4. p değerlerine karşı $R^2{}_p$ değerlerinin grafiği (Montgomery ve ark., 2001). 73

Şekil 4.5. p değerlerine karşı $MS_E(p)$ değerlerinin grafiği (Montgomery ve ark., 2001). 75

Şekil 4.6. p değerlerine karşı C_p grafiği (Montgomery ve ark., 2001). 76

Şekil 4.7. Hald çimento verisinin SAS bilgisayar programıyla yapılan analiz sonucu (İleriye Doğru Seçim Yöntemi) (Montgomery ve ark., 2001). 80

Şekil 4.8. Hald çimento verisinin SAS bilgisayar programıyla yapılan analiz sonucu (Geriye Doğru Ayıklama Yöntemi) (Montgomery ve ark., 2001). 82

ŞEKİLLER DİZİNİ **SAYFA NO**

Şekil 4.9. Hald çimento verisinin SAS bilgisayar programıyla yapılan analiz sonucu (Adımsal Regresyon Yöntemi) (Montgomery ve ark., 2001). 84

Şekil 5.1. Verilen erişkin iki modelin ikili string çifti için çaprazlama yoluyla çiftleştirme işlemine bir örnek (Bozdogan, 2003). 106

Şekil 5.2. Verilen erişkin iki modelin ikili string çifti için tek nokta çaprazlama yoluyla çiftleştirme işlemine bir örnek (Bozdogan, 2003). 107

Şekil 5.3. Verilen erişkin iki modelin ikili string çifti için iki nokta çaprazlama yoluyla çiftleştirme işlemine bir örnek (Bozdogan, 2003). 108

Şekil 5.4. Verilen erişkin iki modelin ikili string çifti için düzgün çaprazlama yoluyla çiftleştirme işlemine bir örnek (Bozdogan, 2003). 108

Şekil 5.5. ICOMP ile hesaplanan tüm modellerin oluşturduğu yapının üç boyutlu grafiği (Bozdogan, 2003). 115

Şekil 5.6. Vücut yağ veririsi için Genetik Algoritmanın 100 kez çalışmasının bir özeti (Bozdogan, 2003). 116

Şekil 5.7. Vücut yağ verisi için Genetik Algoritmanın 100 kez çalıştırılması sonucunda ICOMP(IFIM) ile hesaplanan tüm modellerin oluşturduğu yapının üç boyutlu grafiği (Bozdogan, 2003). 116

1. GİRİŞ

1.1. Çoklu Lineer Regresyon Modeli

Bir yanıt (bağımlı) değişkendeki toplam değişimi açıklamak amacıyla birden fazla regresör (açıklayıcı) değişken kullanılarak oluşturulan regresyon modeline çoklu regresyon modeli denir. Çoklu lineer regresyon modelleri genelde regresyon probleminin çözümüne yaklaşım fonksiyonu olarak kullanılırlar (Montgomery ve ark., 2001).

1.2. En İyi Çoklu Lineer Regresyon Modelinin Seçilmesi

Çoklu lineer regresyon modelinde, y yanıt değişkendeki toplam değişimi açıklayan en iyi regresyon modelinin seçilmesi "değişken seçimi" ya da "en iyi alt küme modelinin seçimi" olarak adlandırılır (Draper ve Smith, 1998). k tane açıklayıcı değişken ya da regresör içeren çoklu lineer regresyon modeli için 2^k tane aday model (alt küme modeli) vardır (Gunst ve Mason, 1980). En iyi regresyon modelinin belirlenmesinin iki amacı vardır: Birincisi, modele katkısı istatistiksel olarak anlamsız değişkenleri çıkararak, oluşturulan modelin değişken sayısının azaltılması istenir. Böylece işlemler için gereken süre ve maliyet azalır. İkincisi ise modelin olası birçok regresör içermesi istenir. Çünkü değişkenlerdeki bilgi içeriği, tahmin edilen yanıt değerlerini etkiler (Montgomery ve ark., 2001).

1.3. En İyi Çoklu Lineer Regresyon Modelinin Seçiminde Uygulanacak Kriterler – Klasik Yöntem

En iyi regresyon modelinin belirlenmesinde klasik yöntem uygulandığında çoklu belirleyicilik katsayısı R^2 veya düzeltilmiş çoklu belirleyicilik katsayısı $R^2_{Düzeltilmiş}$ ve hata kareleri ortalamaları (HKO) kullanılabilir (Draper ve Smith, 1998). Eşit sayıda açıklayıcı değişken içeren modellerin karşılaştırılmasında çoklu

belirleyicilik katsayısı R^2 ve farklı sayıda açıklayıcı değişken içeren modellerin karşılaştırılmasında düzeltilmiş çoklu belirleyicilik katsayısı $R^2_{Düzeltilmiş}$ değerleri kullanılır. En iyi regresyon modelinin belirlenmesinde R^2'si veya $R^2_{Düzeltilmiş}$'si yüksek, HKO'sı düşük olan ve az sayıda açıklayıcı değişken içeren model tercih edilir (Montgomery ve ark., 2001).

1.4. En İyi Çoklu Lineer Regresyon Modelinin Seçiminde Uygulanacak Kriterler – Adımsal Yöntemler

Bazı durumlarda mevcut açıklayıcı değişkenler, yanıt değişkendeki toplam değişimi açıklamada yetersiz kalabilir. Böyle durumlarda regresyon modeline yeni açıklayıcı değişken ya da değişkenler eklenebilir. Bazı durumlarda ise mevcut açıklayıcı değişkenlerden bazıları yanıt değişkendeki toplam değişimi açıklamada istatistiksel olarak etkileri ya da katkıları bulunmadığından çoklu lineer regresyon modelinden silinebilir ya da çıkarılabilir (Chatterjee ve ark., 2000).

Çoklu lineer regresyon modelindeki açıklayıcı değişken sayısının artması durumunda ileriye doğru seçim, geriye doğru ayıklama ya da adımsal regresyon gibi yöntemler uygulanabilir (Miller, 1990).

En iyi regresyon modelinin belirlenmesinde adımsal yöntemler uygulandığında klasik yöntemdeki kriterler uygulanır.

1.5. Çoklu Lineer Regresyon Modellerinde Genetik Algoritmanın Uygulanması

Çoklu lineer regresyon modelindeki açıklayıcı değişken sayısının fazla olması durumunda ne geleneksel yöntemler ne de adımsal yöntemler kullanılamamaktadır (Bozdoğan, 2003). Çoklu lineer regresyon modelindeki açıklayıcı değişken sayısının fazla olması durumunda çoklu lineer regresyon modelinin oluşturulmasında Genetik Algoritma kullanılabilir (Wasserman ve Sudjianto, 1994; Wallet ve ark., 1996).

1.6. Çoklu Lineer Regresyon Modellerinde Genetik Algoritma Uygulanırken Bilgi Kriterlerinin Kullanılması

Çoklu lineer regresyon modelinde model seçimi için Genetik Algoritma uygulanırken Akaike'nin Bilgi Kriteri (Akaike 1973, 1987, 1994; Sclove 1987; Bozdogan 2000), Mallows'un C_p istatistiği (Mallows 1964, 1966, 1973; Montgomery ve ark., 2001) ya da Bozdogan'ın Bilgi karmaşıklığı (Bozdogan 1987, 1990, 2000, 2003, 2004; Bozdogan ve Ueno 2000) gibi bir bilgi kriteri kullanılabilir.

2. ÖNCEKİ ÇALIŞMALAR

2.1. Çoklu Lineer Regresyon Modeli İle İlgili Çalışmalar

Bir yanıt (bağımlı) değişken ve birden fazla regresör (açıklayıcı) değişken içeren regresyon modelleri istatistiksel analizin temel konularındandır (Montgomery ve ark., 2001). Yanıt değişken y, k tane açıklayıcı değişken $x_1, x_2, ..., x_k$ ile ilişkili olabilir. Bu nedenle,

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + \varepsilon \quad (2.1)$$

eşitliği k tane açıklayıcı değişkenli çoklu lineer regresyon modeli olarak adlandırılır (Montgomery ve ark., 2001). Burada $\beta_0, \beta_1, \beta_2, ..., \beta_k$ parametreleri regresyon katsayılarını ve ε hata terimini göstermektedir. Bu modelde $j = 1, 2, ..., k$ olmak üzere x_j açıklayıcı değişkenlerinin k boyutlu uzayında bir hiper düzlem belirtir (Montgomery ve ark., 2001). β_j parametresi, j. açıklayıcı değişken dışındaki tüm açıklayıcı değişkenler sabit tutulduğunda, x_j'deki bir birimlik değişim nedeniyle yanıt değişken y'de olabilecek ya da beklenen değişim miktarını belirtir. Bu nedenle, $\beta_0, \beta_1, \beta_2, ..., \beta_k$ parametreleri genelde kısmi regresyon katsayıları olarak adlandırıldı (Draper ve Smith, 1998).

En küçük kareler yöntemi, (2.1)'deki eşitlikteki çoklu lineer regresyon modelinin regresyon katsayılarını tahmin etmek için kullanıldı (Gunst ve Mason, 1980).

Çoklu lineer regresyon modeli matris gösterimiyle,

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon} \quad (2.2)$$

şeklinde yazıldı (Hocking 1976, 1983; Miller 1990). Burada $\boldsymbol{y}=\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$, $n\times 1$ tipinde olmak üzere gözlemlerin vektörünü; $\boldsymbol{X}=\begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & x_{n1} & x_{n2} & & x_{nk} \end{bmatrix}$, $n\times p$ tipinde olmak üzere açıklayıcı değişkenlerin düzeylerinin matrisini; $\boldsymbol{\beta}=\begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}$, $p\times 1$ tipinde olmak üzere regresyon katsayılarının vektörünü ve $\boldsymbol{\varepsilon}=\begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$, $n\times 1$ tipinde olmak üzere rastgele hataların vektörünü göstermektedir.

(2.2)'teki matris formundaki çoklu lineer regresyon modelinde hatalar, sıfır ortalamalı E($\boldsymbol{\varepsilon}$)=0 ve σ^2 sabit varyanslı V($\boldsymbol{\varepsilon}$)=σ^2 normal dağılıma sahiptir. Çoklu lineer regresyon modelinde $\varepsilon \sim N(0,\sigma^2 I)$ varsayımı yapılarak $\boldsymbol{\beta}$ parametre vektörünün $\tilde{\boldsymbol{\beta}}$ en çok olabilirlik tahmin edicisi elde edildi (Gunst ve Mason, 1980).

Çoklu lineer regresyon modelinde hataların normal dağılıma sahip olduğu varsayımı yapılarak regresyon katsayılarıyla ilgili hipotezler test edildi (Myers 1990; Montgomery ve ark., 2001). Çoklu lineer regresyon modelinde regresyon katsayıları ve ortalama yanıt için güven aralıkları oluşturuldu (Myers 1990; Montgomery ve ark., 2001).

2.2. En İyi Çoklu Lineer Regresyon Modelinin Seçilmesi İle İlgili Çalışmalar – Klasik Yöntem ve Adımsal Yöntemler

Çoklu lineer regresyon modelinde yanıt değişkendeki toplam değişimi açıklamada regresörlerin bir kümesinin belirlenmesi gerekir (Hocking 1972; Hocking ve LaMotte 1973; Cox ve Snell 1974). Regresyon modeli için istatistiksel olarak anlamlı ya da önemli regresörlerin uygun alt kümesini belirleme işlemine "değişken seçimi problemi" adı verildi (Draper ve Smith, 1998).

Çoklu lineer regresyon modeli oluşturulurken genellikle regresörlerin doğru fonksiyonel biçiminin bilindiği, veride aykırı ya da sapan değerlerin ve etkili gözlemlerin bulunmadığı varsayıldı. Modelin doğru fonksiyonel biçiminin, verideki sapan ya da aykırı gözlem değerlerinin ve verideki etkin gözlem değerlerinin belirlenmesi problemlerinin eş zamanlı çözülmesi gerekse bile çoğu kez ardışık yaklaşım kullanıldı. Önce değişken seçimi stratejisi kullanıldı. Sonra sonuçta bulunan alt küme modeli doğru fonksiyonel belirtisi için sapan değerler için ve etkili gözlemler için kontrol edildi. Bu, birinci adımın tekrarlanması gerektiğini belirtebilir. Yeterli bir model oluşturmak için bir çok ardışık işlem gerekebilir (Montgomery ve ark., 2001).

Geleneksel yöntemlerde en iyi regresyon modelinin belirlenmesinde çoklu belirleyicilik katsayısı R^2 veya düzeltilmiş çoklu belirleyicilik katsayısı $R^2_{Düzeltilmiş}$ ve hata kareleri ortalamaları (HKO) kullanıldı (Draper ve Smith, 1998). En iyi regresyon modelinin belirlenmesinde R^2'si veya $R^2_{Düzeltilmiş}$'si yüksek, HKO'sı düşük olan ve az sayıda açıklayıcı değişken içeren model tercih edildi (Montgomery ve ark., 2001).

Bazı durumlarda ise mevcut açıklayıcı değişkenlerden bazıları yanıt değişkendeki toplam değişimi açıklamada etkileri ya da katkıları olmadığından çoklu lineer regresyon modelinden silinebilir ya da çıkarılabilir. Çoklu lineer regresyon modelindeki açıklayıcı değişken sayısının artması durumunda ileriye doğru seçim ya da geriye doğru ayıklama gibi adımsal regresyon analizi yöntemleri uygulandı (Chatterjee ve ark., 2000).

Regresyonda değişken seçimi ya da en iyi modelin oluşturulmasıyla ilgili çalışmalar Hocking (1972, 1976), Cox ve Snell (1974), Myers(1990), Hocking ve LaMotte (1973), Thompson (1978a, 1978b) tarafından yapılmıştır. Alt küme modelindeki regresyon katsayılarının ve $\hat{\sigma}^2$ tahminlerinin özellikleri, Hocking (1974, 1976), Narula ve Ramberg (1972), Walls ve Weeks (1969) tarafından araştırılmıştır.

Bazı regresyon modelleri tarihsel kayıtlardan alınmış rastgele verilerden oluşmaktadır (Box ve ark., 1978). Rastgele veriler genelde aykırı ya da sapan değerleri, etkin gözlemleri, veri toplamadaki değişimlerden oluşan tutarsızlıkları ve zamana karşı bilgi-işlem sistem hatalarını içerir. Verideki bu hatalar, değişken seçimi sürecinde büyük etki yaratabilir ve doğru modeli belirleyememe problemine neden olabilir. Rastgele verideki en genel problem, kontrol edilmiş regresörleri bulmaktır. Kontrol edilmiş regresörler daha tutarlıdır. Ayrıca etkili değişkenlerdir. Regresörler, yanıtı doğru sınırlarda tutmak için kontrol edilmelidir. Verideki bu rastgele hataların etkileri en küçük kareler uyumunda önemsiz görünebilir.

Regresyon analizinde değişken seçimi probleminin iki aşamalı çözümü vardır. Birinci aşamada alt küme modelleri üretilir. İkinci aşamada ise bir alt kümenin diğerinden daha iyi olup olmadığına karar verilir (Berk, 1978).

Regresyon modelinin uygunluğun bir ölçütü, R^2 çoklu belirleyicilik katsayısıdır (Montgomery ve ark., 2001). Modelde β_0 sabit terimi bulunan p terimli ve $p-1$ regresörlü bir alt küme modeli için çoklu belirleyicilik katsayısı R_p^2 ile gösterildi (Montgomery ve ark., 2001). Bir alt küme regresyon modeli için R^2'nin optimum değerini arama yerine R^2'nin memnun edici ve beklentileri karşılayan değeri aranmalıdır. Aitkin(1974), bu probleme bir çözüm olarak tam model için R^2'den anlamlı olarak farklı olmayan yeni bir çoklu belirleyicilik katsayısı oluşturdu. R^2'nin yorumlanmasındaki zorluklardan kaçınmak için, bazı araştırmacılar düzeltilmiş R^2'yi ya da $\overline{R}_p^2$ kullanmayı tercih etmişlerdir (Haitovski 1969). $\overline{R}_p^2$ istatistiği, R_p^2 istatistiğine göre daha tutarlıdır. $\overline{R}_p^2$, modele eklenen yeni regresörlerden fazla etkilenmez (Edwards 1969; Seber 1977). Modele s tane

regresör eklenirse $\overline{R}^2_{p+s}$'nin değerinin, $\overline{R}^2_p$'nin değerinden daha fazla olabilmesi için gerek ve yeter koşul modele eklenen s tane regresörün önemini test etmek için kısmi F-istatistiğinin değerinin 1'i aşmasıdır (Edwards 1969). Sonuç olarak bir alt kümenin oluşturduğu optimum modeli seçmek için bir kriter, maksimum $\overline{R}^2_p$'ye sahip modeli seçmektir (Montgomery ve ark., 2001).

Bu kritere denk olan başka bir kriter, alt küme regresyon modeli için hata kareler ortalamasıdır, yani $MSE(p)$ dir (Montgomery ve ark., 2001). Mallows (1964, 1966, 1973) çalışmalarında, oluşturulan modelin değerinin hata kareler ortalamasına dayalı bir kriter önermiştir. Görüldüğü gibi altküme regresyon modellerini değerlendirmek için kullanılan bir çok kriter vardır. Model seçimi için kullanılacak kriter kesinlikle modelin kastedilen kullanımıyla ilgili olmalıdır. Regresyonun; (1) veri tanımlama, (2) kestirim ve tahmin, (3) parametre tahmini ve (4) kontrol olmak üzere bir çok olası kullanımı vardır. Amaç, verilmiş bir yöntem için iyi bir tanımlama elde etmekse veya karmaşık bir sistemin modelini elde etmekse, hata kareler toplamı küçük olan regresyon denklemleri için bir araştırma gösterilmiştir. k tane aday regresörlerin tamamını kullanarak hata kareler toplamı SSE minimum yapıldığında, SSE sonuçlarında küçük artmalar olabileceğinden bazı değişkenlerin modelden çıkarılması, silinmesi ya da elenmesi önerilir. y'deki toplam değişim açıklanırken, birkaç regresörlü sistemin uygun olduğu söylenir (Boyce ve ark., 1974).

Çoğu kez regresyon denklemleri, gözlemlerin ön tahmini veya yanıtın ortalamasının tahmini için kullanılır. Genel olarak, kestirimin hata kareler ortalamasının minimum yapıldığı regresörler seçilir. Bu da az etkili regresörlerin modelden silineceği anlamına gelir. Bir alt küme üretme yöntemi tarafından oluşturulan aday denklemleri değerlendirmek için $PRESS_p$ istatistiği kullanıldı (Chatterjee ve ark., 2000; Montgomery ve ark., 2001). $PRESS_p$'nin küçük değerine dayalı bir alt küme regresyon modeli seçilebilir. $PRESS_p$, özellikle tahmin problemi için sezgisel başvurmaya sahip olduğundan, hata kareler toplamının basit bir fonksiyonu değildir. Bu kritere dayalı değişken seçimi için bir algoritma

geliştirilmesi kolay değildir. $PRESS_p$ istatistiği alternatif modelleri ayırt etmede kullanışlıdır.

Parametre tahminiyle ilgileniliyorsa hem değişken silmenin sonucundaki yanlılıklar, hem de tahmin edilen katsayıların varyansları göz önünde bulundurulmalıdır. Regresörler yüksek içilişkili olduğunda, regresyon katsayılarının en küçük kareler tahminleri son derece zayıf olur. Regresyon modeli kontrol için kullanıldığında, parametrelerin doğru tahminleri çok önemlidir. En son denklemde kullanmak için değişkenlerin alt kümesini bulmada, aday regresörlerin çeşitli kombinasyonlarıyla model oluşturma dikkate alınmalıdır. Olası bütün regresyonlar yönteminde, sabit terimli model (Bu model regresör içermemektedir.), bir-aday regresör içeren model, iki-aday regresör içeren model, ..., k-aday regresör içeren model gibi bütün regresyon denklemlerinin oluşturulması gerekir (Montgomery ve ark., 2001).

Oluşturulan bu aday modeller değişik kriterlere göre değerlendirilir ve "en iyi regresyon modeli" seçilir. β_0 sabit teriminin bütün denklemlere dahil edildiğini varsayalım. k tane regresör için 2^k tane toplam aday regresyon denklemi vardır. R_p^2, $\overline{R}_p^2$, $MSE(p)$ ve C_p istatistiklerinin değerlerine bakılarak "en iyi model" belirlenir.

Tüm aday regresyon modellerini hesaplayıp değerlendirmek zor olabileceğinden, sadece az sayıdaki alt küme regresyon modellerini değerlendirmek için her seferinde bir tane regresör ekleyerek veya çıkararak yapılan çeşitli yöntemler geliştirilmiştir (Miller, 1990). Bu yöntemler adımsal türdeki yöntemlere ilişkindir. Bunlar üç ana grupta sınıflandırılabilir: Birincisi ileriye doğru seçim yöntemi (Montgomery ve ark., 2001). İkincisi geriye doğru ayıklama yöntemi (Montgomery ve ark., 2001). Üçüncüsü, 1-inci ve 2-inci yöntemin birleşimi olan adımsal regresyondur (Montgomery ve ark., 2001).

Çoklu lineer regresyon analizinde, lojistik regresyon analizinde ya da sıralı lojistik regresyon analizinde olduğu gibi regresyon tipi modellerde model oluşturma ve açıklayıcı değişkenlerin uygun alt küme seçimi veri madenciliğinde merkezi ve önemli bir problemdir. Çoğu kez açıklayıcı değişkenlerin bir alt kümesi verildiğinde

bir miktarsal, ikili veya sıralı düzeyde yanıt değişken ile çalışır. Böyle durumlarda açıklayıcı değişkenlerin hangilerinin yanıt değişkendeki toplam değişimi açıklamak ve regresyon katsayılarını yorumlamak için kullanılacağını belirlemek önemlidir.

İstatistiksel analiz için bir çok istatistiksel paket program, en iyi alt küme modelini seçmek için geriye doğru ayıklama ve ileriye doğru seçim gibi adımsal seçim yöntemlerini sağlar / içerir (Wilkinson, 1989). Bununla birlikte, regresyon analizinde geriye doğru ayıklama ve ileriye doğru seçim adımsal yöntemlerinin her ikisi de k değişkeninin bir kümesinden açıklayıcı (predictor) değişkenlerin en iyi alt kümesini her zaman bulmaz (Mantel, 1970). Geriye doğru ayıklama ve ileriye doğru seçim adımsal yöntemleri hakkında en önemli kritikler ya da eleştiriler: Birincisi, algoritmada hangi değişkenlerin modele dahil edileceği veya modelden çıkarılacağı sıralaması için teorik düzenlemelerin bulunmaması / olmaması (Boyce ve ark., 1974; Wilkinson 1989). İkincisi, analizde modele dahil edilecek ve modelden çıkarılacak değişkenlerin prior olasılıklarının seçimiyle ilgili herhangi bir teorik düzenlemelerin bulunmaması/ olmamasıdır. Üçüncüsü, adımsal arama arasıra da olsa en iyi modeli veya özel bir boyuttaki en iyi alt kümeyi bulur (Mantel 1970; Hocking 1976, 1983; Moses 1986). Dördüncüsü, yerel araştırmaya başvurulduğundan adımsal seçim geniş çözüm uzayının küçük bir alanından oldukça sınırlı bir örneklemi sağlar. Adımsal seçim en iyisinde sadece yeterli modeli oluşturur (Sokal ve Rohlf, 1981).

2.3. En İyi Çoklu Lineer Regresyon Modelinin Seçilmesi İle İlgili Çalışmalar – Genetik Algoritma

Çoklu lineer regresyon modelindeki açıklayıcı değişken sayısının fazla olması durumunda ne geleneksel yöntemler ne de adımsal yöntemler kullanılamamaktadır (Wasserman ve Sudjianto 1994; Bozdogan 2003). Çoklu lineer regresyon modelindeki açıklayıcı değişken sayısının fazla olması durumunda çoklu lineer regresyon modelinin oluşturulmasında Genetik Algoritma kullanılabilir (Wallet ve ark., 1996).

Çoklu lineer regresyon modelindeki açıklayıcı değişken sayısı k'nın fazla olması durumunda, örneğin $k = 10$ olsun. Aday model sayısı $2^k = 2^{10} = 1024$

olacaktır. Bu durumda ne geleneksel yöntemler, ne de adımsal yöntemler kullanılamamaktadır. Çoklu lineer regresyon modelindeki açıklayıcı değişken sayısının fazla olması durumunda çoklu lineer regresyon modelinin oluşturulmasında Genetik Algoritma kullanılabilir (Wasserman ve Sudjianto, 1994; Wallet ve ark., 1996).

Genetik algoritmalar (GA); evrim, gelişim ya da değişim hesaplamalarının bir parçasıdır. Genetik algoritmalar, Darwin'in evrim teorisinden esinlenerek oluşturulur. Genetik algoritmalar, yapay zekanın çok hızlı gelişen bir alanıdır (Goldberg, 1989). Genetik algoritmalar, genellikle bir problemin çözümünü kolaylaştırmak için kullanılır. Bir problemin çözümünde genetik algoritmaların kullanılması ilk defa John Holland tarafından ortaya atılmıştır. Daha sonra kendisi, öğrencileri ve meslektaşları tarafından geliştirilmiştir. John Holland, bu çalışmalar sonucunda 1975 yılında "Doğal ve Yapay Sistemlerde Adaptasyon / Uyum" (Adaption in Natural and Artificial Systems) adlı kitabı yazmıştır (Bozdogan, 2003). 1992 yılında John Koza, belirli işleri yapabilmek veya yerine getirmek amacıyla, program geliştirmek için genetik algoritmayı kullandı. Bu yöntemini de "Genetik Programlama (GP)" olarak adlandırdı. Genetik programlamada LISP (LISt Processing) programlama dili kullanıldı. Bunun nedeni LISP programlama dilinin, genetik algoritmalarda da kullanılan soyağacı (parse tree) yapısını daha kolay ve etkin işleyebilmesidir. Her problemin çözümünde problemin yapısına göre bir genetik algoritma oluşturulabilir (Michalewicz, 1992). Genetik algoritma, bir problemin çözümü için bir yöntem değildir. Bununla birlikte genetik algoritma bir problemin çözümünü elde etmek için izlenen yol olarak ifade edilebilir (Bauer, 1994).

Genetik algoritmalar, popülasyon (erişkinlerin ya da yetişkinlerin bir kitlesi) denilen ve kromozomlar ile gösterilen çözümlerin bir kümesi ile başlatılır. Bir popülasyondan çözümler alınır. Bu çözümler daha sonra yeni bir popülasyon oluşturmak için kullanılır. Bu işlem yeni popülasyonun eski popülasyondan daha iyi olacağı varsayımından hareketle yapılır. Yeni çözümleri (nesilleri) oluşturmak için seçilen çözümler uygunluk ya da uyumluluk değerlerine göre seçilir (Goldberg, 1989).

Genetik algoritmada kullanılan işlemler yeni bir populasyonu oluşturmak için kullanılır. Bu işlemler tamamen uygunluk fonksiyonuna bağlı olarak gerçekleşir.

Regresyon analizinde mevcut problemlerin yukarıdaki açıklamalarına dayalı olarak bu çalışmanın amacı çoklu regresyon modellerinde alt küme seçim için bilgi tabanlı model seçim kriterine ve genetik algoritmaya (GA) dayalı hesaplama bakımından uygulanabilir akıllı veri madenciliğini tanıtmak ve geliştirmektir (Bearse ve Bozdogan, 2002). Bu yaklaşım aynı zamanda üç yönlü hibrid olarak lojistik regresyon ve sıralı lojistik regresyon modellerine genişletilebilir. Sıralı lojistik regresyon modellerinde en iyi açıklayıcı değişkenlerin alt küme seçimi için de kullanılabilir (Lanning ve Bozdoğan, 2003).

Bir genetik algoritma geniş sayıda mümkün/olası çözümlerin bulunduğu problemin çözümüne uygulanabilen ve biyolojik değişim/dönüşüm ve doğal seçime dayalı stokastik (rastgele) arama algoritmasıdır. Genetik algoritmalar mühendislik, ekonomi, oyun teorisi (Holland, 1992), hesaplama bilimleri (Forrest, 1993), pazarlama (Bauer, 1994) ve biyoloji (Sumida ve ark., 1990) gibi geniş bir alanda kullanılabilir. Geleneksel optimizasyon yaklaşımından farklı olarak genetik algoritma amaç fonksiyonunun gradyanlarını hesaplamaya gereksinim duymaz ve bir yerel optimuma sınırlanmaz (Goldberg, 1989). Bir genetik algoritma bir ikili string halindeki kodların bir dizisi olarak bilgilendirir. İkili stringler verilen probleme farklı çözümleri gösterir. Bu stringler bir kromozom üzerindeki genler tarafından kodlanan genetik bilgiye analog modellerdir. Bir string problemi çözmek için özel yeteneği için uyum / uygunluk değerlerine göre hesaplanabilir. Uyum değerleri tabanında stringler, her bir çalıştırmadan sonra ve analizde problemin çözümü için kullanılır ya da atılır. Bir çok çalıştırmadan sonra en iyi çözüm belirlenir / tespit edilir. Herhangi bir genetik algoritmadaki zorluk, her bir çözümü hesaplamak için temel olarak uygun bir uyum fonksiyonunun seçimidir.

Çoklu regresyon analizine göre uyum değeri en iyi alt kümenin araştırılmasında alt küme modellerin karşılaştırılması için bir alt küme seçim kriteridir. Bu bilgisel model seçimi kriteri kullanılarak kolaylıkla belirlenebilir.

Genel olarak istatistiksel modelleme ve model hesaplama problemlerinde model karmaşıklığı kavramı önemli bir rol oynar. Karmaşıklık bağlantı yapıları

olarak tasarımlar ve model bileşenlerinin etkileşimlerini içerir. "Genel" model karmaşıklığının bir ölçümü olmaksızın model davranışını tahmin etmek ve modelin kalitesini değerlendirmek zordur. Bu detaylı istatistiksel analize ve verilen sonlu bir örneklem için yarışan modellerin tümü arasında en iyi modeli seçmek için hesaplamalara gereksinim duyar.

Yakın zamanda Akaike'nin (1973) orijinal AIC'kine dayalı bir çok model-seçimi prosedürü önerilmiştir (Sclove, 1987). Model seçiminde AIC'in kullanılmasında Akaike'ye (1987) göre parametre tahminlerinin doğruluğu bir genel kriter ile ölçülür. AIC'dekine benzer şekildeki incelemelerden hareketle işlemler yapılmıştır. Bununla birlikte yeni prosedür, Van Emden (1971)'in bilgi-tabanlı kovaryans karmaşıklık indeksinin bir genelleştirilmesi yoluyla bir elemanın veya rasgele vektörlerin yapısal karmaşıklığı üzerinde *ICOMP*'a dayandırılmıştır. *ICOMP*'un oluşturulması ve geliştirilmesi orijinal olarak Van Emden (1971) tarafından tanımlanan kovaryans karmaşıklık indeksinin bir genelleştirmesine dayalıdır. Direk olarak serbest parametrelerin sayısını cezalandırma yerine *ICOMP* modelin kovaryans karmaşıklığını cezalandırır. *ICOMP*'un en genel formu *ICOMP*(*IFIM*) dir (Bozdogan, 2003). *ICOMP*(*IFIM*), maksimum likelihood tahminlerin iyi-bilinen asimptotik optimallik özelliğini açıklar ve bir modelin inverse-fisher bilgi matrisinin (*IFIM*) bilgi tabanlı karmaşıklığını kullanır. Bu, Cramér-Rao alt sınır matrisi olarak bilinir (Cramér 1946; Rao 1945, 1947, 1948).

Karmaşıklık, istatistiksel modellerin bir genel özelliğidir ve modellerin olasılık tanımlarından / özelliklerinden, yapısından veya özel içeriğinden çoğunlukla bağımsızdır. Literatürde, karmaşıklık kavramı bir çok değişik içerikte kullanılmıştır. Van Emden (1971)'e göre tasarım anlaşılması zor olduğundan genel olarak istatistikte karmaşıklığın tek bir tanımı yoktur. Karmaşıklığın bir çok yönü vardır ve "Kolmogorov karmaşıklığı" (Cover ve ark., 1989), "Shannon Karmaşıklığı" (Rissanen 1987, 1989) gibi bir çok adlar altında tanımlanır. Bilgi teorik kodlama teorisinde Rissanen (1986, 1987, 1989), modellerin sınıfları tarafından ortaya çıkarılabilen veriler için en kısa kod uzunluğu cinsinden karmaşıklığı tanımlayan Kolmogorov (1983)'dekine benzer şekilde karmaşıklığı tanımlamıştır ve onu Stokastik Karmaşıklık (SC) olarak adlandırmıştır. Wallace ve Freemen (1987),

Wallace ve Dowe (1993) ve Baxter (1996) karmaşıklığı, Minimum Mesaj Uzunluğu (Minimum Message Length-MML) cinsinden tanımlamıştır. Minimum mesaj uzunluğu, veriyi kapsayan bir mesajı sıkıştırma yeteneğine göre modellerin hesaplanmasına dayalıdır.

Karmaşıklığın anlaşılması ve verinin ışığında belirsizliğini çalışmak için (tümevarımsal) sonuç çıkarmak genel model oluşturma teorisinde çok gereklidir. İstatistiksel modeller ve yöntemler tam olarak tümdengelimli değildir. Çünkü insanlar çoğu zaman belirsizlik durumunda sonuç çıkarır. Tümevarımsal sonuç çıkarma, bir hipotezden veya model uzayından bir parametreyi veya bir modeli seçme problemidir. Çalışılan veriyi en iyi açıklar (Baxter, 1996). Akaike (1994)'de incelendiği gibi belirsizlik altında sonuç çıkarma Pierce (1955) tarafından çalışılmıştır. Pierce bunu kaçırma mantığı ya da kısaca kaçırma olarak adlandırdı. Kaçırma, sonuç çıkarmanın bir yoludur, genel prensipler ve yeni gerçekleri elde etmek için gözlenen gerçekleri kullanır. Hepsinin bir belirsizlik derecesi vardır. Kaçırma nümerik fonksiyonları kullanarak yerini alır ve bilgi teorik model seçim kriteri gibi büyüklükleri ölçer. Pierce bilimsel çalışmanın orijinal bölümünün çoğunun kaçırma aşamasıyla veya uygun hipotezlerin seçim aşamasıyla ilgili olduğu konusunda ısrar etmiştir. Bu nedenle karmaşıklığın tasarımı yardımıyla kaçırmaya dayalı sonuç çıkarma için bir sistematik prosedür geliştirmek öğrenme ve değişim/evrimleşme işlemini anlamak öncelikle yapılması gereken işlemdir (Von Neumann, 1966). Bu çerçevede istatistiksel modelleme ve model oluşturma kaçırma bilimidir. Bu nedenle karmaşıklığın çalışılması uygun hipotezlerin model seçimi veya veri madenciliği işi içinde modeller için oldukça önemlidir.

Çoklu lineer regresyon analizi için genel olarak genetik algoritmada kullanılan uyum fonksiyonu için bir model seçim kriteri kullanılır. Bu çalışmada karmaşık bilgi kriteri *ICOMP* (Information COMPlexity) kriteri kullanılacaktır. Analizciler ya da araştırmacılar gereksinimlerine veya önceliklerine dayalı olarak herhangi bir uygun model seçim kriterini seçebilirler.

Bu adım eşleştirme ya da çiftleştirme havuzunda (mating pool) birleştirme için modellerin *ICOMP*(*IFIM*) değerlerine dayalı olarak modelleri seçme işleminden oluşur. Burada *IFIM* (Inverse Fisher Information Matrix), modellerin

inverse Fisher bilgi matrislerini göstermektedir (Bozdogan, 2003). Popülasyonda ya da erişkinlerin veya yetişkinlerin oluşturduğu N tane modelde olası altküme modellerinden her biri için $ICOMP(IFIM)$ değerleri hesaplanır.

3. ÇOKLU LİNEER REGRESYON MODELİ

Bir yanıt (bağımlı) değişkendeki toplam değişimi açıklamak amacıyla birden fazla regresör (açıklayıcı) değişken kullanılarak oluşturulan regresyon modeline çoklu regresyon modeli denir. Bu bölümde çoklu regresyon modellerinin oluşturulması ve analizi incelenecektir. Ayrıca çoklu regresyon modelinin yeterlik ölçüleri ele alınacaktır.

3.1. Çoklu Lineer Regresyon Modeli Hakkında Genel Bilgiler

Bir testerenin etkin yaşam süresi; testerenin kesme hızına ve kesme derinliğine bağlı olsun (Montgomery ve ark., 2001). Yanıt değişken y: testerenin etkin yaşam süresini; açıklayıcı değişkenler x_1: kesme hızını ve x_2: kesme derinliğini göstermektedir. Bu durumda yanıt değişken ile açıklayıcı değişkenler arasındaki ilişkiyi açıklayan çoklu regresyon modeli,

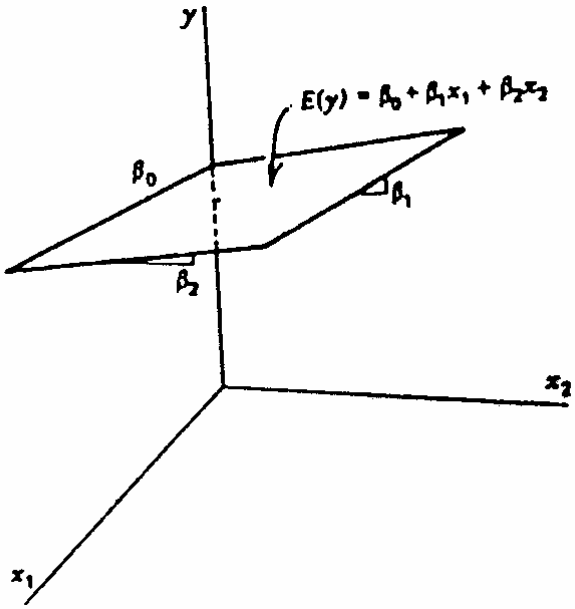
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon \qquad (3.1)$$

şeklindedir. Burada y, yanıt değişkeni; x_1 ve x_2, açıklayıcı değişkenleri; β_0, β_1 ve β_2 bilinmeyen parametreleri ya da regresyon katsayılarını ve ε, hata terimini göstermektedir.

Yanıt değişken ile açıklayıcı değişkenler arasındaki ilişkiyi açıklayan çoklu regresyon modeli β_0, β_1 ve β_2 bilinmeyen parametrelerinde ya da regresyon katsayılarında lineer olduğundan bu modele, çoklu lineer regresyon modeli de denilir.

(3.1)'deki eşitlikteki çoklu lineer regresyon modeli iki boyutlu uzayda bir düzlem belirtir (Montgomery ve ark., 2001). β_0 parametresi, regresyon düzleminin sabitidir. x_1 ve x_2 açıklayıcı değişkenlerinin değişim aralığı $x_1 = x_2 = 0$ değerini içeriyorsa $\beta_0 = \bar{y}$ olur. Diğer durumda β_0 'ın hiçbir fiziksel açıklaması yoktur. β_1,

x_2 sabit tutulduğunda x_1'deki bir birimlik değişim nedeniyle yanıt değişkende olabilecek ya da beklenen değişim miktarını belirtir. Benzer biçimde, β_2, x_1 sabit tutulduğunda x_2'deki bir birimlik değişim nedeniyle yanıt değişkende olabilecek ya da beklenen değişim miktarını belirtir.



Şekil 3.1. İki boyutlu uzayda çoklu lineer regresyon modeli bir regresyon düzlemi belirtir (Montgomery ve ark., 2001).

Genel olarak, yanıt değişken y, k tane açıklayıcı değişken x_1, x_2, ..., x_k ile ilişkili olabilir. Bu nedenle,

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + \varepsilon \qquad (3.2)$$

eşitliği k tane açıklayıcı değişkenli çoklu lineer regresyon modeli olarak adlandırılır. Burada β_0, β_1, β_2, ..., β_k parametreleri regresyon katsayılarını ve ε hata terimini göstermektedir. Bu model $j = 1, 2, ..., k$ olmak üzere x_j açıklayıcı değişkenlerinin k-boyutlu uzayında bir hiper düzlem belirtir (Montgomery ve ark., 2001). β_j parametresi, j. açıklayıcı değişken dışındaki tüm açıklayıcı değişkenler sabit

tutulduğunda, x_j 'deki bir birimlik değişim nedeniyle yanıt değişken y 'de olabilecek ya da beklenen değişim miktarını belirtir. Bu nedenle, $\beta_0, \beta_1, \beta_2, ..., \beta_k$ parametreleri genelde kısmi regresyon katsayıları olarak adlandırılır (Draper ve Smith, 1998).

Çoklu lineer regresyon modelleri genelde regresyon probleminin çözümüne yaklaşım fonksiyonu olarak kullanılırlar. y yanıt değişkeni ile $x_1, x_2, ..., x_k$ açıklayıcı değişkenleri arasındaki gerçek fonksiyonel ilişki bilinmez. Bununla birlikte açıklayıcı değişkenlerin lineer regresyon modeli, bir uygun yaklaşım olarak kullanılır. Yapısal olarak (3.2)'den daha karmaşık modeller de, çoklu lineer regresyon yöntemleri kullanılarak analiz edilebilir. Örneğin,

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \varepsilon \tag{3.3}$$

kübik polinom modelinde $x_1 = x$, $x_2 = x^2$ ve $x_3 = x^3$ olarak alınsın. Bu durumda (3.3)'teki eşitlikteki model,

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon \tag{3.4}$$

şekilde yazılabilir. (3.4)'deki eşitlik üç açıklayıcı değişken içeren çoklu lineer regresyon modelidir. Etkileşim terimleri içeren modeller de çoklu lineer regresyon yöntemleri kullanılarak analiz edilebilir. Örneğin etkileşim terimi içeren model,

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \varepsilon \tag{3.5}$$

şeklinde olsun. Etkileşim terimi içeren modelde $x_3 = x_1 x_2$ ve $\beta_3 = \beta_{12}$ olarak alınırsa (3.5)'teki etkileşim terimi içeren model,

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon \tag{3.6}$$

şekilde yazılabilir. (3.6)'daki eşitlik yine üç açıklayıcı değişken içeren çoklu lineer regresyon modelidir. Genel olarak β parametrelerinde ya da regresyon katsayılarında lineer olan herhangi bir regresyon modeli, oluşturduğu yüzeyin şekli ne olursa olsun, bir lineer regresyon modelidir (Montgomery ve ark., 2001).

3.2. Çoklu Lineer Regresyon Modelindeki Parametrelerinin Tahmin Edilmesi

Bu kısımda çoklu lineer regresyon modelindeki regresyon katsayılarını tahmin etmek için kullanılan En Küçük Kareler (EKK) ve En Çok Olabilirlik (EÇO) yöntemleri incelenecektir.

3.2.1. Regresyon Katsayılarının En Küçük Kareler Yöntemiyle Tahmin Edilmesi

En küçük kareler yöntemi, (3.2)'deki eşitlikteki çoklu lineer regresyon modelinin regresyon katsayılarını tahmin etmek için kullanılır (Draper ve Smith, 1998). n ve k sırasıyla gözlem sayısını ve regresyon modelindeki açıklayıcı değişken sayısını göstersin. $n > k$ olmak üzere n tane gözlemin bulunduğunu varsayalım. y_i, i-inci gözlenmiş yanıt değerini ve x_{ij}, j-inci açıklayıcı değişkeninin i-inci gözlenmiş değerini ya da seviyesini göstersin. Veriler, Tablo 3.1'deki gibi düzenlenir.

Tablo 3.1. Çoklu lineer regresyon modeli için veriler.

Gözlem i	y	x_1	x_2	$\cdots$	x_k
1	y_1	x_{11}	x_{12}	$\cdots$	x_{1k}
2	y_2	x_{21}	x_{22}	$\cdots$	x_{2k}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	y_n	x_{n1}	x_{n2}	$\cdots$	x_{nk}

Çoklu regresyon modelindeki ε hata terimi için $E(\varepsilon)=0$, $V(\varepsilon)=\sigma^2$ ve hataların ilişkisiz olduğu yani $i \neq j$ için $Cov(\varepsilon_i, \varepsilon_j)=0$ varsayımı yapılır (Montgomery ve ark., 2001).

(3.2)'deki eşitlikteki çoklu lineer regresyon modeline karşılık gelen örneklem modeli açık biçimde,

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik} + \varepsilon_i \qquad i = 1,2,...,n \tag{3.7}$$

şeklinde ya da kapalı biçimde,

$$y_i = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} + \varepsilon_i \qquad i = 1,2,...,n \tag{3.8}$$

şeklinde kapalı biçimde yazılabilir.

β_0, β_1, β_2, ..., β_k parametreleri ya da regresyon katsayıları cinsinden en küçük kareler fonksiyonu $S(\beta_0, \beta_1, ..., \beta_k)$,

$$S(\beta_0, \beta_1, ..., \beta_k) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{k} \beta_j x_{ij} \right)^2 \tag{3.9}$$

olarak tanımlanır. En küçük kareler yönteminde: S fonksiyonunu, $\beta_0, \beta_1, ..., \beta_k$ parametrelerine göre minimum yapan $\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_k$ parametre tahminleri hesaplanır.

$\beta_0, \beta_1, ..., \beta_k$ parametrelerinin sırasıyla $\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_k$ en küçük kareler tahmin edicileri,

$$\left.\frac{\partial S}{\partial \beta_0}\right|_{\hat{\beta}_0,\hat{\beta}_1,\ldots,\hat{\beta}_k} = -2\sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \sum_{j=1}^{k}\hat{\beta}_j x_{ij}\right) = 0 \tag{3.10}$$

ve

$$\left.\frac{\partial S}{\partial \beta_j}\right|_{\hat{\beta}_0,\hat{\beta}_1,\ldots,\hat{\beta}_k} = -2\sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \sum_{j=1}^{k}\hat{\beta}_j x_{ij}\right)x_{ij} = 0, \tag{3.11}$$

eşitliklerini sağlarlar. (3.10) ve (3.11)'deki eşitlikler sadeleştirilerek,

$$\begin{aligned}
&n\hat{\beta}_0 + \hat{\beta}_1\sum_{i=1}^{n}x_{i1} + \hat{\beta}_2\sum_{i=1}^{n}x_{i2} + \ldots + \hat{\beta}_k\sum_{i=1}^{n}x_{ik} = \sum_{i=1}^{n}y_i \\
&\hat{\beta}_0\sum_{i=1}^{n}x_{i1} + \hat{\beta}_1\sum_{i=1}^{n}x_{i1}^2 + \hat{\beta}_2\sum_{i=1}^{n}x_{i1}x_{i2} + \ldots + \hat{\beta}_k\sum_{i=1}^{n}x_{i1}x_{ik} = \sum_{i=1}^{n}x_{i1}y_i \\
&\qquad\qquad\vdots \qquad\qquad\qquad\qquad\qquad\qquad\vdots \\
&\hat{\beta}_0\sum_{i=1}^{n}x_{ik} + \hat{\beta}_1\sum_{i=1}^{n}x_{ik}x_{i1} + \hat{\beta}_2\sum_{i=1}^{n}x_{ik}x_{i2} + \ldots + \hat{\beta}_k\sum_{i=1}^{n}x_{ik}^2 = \sum_{i=1}^{n}x_{ik}y_i
\end{aligned} \tag{3.12}$$

biçiminde en küçük kareler normal denklemleri elde edilir. Bilinmeyen regresyon katsayılarının her biri için birer tane olmak üzere toplam $p = k+1$ tane normal denklem vardır. (3.12)'deki eşitlikteki normal denklemlerin çözümü, $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ en küçük kareler tahminlerini verir.

3.2.2. Çoklu Lineer Regresyon Modelinde Matris Gösteriminin Kullanılması ve Regresyon Katsayılarının En Küçük Kareler Yöntemiyle Tahmin Edilmesi

Çoklu lineer regresyon modeli matris gösterimiyle,

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon} \tag{3.13}$$

şeklinde yazılabilir (Hocking 1976, 1983; Miller 1990). Burada $\boldsymbol{y}=\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$, $n\times 1$ tipinde olmak üzere gözlemlerin vektörünü; $\boldsymbol{X}=\begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & \mathrm{x}_{1k} \\ 1 & \mathrm{x}_{21} & \mathrm{x}_{22} & \cdots & \mathrm{x}_{2k} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & \mathrm{x}_{n1} & \mathrm{x}_{n2} & & \mathrm{x}_{nk} \end{bmatrix}$, $n\times p$ tipinde olmak üzere açıklayıcı değişkenlerin düzeylerinin matrisini; $\beta=\begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{\boldsymbol{k}} \end{bmatrix}$, $p\times 1$ tipinde olmak üzere regresyon katsayılarının vektörünü ve $\varepsilon=\begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$, $n\times 1$ tipinde olmak üzere rastgele hataların vektörünü göstermektedir. (3.13)'teki matris formundaki çoklu lineer regresyon modelindeki β parametre vektörünün $\hat{\beta}$ en küçük kareler tahmin edicisi,

$$S(\boldsymbol{\beta})=\sum_{i=1}^{n}\varepsilon_i^2=\sum_{i=1}^{n}\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}=(\boldsymbol{y}-\boldsymbol{X\beta})'(\boldsymbol{y}-\boldsymbol{X\beta}) \tag{3.14}$$

fonksiyonunun minimum yapılmasıyla elde edilir. $S(\boldsymbol{\beta})$,

$$\begin{aligned} S(\boldsymbol{\beta}) &= \boldsymbol{y'y}-\boldsymbol{\beta'X'y}-\boldsymbol{y'X\beta}+\boldsymbol{\beta'X'X\beta} \\ &= \boldsymbol{y'y}-2\boldsymbol{\beta'X'y}+\boldsymbol{\beta'X'X\beta} \end{aligned} \tag{3.15}$$

şekilde yazılabilir. Burada $\boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{y}$, 1×1 tipinde matristir veya skalerdir. Bunun transpozu olan $(\boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{y})' = \boldsymbol{y}'\boldsymbol{X}\boldsymbol{\beta}$ de benzer biçimde 1×1 tipinde matristir veya skalerdir. $\boldsymbol{\beta}$ parametre vektörünün $\hat{\boldsymbol{\beta}}$ en küçük kareler tahmin edicisi,

$$\frac{\partial S}{\partial \boldsymbol{\beta}}\Big|_{\hat{\boldsymbol{\beta}}} = -2\boldsymbol{X}'\boldsymbol{y} + 2\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}} = 0 \tag{3.16}$$

eşitliğini sağlar. (3.16)'daki eşitlik yeniden düzenlendiğinde,

$$\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}'\boldsymbol{y} \tag{3.17}$$

eşitliği elde edilir. (3.17)'deki eşitlik, en küçük kareler normal denklemleridir. Normal denklemleri çözmek için, (3.17)'deki eşitliğin her iki tarafı $\boldsymbol{X}'\boldsymbol{X}$ matrisinin tersi ile çarpılır. Böylece, $\boldsymbol{\beta}$ parametre vektörünün $\hat{\boldsymbol{\beta}}$ en küçük kareler tahmin edicisi,

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} \tag{3.18}$$

olarak bulunur. Burada $\boldsymbol{X}$ açıklayıcı değişkenlerin matrisindeki sütunlar lineer olarak bağımsız iseler, yani $\boldsymbol{X}$ matrisinin hiçbir sütunu diğer sütunların lineer bir kombinasyonu değilse, $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ matrisi her zaman hesaplanabilir (Montgomery ve ark., 2001).

(3.12)'deki eşitlikteki normal denklemler, (3.17)'deki normal denklemlere benzerdir. Bu (3.17)'deki eşitlik ayrıntılı olarak yazıldığında görülebilir.

(3.17)'deki eşitlikteki $\boldsymbol{X}'\boldsymbol{X}$, $p\times p$ tipinde simetrik bir matristir. $\boldsymbol{X}'\boldsymbol{X}$ matrisi özel bir yapıya sahiptir. $\boldsymbol{X}'\boldsymbol{X}$ matrisinin ana köşegenindeki elemanlar, $\boldsymbol{X}$ matrisinin sütunlarındaki elemanların kareleri toplamıdır. $\boldsymbol{X}'\boldsymbol{X}$ matrisinin ana köşegeni dışındaki elemanlar, $\boldsymbol{X}$ matrisinin sütunlarındaki elemanların çapraz çarpımlarının

toplamıdır. $\boldsymbol{X'y}$ matrisinin elemanları, $\boldsymbol{X}$ matrisinin sütunlarındaki elemanlarla y_i gözlemlerinin çapraz çarpımlarının toplamıdır.

$$\begin{bmatrix} n & \sum_{i=1}^{n} x_{i1} & \sum_{i=1}^{n} x_{i2} & \cdots & \sum_{i=1}^{n} x_{ik} \\ \sum_{i=1}^{n} x_{i1} & \sum_{i=1}^{n} x_{i1}^2 & \sum_{i=1}^{n} x_{i1}x_{i2} & \cdots & \sum_{i=1}^{n} x_{i1}x_{ik} \\ \vdots & \vdots & \vdots & & \vdots \\ \sum_{i=1}^{n} x_{ik} & \sum_{i=1}^{n} x_{ik}x_{i1} & \sum_{i=1}^{n} x_{ik}x_{i2} & \cdots & \sum_{i=1}^{n} x_{ik}^2 \end{bmatrix} \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_k \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_{i1}y_i \\ \vdots \\ \sum_{i=1}^{n} x_{ik}y_i \end{bmatrix} \quad (3.19)$$

Açıklayıcı değişkenlerin $\boldsymbol{x'} = [1, x_1, x_2, ..., x_k]$ düzeyinde oluşturan regresyon modelinin değeri,

$$\hat{\boldsymbol{y}} = \boldsymbol{x'}\hat{\boldsymbol{\beta}} = \hat{\beta}_0 + \sum_{j=1}^{k} \hat{\beta}_j x_j \quad (3.20)$$

olarak elde edilir. Gözlenen y_i değerleriyle oluşturulan $\hat{y}_i$ değerleri arasındaki bağıntı ise,

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}(\boldsymbol{X'X})^{-1}\boldsymbol{X'y} = \boldsymbol{Hy} \quad (3.21)$$

şeklindedir. Burada $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X'X})^{-1}\boldsymbol{X'}$ matrisi 'şapka' matrisi olarak adlandırılır. $\boldsymbol{H}$ şapka matrisi, gözlenen y_i değerlerinin vektörünü oluşturulan $\hat{y}_i$ değerlerinin bir vektörü olarak düzenler. $\boldsymbol{H}$ şapka matrisi regresyon analizinde önemli rol oynar (Montgomery ve ark., 2001).

Gözlenen y_i değerleriyle oluşturulan $\hat{y}_i$ değerleri arasındaki fark

$$e_i = y_i - \hat{y}_i \quad (3.22)$$

dir. Buna rezidü ya da artık veya kalan denir. n tane rezidü matris gösterimiyle,

$$\boldsymbol{e} = \boldsymbol{y} - \hat{\boldsymbol{y}} \tag{3.23}$$

şeklinde yazılabilir. $\boldsymbol{e}$ rezidü vektörü,

$$\boldsymbol{e} = \boldsymbol{y} - \boldsymbol{X}\hat{\beta} = \boldsymbol{y} - \boldsymbol{H}\boldsymbol{y} \tag{3.24}$$

şeklinde ya da

$$\boldsymbol{e} = \left(\boldsymbol{I} - \boldsymbol{H}\right)\boldsymbol{y} \tag{3.25}$$

biçiminde ifade edilebilir.

Bir yanıt ve iki açıklayıcı değişken içeren çoklu lineer regresyon modelini oluşturma örneği, Örnek 3.1'de verilmiştir.

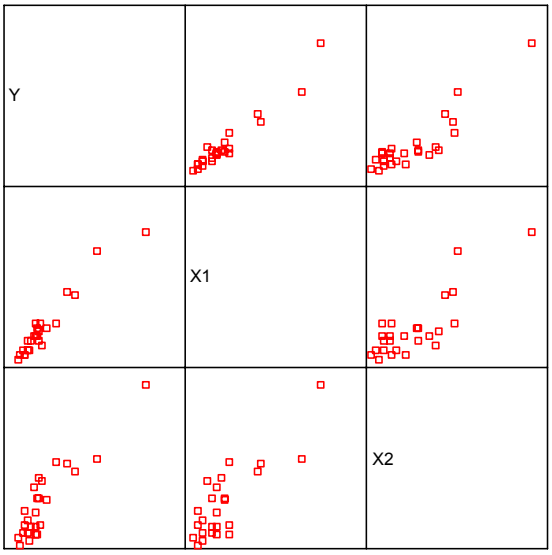
Örnek 3.1. Alkolsüz içecek firmasında çalışan bir endüstri mühendisi, içecek dağıtım sisteminde parayla çalışan makinelerin servis performansını analiz etmek istemektedir. Bu amaçla teslim süresiyle ilgilenmektedir. Teslim süresini etkileyen iki önemli değişken (faktör) olduğunu düşünmektedir. Bunlar ürün miktarı ve uzaklık olarak düşünülmektedir. Endüstri mühendisi bu amaçla 25 gözlem toplamıştır (Montgomery ve ark., 2001). Veriler Tablo 3.2'de verilmiştir.

Bu örnekte yanıt değişken y (teslim süresi)'deki toplam değişimi açıklamak için x_1 (ürün miktarı) ve x_2 (uzaklık) açıklayıcı değişkenleri kullanılacaktır. Bu amaçla veriye $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$ şeklinde çoklu lineer regresyon modeli oluşturulacaktır. Öncelikle yanıt değişken y (teslim süresi) ile açıklayıcı değişkenler x_1 (ürün miktarı) ve x_2 (uzaklık) arasındaki ilişkiyi görsel ya da grafiksel olarak kontrol etmek için matris grafiği oluşturulur.

Tablo 3.2. İçecek teslim/dağıtım verisi (Montgomery ve ark., 2001).

Gözlem No	Teslim süresi (Dakika)	Ürün miktarı (Kutu sayısı)	Uzaklık (Feet)	Gözlem No	Teslim süresi (Dakika)	Ürün miktarı (Kutu sayısı)	Uzaklık (Feet)
	y	x_1	x_2		y	x_1	x_2
1	16.68	7	560	14	19.75	6	462
2	11.50	3	220	15	24.00	9	448
3	12.03	3	340	16	29.00	10	776
4	14.88	4	80	17	15.35	6	200
5	13.75	6	150	18	19.00	7	132
6	18.11	7	330	19	9.50	3	36
7	8.00	2	110	20	35.10	17	770
8	17.83	7	210	21	17.90	10	140
9	79.24	30	1460	22	52.32	26	810
10	21.50	5	605	23	18.75	9	450
11	40.33	16	688	24	19.83	8	635
12	21.00	10	215	25	10.75	4	150
13	13.50	4	255				

Yanıt değişken y ile açıklayıcı değişkenler x_1 ve x_2 arasındaki ilişkiyi grafiksel olarak kontrol etmek için oluşturulan matris grafiği Şekil 3.2'de verilmiştir.



Şekil 3.2. İçecek teslim/dağıtım verisindeki yanıt değişken y ile açıklayıcı değişkenler x_1 ve x_2 arasındaki ilişki için oluşturulan matris grafiği.

İçecek teslim/dağıtım verisindeki yanıt değişken y ile açıklayıcı değişkenler x_1 ve x_2 arasındaki ilişki için analiz sonuçlarının bilgisayar çıktısı Tablo 3.3'te verilmiştir.

Tablo 3.3. İçecek teslim/dağıtım verisindeki yanıt değişken y ile açıklayıcı değişkenler x_1 ve x_2 arasındaki ilişki için analiz sonuçlarının bilgisayar çıktısı.

Correlations

		Y	X1	X2
Pearson Correlation	Y	1,000	,965	,892
	X1	,965	1,000	,824
	X2	,892	,824	1,000
Sig. (1-tailed)	Y	,	,000	,000
	X1	,000	,	,000
	X2	,000	,000	,
N	Y	25	25	25
	X1	25	25	25
	X2	25	25	25

Şekil 3.2'deki matris grafiğine ve Tablo 3.3'teki analiz sonuçlarının bilgisayar çıktısına göre y ile x_1, y ile x_2 ve x_1 ile x_2 arasındaki ilişki pozitif yönde lineerdir. İçecek teslim/dağıtım verisindeki y, x_1 ve x_2 için tanımlayıcı istatistiklerin analiz sonuçlarının bilgisayar çıktısı Tablo 3.4'te verilmiştir.

Tablo 3.4. İçecek teslim/dağıtım verisindeki y, x_1 ve x_2 için tanımlayıcı istatistiklerin analiz sonuçlarının bilgisayar çıktısı.

Descriptive Statistics

	Mean	Std. Deviation	N
Y	22,3840	15,5249	25
X1	8,76	6,88	25
X2	409,28	325,19	25

İçecek teslim/dağıtım verisindeki x_1 ve x_2 açıklayıcı değişkenlerinin y yanıt değişkendeki toplam değişimi açıklama oranı için analiz sonuçlarının bilgisayar çıktısı Tablo 3.5'te verilmiştir.

Tablo 3.5. İçecek teslim/dağıtım verisindeki x_1 ve x_2 açıklayıcı değişkenlerinin y yanıt değişkendeki toplam değişimi açıklama oranı için analiz sonuçlarının bilgisayar çıktısı.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,980[a]	,960	,956	3,2595

a. Predictors: (Constant), X2, X1

Tablo 3.5'teki sonuçlara göre içecek teslim/dağıtım verisindeki x_1 ve x_2 açıklayıcı değişkenlerinin y yanıt değişkendeki toplam değişimi açıklama oranı %96'dır.

İçecek teslim/dağıtım verisi için regresyonun önemliliği testinin varyans analizi tablosu Tablo 3.6'da verilmiştir.

Tablo 3.6. İçecek teslim/dağıtım verisi için regresyonun önemliliği testinin varyans analizi tablosu.

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	5550,811	2	2775,405	261,235	,000[a]
	Residual	233,732	22	10,624		
	Total	5784,543	24			

a. Predictors: (Constant), X2, X1

b. Dependent Variable: Y

İçecek teslim/dağıtım verisi için regresyonun önemliliği testinin varyans analizi tablosundaki sonuca göre y yanıt değişkendeki toplam değişimi açıklamada x_1 ve x_2 açıklayıcı değişkenlerine gereksinim vardır.

İçecek teslim/dağıtım verisi için oluşturulan regresyonun modelinde β parametre vektörünün (3.18)'deki $\hat{\beta}$ en küçük kareler tahmin edicisi kullanılarak elde edilen parametre tahmin değerleri Tablo 3.7'de verilmiştir.

Tablo 3.7. İçecek teslim/dağıtım verisi için oluşturulan regresyonun modelindeki parametre tahmin değerleri.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	2,341	1,097		2,135	,044	,067	4,616
	X1	1,616	,171	,716	9,464	,000	1,262	1,970
	X2	1,438E-02	,004	,301	3,981	,001	,007	,022

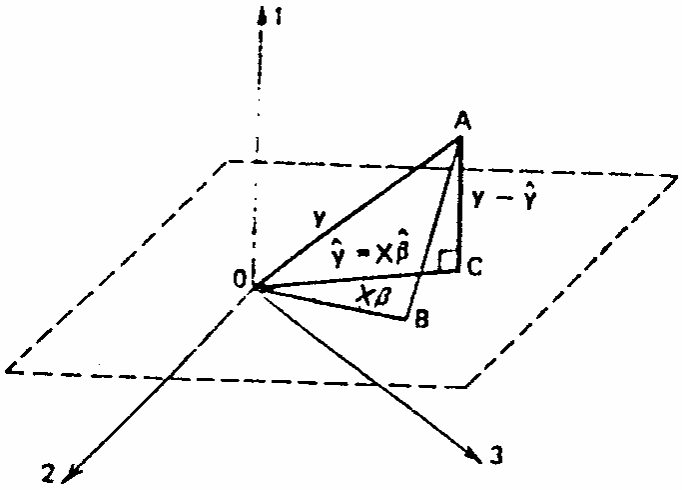
a. Dependent Variable: Y

Tablo 3.7'deki sonuçlara göre içecek teslim/dağıtım verisi için oluşturulan regresyonun modeli,

$$\hat{y} = 2.341 + 1.616x_1 + 0.014x_2 \tag{3.26}$$

dir.

3.2.3. En Küçük Kareler Yönteminin Geometrik Yorumu



Şekil 3.3. En küçük kareler yönteminin bir geometrik yorumu (Montgomery ve ark., 2001).

Gözlemlerin vektörü $\boldsymbol{y}' = [y_1, y_2, ..., y_n]$, Şekil 3.3'teki gibi orijinden A noktasına kadar tanımlansın. y_1, y_2 ,..., y_n'ler $n-$boyutlu örneklem uzayının

koordinatlarını oluşturmaktadır. Şekil 3.3'teki örneklem uzayı üç-boyutludur. $\boldsymbol{X}$ matrisi, $p = k + 1$ tane $(n \times 1)$ tipinde sütun vektöründen oluşur. Yani $\boldsymbol{X}$ matrisi $(n \times p)$ tipindedir. $\mathbf{1}$, 1'lerden oluşan sütun vektörü olsun. Yani $\mathbf{1}' = [1, 1, ..., 1]$ olsun. Bu durumda $\boldsymbol{X}$ matrisinin sütunları, $\mathbf{1}, \boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_k$ vektörlerinden oluşmaktadır. Örneklem uzayda bütün bu sütunlar orijinden bir vektör tanımlar. Bu p tane vektör tahmin uzayı da denilen p - boyutlu bir alt uzay oluşturur. $p = 2$ için tahmin uzayı Şekil 2'de gösterilmiştir. Bu alt uzaydaki herhangi bir nokta, $\mathbf{1}, \boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_k$ vektörlerinin bir lineer kombinasyonu olarak belirtilebilir. Böylece alt uzaydaki herhangi bir nokta $\boldsymbol{X\beta}$ formundadır. $\boldsymbol{X\beta}$ vektörü, Şekil 3.3'teki B noktasını belirtsin. B noktasından A noktasına olan karesi alınmış uzaklık,

$$S(\boldsymbol{\beta}) = (\boldsymbol{y} - \boldsymbol{X\beta})'(\boldsymbol{y} - \boldsymbol{X\beta}) \tag{3.27}$$

dir. Bu nedenle $\boldsymbol{y}$ vektörü tarafından tanımlanan A noktasının tahmin uzayına olan karesi alınmış uzaklığını azaltmak, tahmin uzayındaki A noktasına en yakın noktayı bulmayı gerektirir. Karesi alınmış uzaklığın minimum olması için tahmin uzayında bulunacak nokta, A noktasından tahmin uzayına dik olan çizginin ayağı olmalıdır. Bu da Şekil 3.3'teki C noktasıdır. Bu nokta $\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$ vektörü tarafından tanımlanır. $\boldsymbol{y} - \hat{\boldsymbol{y}} = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}$ tahmin uzayına dik olduğundan $\boldsymbol{X}'(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) = 0$ veya $\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}'\boldsymbol{y}$ yazılabilir. Bu ise en küçük kareler normal denklemi olarak bilinen denklemdir.

3.2.4. En Küçük Kareler Tahmin Edicilerinin Özellikleri

$\boldsymbol{\beta}$ parametre vektörünün $\hat{\boldsymbol{\beta}}$ en küçük kareler tahmin edicisinin istatistiksel özelliklerinden birinci olarak yanlılığını inceleyelim.

$\boldsymbol{I}$, birim matrisi göstersin. E($\boldsymbol{\varepsilon}$)=0 ve $(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X} = \boldsymbol{I}$ olduğundan,

$$E(\hat{\boldsymbol{\beta}})=E[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}]$$
$$=E[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon})]$$
$$=E[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta}+(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\varepsilon}]$$
$$=\boldsymbol{\beta} \qquad (3.28)$$

olur. Böylece $\hat{\boldsymbol{\beta}}$ en küçük kareler tahmin edicisi, $\boldsymbol{\beta}$ parametre vektörünün yansız tahmin edicisidir.

$\boldsymbol{\beta}$ parametre vektörünün $\hat{\boldsymbol{\beta}}$ en küçük kareler tahmin edicisinin istatistiksel özelliklerinden ikinci olarak varyansını inceleyelim.

$\hat{\boldsymbol{\beta}}$ en küçük kareler tahmin edicisinin varyans özelliği kovaryans matris ile ifade edilmiştir,

$$Cov(\hat{\boldsymbol{\beta}})=E([\hat{\beta}-\boldsymbol{E}(\hat{\beta})][\hat{\beta}-\boldsymbol{E}(\hat{\beta})]') \qquad (3.29)$$

matrisi $p\times p$ tipinde bir matrisi olup, ana köşegendeki $j-$inci elemanı $\hat{\beta}_j$'nin varyansını ve ana köşegen dışındaki $ij-$inci elemanı $\hat{\beta}_i$ ve $\hat{\beta}_j$ arasındaki kovaryansı belirtir. $\hat{\boldsymbol{\beta}}$'nın kovaryans matrisi,

$$Cov(\hat{\boldsymbol{\beta}})=\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1} \qquad (3.30)$$

şeklindedir. Bu nedenle $\boldsymbol{C}=(\boldsymbol{X}'\boldsymbol{X})^{-1}$ olarak alındığında $\hat{\beta}_j$'nin varyansı, $\sigma^2 C_{jj}$ dir. $\hat{\beta}_i$ ve $\hat{\beta}_j$ arasındaki kovaryans ise $\sigma^2 C_{ij}$ olur.

Gauss-Markov teoreminden $\hat{\boldsymbol{\beta}}$ en küçük kareler tahmin edicisi, $\boldsymbol{\beta}$'nın en iyi lineer yansız tahmin edicisidir (Montgomery ve ark., 2001).

Ayrıca, ε_i hatalarının normal dağılıma sahip olduğu varsayıldığında, $\hat{\boldsymbol{\beta}}$, $\boldsymbol{\beta}$'nın en çok olabilirlik tahmin edicisi olur. Maksimum likehood tahmin edicisi, $\boldsymbol{\beta}$'nın minimum varyanslı yansız tahmin edicisidir (Draper ve Smith, 1998).

3.2.5. σ^2'nin Tahmini

Regresyon analizinde hata kareler toplamı (HKT),

$$SS_E = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{3.31}$$

olarak tanımlanır. Hata kareler toplamından, σ^2'nin bir tahmin edicisi elde edilebilir. (3.31)'deki eşitlikten $SS_E = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$ dir. (3.22)'den $e_i = y_i - \hat{y}_i$ dir. Bu nedenle,

$$SS_E = \sum_{i=1}^{n} e_i^2 = \boldsymbol{e'e} \tag{3.32}$$

olur. (3.32)'deki eşitlikte $\boldsymbol{e = y - X\hat{\beta}}$ alındığında hata kareler toplamı,

$$\begin{aligned} SS_E &= \left(\boldsymbol{y - X\hat{\beta}}\right)' \left(\boldsymbol{y - X\hat{\beta}}\right) \\ &= \boldsymbol{y'y - \hat{\beta}'X'y - y'X\hat{\beta} + \hat{\beta}'X'X\hat{\beta}} \\ &= \boldsymbol{y'y - 2\hat{\beta}'X'y + \hat{\beta}'X'X\hat{\beta}} \end{aligned} \tag{3.33}$$

olur. (3.33)'teki eşitlikte $\boldsymbol{X'X\hat{\beta} = X'y}$ olduğundan,

$$SS_E = \boldsymbol{y'y - \hat{\beta}'X'y} \tag{3.34}$$

olarak elde edilir. Regresyon modelinde tahmin edilen parametre sayısı p olduğunda, hata kareleri toplamının serbestlik derecesi $n-p$ olur. Hata kareler ortalaması ise,

$$MS_E = \frac{SS_E}{n-p} \tag{3.35}$$

olarak tanımlanır. MS_E'nin beklenen değeri σ^2 olduğundan, MS_E, σ^2'nin yansız bir tahmin edicisi olup,

$$\hat{\sigma}^2 = MS_E \tag{3.36}$$

biçiminde ifade edilir.

3.2.6. Regresyon Katsayılarının En Çok Olabilirlik Yöntemiyle Tahmin Edilmesi

Çoklu lineer regresyon modelinde regresyon katsayılarını en çok olabilirlik yöntemiyle tahmin etmek için hata terimlerinin sıfır ortalamalı ve σ^2 varyanslı normal dağılıma sahip olduğu varsayılır (Draper ve Smith, 1998). Regresyon modelindeki parametreler için en çok olabilirlik tahmin edicilerinin en küçük kareler tahmin ediciler olduğu gösterilebilir (Montgomery ve ark., 2001).

Çoklu lineer regresyon modeli (3.13)'deki eşitlikteki gibi matris biçiminde gösterilsin. Burada hatalar, sıfır ortalamalı E(ε)=0 ve σ^2 sabit varyanslı V(ε)=σ^2 normal dağılıma sahiptir. Yani $\varepsilon \sim N\left(0, \sigma^2 I\right)$ şeklindedir. Hatalar için normal yoğunluk fonksiyonu,

$$f\left(\varepsilon_i\right) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}\varepsilon_i^2} \tag{3.37}$$

dir. Olabilirlik fonksiyonu, $\varepsilon_1, \varepsilon_2, ..., \varepsilon_n$ hata terimlerinin ortak olasılık yoğunluk fonksiyonudur. Bu durumda $\prod_{i=1}^{n} f(\varepsilon_i)$ yazılabilir. $L(\boldsymbol{\varepsilon}, \boldsymbol{\beta}, \boldsymbol{\sigma}^2)$ ile gösterilen olabilirlik fonksiyonu,

$$L(\boldsymbol{\varepsilon}, \boldsymbol{\beta}, \boldsymbol{\sigma}^2) = \prod_{i=1}^{n} f(\varepsilon_i) = \frac{1}{\sqrt{(2\pi)^n \sigma^n}} e^{-\frac{1}{2}\frac{1}{\sigma^2}\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}} \tag{3.38}$$

olarak hesaplanır. $\boldsymbol{\varepsilon} = \boldsymbol{y} - \boldsymbol{X\beta}$ olduğundan olabilirlik fonksiyonu,

$$\boldsymbol{L}(\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\beta}, \boldsymbol{\sigma}^2) = \frac{1}{\sqrt{(2\pi)^n \boldsymbol{\sigma}^n}} e^{-\frac{1}{2}\frac{1}{\sigma^2}(y-X\beta)'(y-X\beta)} \tag{3.39}$$

şeklinde yazılabilir. (3.39)'daki eşitlikte her iki tarafın doğal logaritması alındığında,

$$\ln \boldsymbol{L}(\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\beta}, \boldsymbol{\sigma}^2) = -\frac{\boldsymbol{n}}{2}\ln(2\pi) - \boldsymbol{n}\ln(\sigma) - \frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X\beta})'(\boldsymbol{y} - \boldsymbol{X\beta}) \tag{3.40}$$

bulunur. σ'nın sabit bir değeri için $(\boldsymbol{y} - \boldsymbol{X\beta})'(\boldsymbol{y} - \boldsymbol{X\beta})$ minimum olduğunda log-olabilirlik fonksiyon maksimum olur. Bu nedenle, $\boldsymbol{\beta}$ parametre vektörünün $\tilde{\boldsymbol{\beta}}$ en çok olabilirlik tahmin edicisi, hataların normal dağıldığı varsayımı altında $\hat{\boldsymbol{\beta}} = (\boldsymbol{X'X})^{-1}\boldsymbol{X'y}$ olarak elde edilen $\hat{\boldsymbol{\beta}}$ en küçük kareler tahmin edicisi ile eşdeğerdir. Yani $\tilde{\boldsymbol{\beta}} = (\boldsymbol{X'X})^{-1}\boldsymbol{X'y}$ dir. σ^2'nin en çok olabilirlik tahmin edicisi ise

$$\tilde{\sigma}^2 = \frac{(y - X\hat{\beta})'(y - X\hat{\beta})}{n} \tag{3.41}$$

dir.

3.3. Çoklu Lineer Regresyon Modelinde Hipotez Testi

Çoklu lineer regresyon modeline parametrelerle ilgili hipotezlerin test edilmesi, regresyon modelinin yeterliliğini ölçmede yararlıdır. Bu kısımda çoklu lineer regresyon modelleriyle ilgili önemli hipotez testleri incelenecektir. Hipotezlerin test edilmesinde, regresyon modelindeki hataların normal dağılıma sahip olduğu varsayımı yapılır (Montgomery ve ark., 2001).

3.3.1. Regresyonun Önemliliğinin Test Edilmesi

Regresyonun önemliliği için test, y yanıt değişkeni ile $x_1, x_2, ..., x_k$ açıklayıcı değişkenler arasında lineer bir ilişkinin bulunup bulunmadığını belirlemede kullanılan bir testtir. Diğer bir ifadeyle $x_1, x_2, ..., x_k$ açıklayıcı değişkenlerine, y yanıt değişkendeki toplam değişimi açıklamada gereksinim olup olmadığının test edilmesidir. Bu durumda uygun hipotezler,

$$H_0: \beta_1 = \beta_2 = ... = \beta_k = 0$$
$$H_1: j = 1, 2, ..., k \text{ olmak üzere en az bir } j \text{ için } \beta_j \neq 0 \qquad (3.42)$$

biçiminde oluşturulabilir. $H_0 : \beta_j = 0$ hipotezinin red edilmesi $x_1, x_2, ..., x_k$ açıklayıcı değişkenlerinden en az birinin modele katkısının olduğu anlamına gelir.

S_{yy} ile gösterilen genel kareler toplamı, SS_R ile gösterilen regresyon kareler toplamı ve SS_E ile gösterilen hata kareler toplamı olarak parçalanabilir (Montgomery ve ark., 2001). Yani

$$S_{yy} = SS_R + SS_E \qquad (3.43)$$

dir. Eğer $H_0 : \beta_j = 0$ hipotezi doğru ise $\frac{SS_R}{\sigma^2}$ ifadesi k serbestlik dereceli Ki-kare dağılımına sahiptir. Yani $\frac{SS_R}{\sigma^2} \sim \chi_k^2$ dir (Montgomery ve ark., 2001). Burada k, çoklu lineer regresyon modelindeki açıklayıcı değişkenlerin sayısıdır. $\frac{SS_E}{\sigma^2}$ ifadesi $n-k-1$ serbestlik dereceli Ki-kare dağılımına sahiptir. Yani $\frac{SS_E}{\sigma^2} \sim \chi_{n-k-1}^2$ dir (Montgomery ve ark., 2001). SS_R ile SS_E birbirinden bağımsızdır (Montgomery ve ark., 2001). $H_0 : \beta_j = 0$ hipotezini test etmek için,

$$F_0 = \frac{SS_R / k}{SS_E / (n-k-1)} = \frac{MS_R}{MS_E} \tag{3.44}$$

test istatistiği kullanılabilir. $F_0 > F_{\alpha,k,n-k-1}$ ise H_0 hipotezi red edilir. H_0 hipotezini test etmek için oluşturulan varyans analizi tablosu Tablo 3.8'de verilmiştir.

Tablo 3.8. Çoklu lineer regresyon modelinde regresyonun önemliliğini test etmek için kullanılan varyans analizi tablosu.

Değişimin Kaynağı	Kareler Toplamı	Serbestlik Derecesi	Kareler Ortalaması	Test İstatistiği
Regresyon	SS_R	k	$MS_R = \frac{SS_R}{k}$	$F_0 = \frac{MS_R}{MS_E}$
Hata	SS_E	$n-k-1$	$MS_E = \frac{SS_E}{n-k-1}$	
Toplam	S_{yy}	$n-1$		

Tablo 3.8'deki SS_E değeri,

$$SS_E = \boldsymbol{y'y - \hat{\beta}'X'y} \tag{3.45}$$

kullanılarak hesaplansın. S_{yy},

$$S_{yy} = \sum_{i=1}^{n} y_i - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n} = \boldsymbol{y'y} - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n} \tag{3.46}$$

olduğundan SS_E,

$$SS_E = \boldsymbol{y'y} - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n} - [\,\hat{\boldsymbol{\beta}}'\boldsymbol{X'y} - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}\,] \tag{3.47}$$

veya

$$SS_E = S_{yy} - SS_R \tag{3.48}$$

şeklinde yazılabilir. Bu nedenle regresyon kareler toplamı,

$$SS_R = \hat{\boldsymbol{\beta}}'\boldsymbol{X'y} - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n} \tag{3.49}$$

olarak elde edilir. SS_E hata kareler toplamı,

$$SS_E = \boldsymbol{y'y} - \hat{\boldsymbol{\beta}}'\boldsymbol{X'y} \tag{3.50}$$

ve S_{yy} genel kareler toplamı ise

$$S_{yy} = \boldsymbol{y'y} - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n} \tag{3.51}$$

olarak bulunur.

Çoklu lineer regresyon modelinde regresyonun önemliliği testi örneği Örnek 3.2'de verilmiştir.

Örnek 3.2. İçecek teslim/dağıtım verisi için regresyonun önemliliğini test edelim.

(3.51)'deki eşitlikten S_{yy}=5784.5426 olur. (3.49)'daki eşitlikten SS_R=5550.8166 bulunur. (3.48)'deki eşitlikten de SS_E=233.7260 olur. Bu durumda varyans analizi tablosu Tablo 3.9'daki gibi olur.

Tablo 3.9. İçecek teslim/dağıtım verisi için oluşturulan çoklu lineer regresyon modelinde regresyonun önemliliğini test etmek için kullanılan varyans analizi tablosu.

Değişimin Kaynağı	Kareler Toplamı	Serbsetlik Derecesi	Kareler Ortalaması	Test İstatistiği
Regresyon	5550.8166	2	2775.4083	261.24
Hata	233.7260	22	10.6239	
Toplam	5784.5426	24		

$H_0 : \beta_1 = \beta_2 = 0$ hipotezini test etmek için hesaplanan test istatistiği,

$$F_0 = \frac{MS_R}{MS_E} = \frac{2775.4083}{10.6239} = 261.24 \qquad (3.52)$$

olur. $F_0 = 261.24 > F_{Tablo} = F_{.05,2,22} = 3.44$ olduğundan, yanıt değişkendeki (teslim süresi) toplam değişimi açıklamada açıklayıcı değişkenlerin (teslim miktarı ve uzaklık) gerekli olduğu sonucuna varılır. Ancak bu, teslim süresinin ürün miktarı ve uzaklığın bir fonksiyonu olduğu anlamına gelmez. Modelin yeterliliği için başka testlere de başvurulmalıdır (Montgomery ve ark., 2001).

3.3.2. Her Bir Regresyon Katsayısı İçin Hipotezin Test Edilmesi

Çoklu lineer regresyon modelinde her bir açıklayıcı değişkenin modele katkısının olup olmadığını test etmek amacıyla her bir regresyon katsayısı için hipotez test edilir. Regresyon katsayısı hakkındaki hipotez testinin sonucuna göre ilgili açıklayıcı değişkenin modele eklenip eklenmeyeceğine karar verilir. Bazı durumlarda açıklayıcı değişkenin modelden çıkarılması modeli daha etkin hale getirebilir.

Bir regresyon modeline yeni bir değişkenin eklenmesi daima regresyon kareler toplamının artmasına ve hata kareler toplamının azalmasına neden olur. Regresyon kareler toplamındaki artışın modele eklenen yeni açıklayıcı değişkenden kaynaklandığını garantilemek gerekir. Bir açıklayıcı değişkenin modele eklenmesi, oluşturulan $\hat{y}$ değerinin varyansını da arttırır. Bu nedenle yanıt değişkendeki toplam değişimi açıklamada doğru açıklayıcı değişkenlerin modele dahil edilmesinde dikkatli olunmalıdır. Ayrıca modelin kullanışlılığını azaltan önemsiz bir açıklayıcı değişkenin modele eklenmesi hata kareler ortalamasını arttırabilir. Herhangi bir β_j regresyon katsayısının önemliliğini test etmek için oluşturulan hipotezler,

$$H_0: \beta_j = 0$$
$$H_1: \beta_j \neq 0 \qquad (3.53)$$

biçiminde oluşturulabilir. $H_0: \beta_j = 0$ hipotezinin red edilmemesi x_j açıklayıcı değişkeninin modelden silinebileceğini gösterir. Bu hipotez için test istatistiği,

$$t_0 = \frac{\hat{\beta}_j}{\sqrt{\hat{\sigma}^2 C_{jj}}} = \frac{\hat{\beta}_j}{\text{se}\left(\hat{\beta}_j\right)} \qquad (3.54)$$

Dir (Montgomery ve ark., 2001). Burada C_{jj}, $\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}$ matrisinde $\hat{\beta}_j$'ye karşılık gelen ana köşegen elemanıdır. $|t_0| > t_{\alpha/2,n-k-1}$ ise H_0 hipotezi ya da sıfır hipotezi red

edilir. $i \neq j$ olmak üzere $\hat{\beta}_j$ regresyon katsayısı modeldeki diğer bütün x_i $(i \neq j)$ açıklayıcı değişkenlerine bağlı olduğundan bu bir marjinal testtir. Dolayısıyla bu test, modeldeki diğer açıklayıcı değişkenler verildiğinde x_j'nin modele katkısını ölçer.

Çoklu lineer regresyon modelinde her bir regresyon katsayısı hakkındaki hipotez testi örneği Örnek 3.3'te verilmiştir.

Örnek 3.3. İçecek teslim/dağıtım verisindeki x_2 uzaklık açıklayıcı değişkeninin modele eklenip eklenmeyeceğini test edelim. Diğer bir ifadeyle y yanıt değişkendeki toplam değişimi açıklamada x_2 uzaklık açıklayıcı değişkenine gereksinim olup olmadığını test edelim. Bunun için oluşturulan hipotezler,

$$H_0: \beta_2 = 0$$
$$H_1: \beta_2 \neq 0 \qquad (3.55)$$

şeklindedir. $(\boldsymbol{X}'\boldsymbol{X})^{-1}$'nin β_2'ye karşılık gelen ana köşegen elemanı $C_{22} = 0.00000123$ tür. (3.54)'deki eşitlikteki t-istatistiği, $t_0 = \dfrac{\hat{\beta}_2}{\sqrt{\hat{\sigma}^2 C_{22}}} = \dfrac{0.01438}{\sqrt{(10.6239)(0.00000123)}} = 3.98$ olarak hesaplanır. $t_{.025,22} = 2.074$ tür. $t_0 = 3.98 > t_{\frac{\alpha}{2}, n-k-1} = 2.074$ olduğundan $H_0: \beta_2 = 0$ red edilir. x_2 uzaklık değişkeninin modele katkısının istatistiksel olarak anlamlı ya da önemli olduğu sonucu çıkarılır.

3.3.3. Regresyon Katsayılarının Bir Alt Kümesi İçin Hipotezin Test Edilmesi

Açıklayıcı değişkenlerin bir altkümesinin modele katkısı incelensin. k tane açıklayıcı değişken bulunan çoklu lineer regresyon modeli göz önüne alınsın. Bu model (3.13)'deki eşitlikteki gibi $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ şeklinde gösterilsin. Burada $\boldsymbol{y}$, $n \times 1$

tipinde gözlenen yanıtların vektörünü; $\boldsymbol{X}$, $n \times p$ tipinde açıklayıcı değişkenlerin düzeylerinin matrisini; $\boldsymbol{\beta}$, $p \times 1$ tipinde parametre vektörünü ve $\boldsymbol{\varepsilon}$, $n \times 1$ tipinde hata vektörünü göstersin. $p = k + 1$ dir. $r < k$ olmak üzere r tane elemandan oluşan açıklayıcı değişkenlerin bir alt kümesinin regresyon modeline katkısı belirlenebilir. Bunun için regresyon katsayılar vektörü $\boldsymbol{\beta}$,

$$\boldsymbol{\beta} = \begin{bmatrix} \boldsymbol{\beta}_1 \\ \cdots \\ \boldsymbol{\beta}_2 \end{bmatrix} \tag{3.56}$$

şeklinde $(p-r) \times 1$ tipinde $\boldsymbol{\beta}_1$ vektörüne ve $r \times 1$ tipinde $\boldsymbol{\beta}_2$ vektörüne parçalanır. Uygun hipotezler,

$$\begin{aligned} &H_0 : \boldsymbol{\beta}_2 = 0 \\ &H_1 : \boldsymbol{\beta}_2 \neq 0 \end{aligned} \tag{3.57}$$

dir. Bu durumda $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ modeli,

$$\boldsymbol{y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon} \tag{3.58}$$

şeklinde yazılabilir. Burada $n \times (p-r)$ tipindeki $\boldsymbol{X}_1$ matrisi, $\boldsymbol{X}$'in $\boldsymbol{\beta}_1$ ile ilgili sütunlarını ve $n \times r$ tipindeki $\boldsymbol{X}_2$ matrisi ise $\boldsymbol{X}$'in $\boldsymbol{\beta}_2$ ile ilgili sütunlarını belirtmektedir. Bu modele, tam model denilmektedir. Tam model için $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$ dir. (3.58)'deki eşitlikteki model için regresyon kareler toplamı,

$$SS_R(\boldsymbol{\beta}) = \hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{y} \tag{3.59}$$

dir. $SS_R(\boldsymbol{\beta})$'nın serbestlik derecesi p dir. (3.58)'deki eşitlikteki model için hata kareler ortalaması,

$$MS_E = \frac{\boldsymbol{y}'\boldsymbol{y} - \hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{y}}{\boldsymbol{n}-\boldsymbol{p}} \tag{3.60}$$

olur. $\boldsymbol{\beta}_2$ vektöründeki terimlerin regresyona katkısını bulmak için H_0: $\boldsymbol{\beta}_2$=0 sıfır hipotezinin doğru olduğunu varsayılır. Bu durumda

$$\boldsymbol{y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\varepsilon} \tag{3.61}$$

indirgenmiş modeli elde edilir. İndirgenmiş modeldeki $\boldsymbol{\beta}_1$'in en küçük kareler tahmin edicisi,

$$\hat{\boldsymbol{\beta}}_1 = (\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'\boldsymbol{y} \tag{3.62}$$

olup bu modele karşılık gelen regresyon kareler toplamı,

$$SS_R(\boldsymbol{\beta}_1) = \hat{\boldsymbol{\beta}}_1' X_1' y \tag{3.63}$$

dir. $SS_R(\boldsymbol{\beta}_1)$'in serbestlik derecesi $p-r$ dir. (3.58)'deki eşitlikteki modelde $\boldsymbol{\beta}_1$ verilmişken $\boldsymbol{\beta}_2$'den kaynaklanan regresyon kareleri toplamı,

$$SS_R(\boldsymbol{\beta}_2/\boldsymbol{\beta}_1) = SS_R(\boldsymbol{\beta}) - SS_R(\boldsymbol{\beta}_1) \tag{3.64}$$

dır. $SS_R(\boldsymbol{\beta}_2/\boldsymbol{\beta}_1)$'in serbestlik derecesi $p-(p-r)=r$ dir. (3.64)'deki eşitlikteki $SS_R(\boldsymbol{\beta}_2/\boldsymbol{\beta}_1)$ kareler toplamına, $\boldsymbol{\beta}_2$'den kaynaklanan ekstra kareler toplamı denir. Çünkü ekstra kareler toplamı, $x_1, x_2, ..., x_{k-r}$ açıklayıcı değişkenlerini içeren modele $x_{k-r+1}, x_{k-r+2}, ..., x_k$ açıklayıcı değişkenlerinin eklenmesiyle elde edilen regresyon kareler toplamındaki artışı ölçer. $SS_R(\boldsymbol{\beta}_2/\boldsymbol{\beta}_1)$, MS_E'den bağımsızdır (Montgomery ve ark., 2001). H_0: $\boldsymbol{\beta}_2$=0 sıfır hipotezini test etmek için

$$F_0 = \frac{SS_R(\boldsymbol{\beta}_2|\boldsymbol{\beta}_1)/r}{MS_E} \tag{3.65}$$

istatistiği kullanılır. $F_0 > F_{\alpha,r,n-p}$ ise H_0 red edilir. Bu durumda $\boldsymbol{\beta}_2$'deki parametrelerden en az birinin sıfır olmadığı ve bu nedenle $\boldsymbol{X}_2$'deki $x_{k-r+1}, x_{k-r+2}, ..., x_k$ açıklayıcı değişkenlerinden en az birinin regresyon modeline katkısının anlamlı olduğu sonucu çıkarılır. (3.65)'deki eşitlikteki F_0 test istatistiği kısmi F-testi olarak ta adlandırılır. Sadece bir x_j değişkeni üzerindeki kısmi F-testi, t-testine eşdeğerdir (Montgomery ve ark., 2001).

Çoklu lineer regresyon modelinde regresyon katsayılarının bir alt kümesi hakkındaki hipotez testi örneği, Örnek 3.4'te verilmiştir.

Örnek 3.4. İçecek teslim/dağıtım verisindeki x_2 uzaklık değişkeninin modele katkısını inceleyelim. Uygun hipotezler, $H_0: \beta_2 = 0$ ve $H_1: \beta_2 \neq 0$ dır. Bu hipotezleri test etmek için β_2'den kaynaklanan ekstra kareler toplamı hesaplanmalıdır. β_2'den kaynaklanan ekstra kareler toplamı,

$$\begin{aligned} SS_R(\beta_2|\beta_1,\beta_0) &= SS_R(\beta_1,\beta_2,\beta_0) - SS_R(\beta_1,\beta_0) \\ &= SS_R(\beta_1,\beta_2|\beta_0) - SS_R(\beta_1|\beta_0) \end{aligned} \tag{3.66}$$

eşitliğiyle hesaplanabilir. (3.66)'daki eşitlikteki $SS_R(\beta_1,\beta_2/\beta_0)$ değeri $SS_R(\beta_1,\beta_2|\beta_0) = \hat{\beta}'X'y - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n} = 5550.8166$ olarak hesaplanır. $SS_R(\beta_1,\beta_2/\beta_0)$'ın serbestlik derecesi 2 dir. İndirgenmiş model $y = \beta_0 + \beta_1 x_1 + \varepsilon$, $\hat{y} = 3.328 + 2.1762x_1$ şeklinde oluşturulur. Bu model için $SS_R(\beta_1/\beta_0)$ regresyon kareler toplamı, $SS_R(\beta_1|\beta_0) = \hat{\beta}_1 S_{xy} = (2.1762)(2473.3440) = 5382.4088$ olarak

hesaplanır. $SS_R(\beta_1/\beta_0)$'ın serbestlik derecesi 1 dir. Bu nedenle $SS_R(\beta_1,\beta_2|\beta_0) = 5550.8166 - 5382.4088 = 168.4078$ olarak bulunur. Bu değer x_1'i içeren modele, x_2'nin eklenmesinden kaynaklanan regresyon kareler toplamındaki artıştır. $H_0 : \beta_2 = 0$ sıfır hipotezini test etmek için F_0 test istatistiği, $F_0 = \dfrac{SS_R(\beta_2|\beta_1,\beta_0)/1}{MS_E} = \dfrac{168.4078/1}{10.6239} = 15.85$ olarak hesaplanır. Dikkat edilirse, hem x_1'i hem de x_2'yi içeren tam modele ait MS_E hata kareler ortalaması, test istatistiğinin paydasında kullanılmaktadır. $F_{.05,1,22} = 4.30$ olduğundan $H_0 : \beta_2 = 0$ sıfır hipotezi red edilir. x_2 uzaklık değişkeninin modele katkısının istatistiksel olarak anlamlı ya da önemli olduğu sonucu çıkarılır.

Bu örnekte bir tek değişken olduğu için burada F-testi, t-testine eşdeğerdir. Bunu görmek için $H_0 : \beta_2 = 0$ sıfır hipotezi t-testiyle hesaplansın. Bu durumda $t_0 = 3.98$ dır. v serbestlik dereceli t rasgele değişkenin karesi, 1 serbestlik dereceli F rasgele değişkeni olduğundan $t_0^2 = (3.98)^2 = 15.84 \cong F_0$ olur.

3.3.4. X Matrisinde Sütunların Ortogonal Olması Özel Durumu

(3.13)'deki $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$ regresyon modelinin (3.58)'deki $\boldsymbol{y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}$ biçimini göz önüne alınsın. Ekstra kareler toplamı yöntemi, $SS_R(\boldsymbol{\beta}_1/\boldsymbol{\beta}_0)$'ı hesaplayarak, $\boldsymbol{X}_1$ koşuluna bağlı olarak $\boldsymbol{X}_2$'deki açıklayıcı değişkenlerin etkisini ölçer. Genel olarak, $\boldsymbol{X}_1$'deki açıklayıcı değişkenler üzerindeki bağımlılık açıklanmadan, $\boldsymbol{\beta}_2$'den kaynaklanan kareler toplamı, $SS_R(\boldsymbol{\beta}_1/\boldsymbol{\beta}_0)$'nin hesaplanmasından bahsedilemez. Bununla birlikte, $\boldsymbol{X}_1$'deki sütunlar, $\boldsymbol{X}_2$'deki sütunlara ortogonal ise $\boldsymbol{\beta}_2$'den kaynaklanan kareler toplamı belirlenebilir. Bunu göstermek için (3.58)'deki model ile ilgili $\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}'\boldsymbol{y}$ normal denklemleri oluşturulsun. Normal denklemler,

$$\begin{bmatrix} X_1'X_1 & \vdots & X_1'X_2 \\ \cdots & & \cdots \\ X_2'X_1 & \vdots & X_2'X_2 \end{bmatrix}\begin{bmatrix} \hat{\beta}_1 \\ \cdots \\ \hat{\beta}_2 \end{bmatrix}=\begin{bmatrix} X_1'y \\ \cdots \\ X_2'y \end{bmatrix} \tag{3.67}$$

şeklinde olur. X_1'deki sütunlar, X_2'deki sütunlara ortogonal ise $X_1'X_2 = 0$ veya $X_2'X_1 = 0$ olur. O halde normal denklemler,

$$\begin{aligned} X_1'X_1\beta_1 &= X_1'y \\ X_2'X_2\beta_2 &= X_2'y \end{aligned} \tag{3.68}$$

şeklinde olur. (3.68)'deki normal denklemlerin çözümü,

$$\begin{aligned} \hat{\beta}_1 &= (X_1'X_1)^{-1}X_1'y \\ \hat{\beta}_2 &= (X_2'X_2)^{-1}X_2'y \end{aligned} \tag{3.69}$$

bulunur. (3.69)'daki eşitlikteki β_1'in en küçük kareler tahmin edicisi, X_2'nin modelde bulunup bulunmadığına bakılmaksızın $\hat{\beta}_1$'dır. Benzer şekilde, β_2'nin en küçük kareler tahmin edicisi, X_1'in modelde olup olmadığına bakılmaksızın $\hat{\beta}_2$'dır.

Tam model için regresyon kareleri toplamı,

$$\begin{aligned} SS_R(\beta) &= \hat{\beta}'X'y \\ &= \begin{bmatrix} \hat{\beta}_1 & \hat{\beta}_2 \end{bmatrix}\begin{bmatrix} X_1'y \\ X_2'y \end{bmatrix} \\ &= \hat{\beta}_1X_1'y + \hat{\beta}_2X_2'y \\ &= y'X_1(X_1'X_1)^{-1}X_1'y + y'X_2(X_2'X_2)^{-1}X_2'y \end{aligned} \tag{3.70}$$

bulunur. Bununla birlikte, normal denklemleri iki küme oluşturmaktadır: Her bir küme için,

$$SS_R(\beta_1) = \hat{\beta}_1' X_1' y = y' X_1 (X_1' X_1)^{-1} X_1' y \quad (3.71)$$

$$SS_R(\beta_2) = \hat{\beta}_2' X_2' y = y' X_2 (X_2' X_2)^{-1} X_2' y \quad (3.72)$$

dir. (3.71) ve (3.72)'deki eşitliklerdeki ifadeler kullanılarak,

$$SS_R(\beta) = SS_R(\beta_1) + SS_R(\beta_2) \quad (3.73)$$

olduğu görülür. Bu nedenle,

$$SS_R(\beta_1/\beta_2) = SS_R(\beta) - SS_R(\beta_2) \cong SS_R(\beta_1) \quad (3.74)$$

ve

$$SS_R(\beta_2/\beta_1) = SS_R(\beta) - SS_R(\beta_1) \cong SS_R(\beta_2) \quad (3.75)$$

olur. Sonuç olarak, $SS_R(\beta_1)$, X_1'deki açıklayıcı değişkenlerin modele olan katkısını koşulsuz olarak ölçmektedir. $SS_R(\beta_2)$, X_2'deki açıklayıcı değişkenlerin modele olan katkısını koşulsuz olarak ölçmektedir. Açıklayıcı değişkenler ortogonal olduklarından her bir açıklayıcı değişkenin etkisi tam olarak belirlenebilir. Deneysel tasarımlar, çoğunlukla ortogonal değişkenler içerecek şekilde tasarlanır.

Ortogonal açıklayıcı değişkenler içeren bir regresyon modeline örnek olarak, $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$ modelini ele alalım. X matrisi,

$$X = \begin{matrix} \beta_0 & \beta_1 & \beta_2 & \beta_3 \end{matrix} \begin{bmatrix} 1 & -1 & -1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 \\ 1 & 1 & 1 & -1 \\ 1 & 1 & -1 & 1 \\ 1 & -1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix}$$

şeklinde olsun. Açıklayıcı değişkenlerin seviyesi 2^3 faktöriyel modeline uymaktadır. X'in sütunları ortogonaldir. Böylece $j = 1,2,3$

olmak üzere $\boldsymbol{SS_R(\beta_j)}$, diğer açıklayıcı değişkenlerden herhangi birinin modele dahil edilip edilmediğine bakılmaksızın, x_j açıklayıcı değişkenlerinin modele katkısını ölçer.

3.3.5. $T\beta = 0$ Genel Lineer Hipotezlerinin Test Edilmesi

Regresyon katsayıları hakkında birçok hipotez birleştirilmiş yaklaşım kullanılarak test edilebilir. Ekstra kareler toplamı yöntemi bu yöntemin özel bir durumudur. Daha genel yöntemlerde hipotezleri test etmek için kullanılan kareler toplamı çoğunlukla iki hata kareleri toplamı arasındaki fark olarak hesaplanır. Şimdi de bu yöntem incelensin (Searle 1971, Graybill 1976, Seber 1977).

$\mathbf{T}$, $m \times p$ tipinde sabitlerin matrisi olmak üzere ilgili hipotezler $H_0 : \boldsymbol{T\beta} = \mathbf{0}$ şeklinde ifade edilsin. $\boldsymbol{T\beta} = \mathbf{0}$'daki m tane denklemden sadece r tanesi bağımsızdır. $\hat{\boldsymbol{\beta}} = (\boldsymbol{X'X})^{-1}\boldsymbol{X'y}$ olmak üzere tam model $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$ olup tam model için hata kareler toplamı, $SS_E(FM) = y'y - \hat{\beta}'X'y$ dır. Tam model için hata kareler toplamının serbestlik derecesi $n - p$ dir. İndirgenmiş modeli elde etmek için kalan $p - r$ tane regresyon katsayısı kullanılır. Tam modeldeki regresyon katsayılarının r tanesi için $\boldsymbol{T\beta} = \mathbf{0}$'daki r tane bağımsız eşitlik çözülür. Bu durumda $\mathbf{Z}$, $n \times (p\text{-}r)$ tipinde bir matrisi ve γ, bilinmeyen regresyon katsayılarının $(p\text{-}r) \times 1$ tipinde bir vektörü olmak üzere indirgenmiş model,

$$\boldsymbol{y} = \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon} \tag{3.76}$$

dir. γ'nin en küçük kareler tahmin edicisi,

$$\hat{\boldsymbol{\gamma}} = (\boldsymbol{Z'Z})^{-1}\boldsymbol{Z'y} \tag{3.77}$$

olur. İndirgenmiş model için hata kareler toplamı,

$$SS_E(RM) = \boldsymbol{y'y} - \hat{\boldsymbol{\gamma}}'\boldsymbol{Z'y} \tag{3.78}$$

bulunur. İndirgenmiş model için hata kareler toplamının serbestlik derecesi de $n-p+r$ dir.

İndirgenmiş model, tam modelden daha az sayıda parametre içerdiğinden $SS_E(RM) \geq SS_E(FM)$ olur. $H_0 : \mathbf{T}\boldsymbol{\beta} = \mathbf{0}$ hipotezini test etmek için hata kareler toplamındaki fark kullanılır. Hata kareleri toplamındaki fark,

$$SS_H = SS_E(RM) - SS_E(FM) \tag{3.79}$$

olur. Hata kareleri toplamının serbestlik derecesi, $n-p+r-(n-p)=r$ dir. Burada SS_H, $H_0 : \mathbf{T}\boldsymbol{\beta} = \mathbf{0}$ hipotezinden kaynaklanan kareler toplamı olarak adlandırılır. Hipotezleri test etmek için

$$F_0 = \frac{SS_H / r}{SS_E(FM)/(n-p)} \tag{3.80}$$

test istatistiği kullanılır (Montgomery ve ark., 2001). $F_0 > F_{\alpha,r,n-p}$ ise $H_0 : \mathbf{T}\boldsymbol{\beta} = \mathbf{0}$ hipotezi red edilir.

3.4. Çoklu Regresyonda Güven Aralıkları

Çoklu lineer regresyon modelinde regresyon katsayıları ve ortalama yanıt için güven aralıkları incelenecektir.

3.4.1. Regresyon Katsayıları İçin Güven Aralıkları

Çoklu lineer regresyon modelinin β_j regresyon katsayıları için güven aralık tahminlerini oluşturmak için ε_i hatalarının, sıfır ortalamalı ve σ^2 varyanslı normal

dağılıma sahip olduğu varsayımı yapılır (Montgomery ve ark., 2001). Bu nedenle y_i gözlemleri, $\beta_o + \sum_j^k \beta_j x_{ij}$ ortalamalı ve σ^2 varyanslı normal dağılıma sahip olur. $\hat{\boldsymbol{\beta}}$ en küçük kareler tahmin edicisi gözlemlerin bir lineer kombinasyonu olduğundan, $\boldsymbol{\beta}$ ortalama vektörlü ve $\sigma^2 (\boldsymbol{X'X})^{-1}$ kovaryans matrisli normal dağılıma sahiptir. Bu nedenle C_{jj}, $(\boldsymbol{X'X})^{-1}$ matrisinin j-inci köşegen elemanı olmak üzere, herhangi bir $\hat{\beta}_j$ regresyon katsayısının marjinal dağılımının, β_j ortalamalı ve $\sigma^2 C_{jj}$ varyanslı normal dağılım olduğunu gösterir (Draper ve Smith, 1998). E($\hat{\beta}_j$)= β_j ve V($\hat{\beta}_j$)= $\sigma^2 C_{jj}$ olduğundan,

$$\frac{\hat{\beta}_j - \beta_j}{\sqrt{\hat{\sigma}^2 C_{jj}}} \qquad j = 0,1,2,...,k \tag{3.81}$$

istatistiklerin her biri, $n-p$ serbestlik dereceli t dağılımına sahiptir (Montgomery ve ark., 2001). Burada $\hat{\sigma}^2$, (3.36)'da elde edilen hata varyansının tahminidir. Bu nedenle $\hat{\beta}_j$, $j = 0,1,2,...,k$ regresyon katsayısı için yüzde $100(1-\alpha)$'lık önem düzeyinde bir güven aralığı,

$$\hat{\beta}_j - t_{\alpha/2,n-p}\sqrt{\hat{\sigma}^2 C_{jj}} \le \beta_j \le \hat{\beta}_j + t_{\alpha/2,n-p}\sqrt{\hat{\sigma}^2 C_{jj}} \tag{3.82}$$

dir. Burada

$$\text{se}\left(\hat{\beta}_j\right) = \sqrt{\hat{\sigma}^2 C_{jj}} \tag{3.83}$$

değerine $\hat{\beta}_j$ regresyon katsayısının standart hatası denir. Standart hata tahminin ne kadar hassas olduğunu ölçer.

3.4.2. Ortalama Yanıt İçin Güven Aralığı

Açıklayıcı değişkenlerin $x_{01}, x_{02}, ..., x_{0k}$ gibi belirli bir düzeyinde ortalama yanıt için bir güven aralığı oluşturulabilir. $\boldsymbol{x}_0$ vektörü,

$$\boldsymbol{x}_0 = \begin{bmatrix} 1 \\ x_{01} \\ x_{02} \\ \vdots \\ x_{0k} \end{bmatrix} \tag{3.84}$$

şeklinde tanımlansın. Bu noktada oluşturulan fonksiyonun değeri,

$$\hat{y}_0 = \boldsymbol{x}_0' \hat{\boldsymbol{\beta}} \tag{3.85}$$

olur. $\hat{y}_0$'ın beklenen değeri,

$$\mathrm{E}(\hat{y}_0) = \boldsymbol{x}_0' \hat{\boldsymbol{\beta}} = y_0 \tag{3.86}$$

olduğundan $\hat{y}_0$, y_0'ın yansız bir tahmin edicisidir. $\hat{y}_0$'ın varyansı ise

$$\mathrm{V}(\hat{y}_0) = \sigma^2 \, \boldsymbol{x}_0' \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{x}_0 \tag{3.87}$$

olur. Bu nedenle açıklayıcı değişkenlerin $x_{01}, x_{02}, ..., x_{0k}$ düzeyinde ortalama yanıt için yüzde$100(1-\alpha)$'lık önem düzeyinde bir güven aralığı,

$$\hat{y}_0 - t_{\frac{\alpha}{2},n-p} \sqrt{\sigma^2 \boldsymbol{x}_0^{'} (\boldsymbol{X}'\boldsymbol{X})^{-1} \boldsymbol{x}_0} \le y_0 \le \hat{y}_0 + t_{\frac{\alpha}{2},n-p} \sqrt{\sigma^2 \boldsymbol{x}_0^{'} (\boldsymbol{X}'\boldsymbol{X})^{-1} \boldsymbol{x}_0} \tag{3.88}$$

Olur (Montgomery ve ark., 2001).

Çoklu lineer regresyon modelinde ortalama yanıt hakkında güven aralığı oluşturulması örneği, Örnek 3.5'te verilmiştir.

Örnek 3.5. İçecek teslim/dağıtım verisindeki açıklayıcı değişkenlerin $x_1 = 8$ kutu ve $x_2 = 275$ feet uzaklık için ortalama yanıt (teslim süresi) için %95'lik güven aralığını oluşturalım.

$x_0 = \begin{bmatrix} 1 \\ 8 \\ 275 \end{bmatrix}$ olmak üzere bu noktada oluşturulan fonksiyonun değeri (3.85)'deki eşitlikten $\hat{y}_0$=19.22 olur. $\hat{y}_0$'ın varyansı (3.87)'deki eşitlikten V($\hat{y}_0$)=0.56794 olur. Bu noktada ortalama yanıt için %95'lik güven aralığı (3.88)'deki eşitlikten $19.22 - 2.074\sqrt{0.56794} \le y_0 \le 2.074\sqrt{0.56794}$ ya da $17.66 \le y_0 \le 20.78$ olarak bulunur.

4. ÇOKLU LİNEER REGRESYON MODELLERİNİN OLUŞTURULMASI VE EN İYİ MODELİN SEÇİLMESİ

Çoklu lineer regresyon modelinde, modele dahil edilen açıklayıcı değişkenlerin ya da regresörlerin etkili ya da yanıt değişkendeki toplam değişimi açıklamada modele katkı sağlayacağı varsayılmıştı. Bununla birlikte regresyon analizinde amaç regresyon modelinin fonksiyonel şeklinin doğruluğunu da garanti etmektir.

Bazı uygulamalarda teorik bilgilerden ve önceki deneyimlerden hareketle, modelde kullanılacak regresörlerin seçimi yapılır. Bununla beraber analizci çoğu problemde yanıt değişkendeki toplam değişimi açıklamada kullanılabileceği düşünülen regresörlerin bir kümesiyle karşılaşır. Bu durumda problem, modelde kullanılması gereken regresörlerin gerçek altkümelerin belirlenmesine indirgenir (Hocking 1972, 1976). Regresyon modeli için anlamlı ya da önemli olan regresörlerin uygun alt kümesini belirleme işlemine "değişken seçimi problemi" adı verilir (Draper ve Smith, 1998).

4.1. Çoklu Lineer Regresyonda En İyi Modelin Seçilmesi

Mevcut regresörlerin sadece bir alt kümesini içeren regresyon modelinin oluşturulmasının iki amacı vardır (Montgomery ve ark., 2001).

1. Oluşturulan modelin olası birçok regresörü içermesi istenir. Böylece bu etkenlerdeki (faktörlerdeki) "bilgi içeriği" $\hat{y}$ oluşturulan ya da tahmin edilen yanıt değerlerini etkiler.
2. Oluşturulan modelin olası birkaç regresör içermesi istenir. Çünkü $\hat{y}$ oluşturulan ya da tahmin edilen yanıt değerlerinin varyansı, regresör sayılarının artışıyla artar. Ayrıca modeldeki regresör sayısının artması daha fazla veri toplama demektir. Bu da işlem için gereken süreyi ve maliyeti arttırır.

Bu iki amaç arasındaki uzlaşma olan bir model oluşturma işlemine "en iyi" regresyon modelini ya da denklemini seçme denir. En iyinin tek bir açıklaması yoktur. Ayrıca değişken seçimi için uygulanan birçok yöntem vardır. Bu yöntemler, çoğu kez aday regresörlerin farklı alt kümelerini en iyi olarak belirtirler.

Örneklem için çoklu lineer regresyon modeli,

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik} + \varepsilon_i \qquad (4.1)$$

ya da

$$y_i = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} + \varepsilon_i \qquad (4.2)$$

şeklinde tanımlanır. Çoklu lineer regresyon modeli k tane açıklayıcı değişken ya da regresör içermektedir. Bu durumda en iyi regresyon modelinin seçimi için mevcut regresörler kullanılarak, 2^k tane aday model oluşturulabilir (Gunst ve Mason, 1980). Örneğin $k = 3$ alınsın. Bu durumda çoklu lineer regresyon modeli $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$ ya da örneklem için regresyon modeli $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i$ biçiminde olur. En iyi regresyon modelinin belirlenmesi için $2^3 = 8$ tane aday model oluşturulabilir. Bunlar: Sabit model:

$$Y = \beta_0 + \varepsilon; \qquad (4.3.a)$$

Bir açıklayıcı değişken içeren modeller:

$$\begin{aligned} &Y = \beta_0 + \beta_1 X_1 + \varepsilon, \\ &Y = \beta_0 + \beta_2 X_2 + \varepsilon, \\ &Y = \beta_0 + \beta_3 X_3 + \varepsilon; \end{aligned} \qquad (4.3.b)$$

İki açıklayıcı değişken içeren modeller:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon,$$
$$Y = \beta_0 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon, \quad (4.3.c)$$
$$Y = \beta_0 + \beta_1 X_1 + \beta_3 X_3 + \varepsilon;$$

Tam model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon \quad (4.3.d)$$

biçimindedir.

En iyi regresyon modelinin belirlenmesinde Çoklu belirleyicilik katsayısı R^2 veya Düzeltilmiş Çoklu belirleyicilik katsayısı $R^2_{Düzeltilmiş}$ ve Hata Kareleri Ortalamaları (HKO) kullanılabilir (Draper ve Smith, 1998). Eşit sayıda açıklayıcı değişken içeren modellerin karşılaştırılmasında Çoklu belirleyicilik katsayısı R^2 ve Farklı sayıda açıklayıcı değişken içeren modellerin karşılaştırılmasında Düzeltilmiş Çoklu belirleyicilik katsayısı $R^2_{Düzeltilmiş}$ değerleri kullanılır. En iyi regresyon modelinin belirlenmesinde R^2'si veya $R^2_{Düzeltilmiş}$'si yüksek, HKO'su düşük olan ve az sayıda açıklayıcı değişken içeren model tercih edilir (Montgomery ve ark., 2001).

Bazı durumlarda mevcut açıklayıcı değişkenler, yanıt değişkendeki toplam değişimi açıklamada yetersiz kalabilir böyle durumlarda regresyon modeline yeni açıklayıcı değişken ya da değişkenler eklenebilir. Bazı durumlarda ise mevcut açıklayıcı değişkenlerden bazıları yanıt değişkendeki toplam değişimi açıklamada etkileri ya da katkıları olmadığından çoklu lineer regresyon modelinden silinebilir ya da çıkarılabilir (Chatterjee ve ark., 2000).

Çoklu lineer regresyon modelindeki açıklayıcı değişken sayısının artması durumunda ileriye doğru seçim ya da geriye doğru ayıklama gibi adımsal regresyon analizi yöntemleri uygulanabilir (Miller, 1990).

Çoklu lineer regresyon modelindeki açıklayıcı değişken sayısının fazla olması durumunda ne geleneksel yöntemler ne de adımsal yöntemler kullanılamamaktadır (Bozdogan, 2003). Çoklu lineer regresyon modelindeki açıklayıcı değişken sayısının

fazla olması durumunda çoklu lineer regresyon modelinin oluşturulmasında Genetik Algoritma kullanılabilir (Wasserman ve Sudjianto, 1994; Wallet ve ark., 1996).

Çoklu lineer regresyon modeli oluşturulurken genellikle regresörlerin doğru fonksiyonel biçiminin bilindiği varsayılır. Ayrıca veride aykırı ya da sapan değerlerin ve etkili gözlemlerin bulunmadığı varsayılır. Bununla birlikte çoğu uygulamada bu gibi varsayımlar geçerli değildir. Bunları kontrol etmenin farklı yöntemleri vardır. Örneğin $\hat{y}_i$ oluşturulan değerleriyle y_i gözlenen değerleri arasındaki doğrusal ilişki çok düşükse oluşturulan modelin biçimi yanlış alınmıştır. Bu da regresörlerin fonksiyonel biçiminin yanlış alındığı anlamına gelir (Montgomery ve ark., 2001). Bazı gözlem değerlerinin verinin genel kümelenme yapısına aykırı olarak uzakta küçük bir küme oluşturması durumunda veride aykırı ya da sapan değerlerin bulunmadığı varsayımı bozulur. Ayrıca veride modelin eğimini değiştirebilecek ve kaldıraç görevi yapabilecek gözlem değerlerinin bulunması da etkili gözlemlerin bulunmadığı varsayımını bozar (Montgomery ve ark., 2001).

Uygulamada bu varsayımlar, çok az veri kümesi ya da çok özel durumlar için sağlanır. Hata ya da rezidü analiziyle, regresörler için doğru fonksiyonel biçimler kontrol edilebilir, sapan ya da aykırı değerler belirlenebilir ve etkin gözlemler kontrol edilebilir. Veride sapan ya da aykırı değerin bulunup bulunmadığı veya etkili ya da yüksek kaldıraç etkisi bulunan gözlemler mutlaka belirlenmelidir. Ayrıca bunların model üzerindeki etkileri de araştırılmalıdır. Tüm bunlar oluşturulan modelin yeterliliğini belirlemek için gereklidir. Oluşturulan modelin yeterliliğinin araştırılması, değişken seçimi problemi ile bağlantılıdır (Cox ve Snell, 1974).

Modelin doğru fonksiyonel biçiminin, verideki sapan ya da aykırı gözlem değerlerinin ve verideki etkin gözlem değerlerinin belirlenmesi problemlerinin eş zamanlı çözülmesi gerekse bile çoğu kez ardışık yaklaşım kullanılır. Önce değişken seçimi stratejisi kullanılır. Sonra sonuçta bulunan alt küme modeli doğru fonksiyonel belirtisi için sapan değerler için ve etkili gözlemler için kontrol edilir. Bu, birinci adımın tekrarlanması gerektiğini belirtebilir. Yeterli bir model oluşturmak için bir çok ardışık işlem gerekebilir.

Bu bölümde açıklanacak değişken seçimi yöntemlerinden hiçbiri, verilmiş bir veri kümesi için en iyi regresyon denklemini oluşturmayı garanti etmez. Aslında bir

tek en iyi denklem yoktur. Bununla birlikte iyi olanlara denk çeşitli denklemler vardır. Çünkü değişken seçimi algoritmaları, yoğun bilgisayar bağlantılıdır. Analizciler, belli bir yöntemin sonuçlarına güvenmemelidir. Analiz sonuçları kontrol edilmelidir. Değişken seçiminde kullanılan yöntemler, verinin yapısını araştırmadaki yöntemler olarak analizciler tarafından kullanılmalıdır. Regresyonda değişken seçimi ya da en iyi modelin oluşturulmasıyla ilgili çalışmalar Cox ve Snell (1974), Hocking (1972), Hocking (1976), Myers(1990), Hocking ve LaMotte (1973), Thompson (1978a) ve Thompson (1978b) tarafından yapılmıştır.

4.2. Yanlış Modelin Belirlenmesinin Sonuçları

Değişken seçiminde doğru olmayan modelin belirlenmesinin sonuçları incelenecektir. k tane aday değişken ya da regresör olsun. Bu regresörler $x_1,\ x_2, ...,\ x_k$ ile gösterilsin. Regresörler için $n \geq k+1$ tane gözlem değeri bulunsun. Yanıt değişken y olsun. k tane regresörü içeren tam model,

$$y_i = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} + \varepsilon_i \qquad i = 1, 2, ..., n \tag{4.4}$$

ya da matris gösterimiyle

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon} \tag{4.5}$$

biçimindedir. Aday regresörlerin listesi, bütün değişkenleri içersin. (4.4)'deki model, β_0 sabit terimini içermektedir. β_0 sabit terimi her zaman modele dahil edilmeye zorlanır. Diğer bir ifadeyle bütün regresyon denklemlerinin bir sabit terim içerdiği varsayılır (Montgomery ve ark., 2001). (4.4)'deki modelden silinen regresörlerin sayısı r olsun. Bu durumda modelde tutulan regresörlerin ya da değişkenlerin sayısı $p = k+1-r$ dir. Modele sabit terim de dahil edildiğinde, alt küme modeli orijinal regresörlerin $p-1 = k-r$ tanesini içerir. Bu durumda (4.5)'deki model,

$$\boldsymbol{y} = \boldsymbol{X}_p\boldsymbol{\beta}_p + \boldsymbol{X}_r\boldsymbol{\beta}_r + \boldsymbol{\varepsilon} \tag{4.6}$$

şekline dönüşür. Burada $\boldsymbol{X}$ matrisi, $\boldsymbol{X}_p$ ve $\boldsymbol{X}_r$ matrislerine parçalanmıştır. $\boldsymbol{X}$, $n \times k$ tipindedir. $\boldsymbol{X}_p$, $n \times p$ tipindedir. $\boldsymbol{X}_p$ matrisinin sütunları sabit terimi ve alt küme modelinde tutulan $p-1$ tane regresörü içermektedir. $\boldsymbol{X}_r$, $n \times r$ tipindedir. $\boldsymbol{X}_r$ matrisinin sütunları tam modelden silinen regresörleri içermektedir. Ayrıca $\boldsymbol{\beta}$ parametre vektörü, $\boldsymbol{\beta}_p$ ve $\boldsymbol{\beta}_r$ olarak parçalanır. Tam model için $\boldsymbol{\beta}$'nın en küçük kareler tahmini edicisi,

$$\hat{\boldsymbol{\beta}}^* = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} \tag{4.7}$$

dir. Hata varyansı σ^2'nin tahmini de

$$\hat{\sigma}_*^2 = \frac{\boldsymbol{y}'\boldsymbol{y} - \hat{\boldsymbol{\beta}}^{*'}\boldsymbol{X}'\boldsymbol{y}}{\boldsymbol{n}-\boldsymbol{k}-1} = \frac{\boldsymbol{y}'[\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X}')^{-1}\boldsymbol{X}']\boldsymbol{y}}{\boldsymbol{n}-\boldsymbol{k}-1} \tag{4.8}$$

olur. $\hat{\boldsymbol{\beta}}^*$ vektörünün parçaları, $\hat{\boldsymbol{\beta}}_p^*$ ve $\hat{\boldsymbol{\beta}}_r^*$ vektörleriyle temsil edilir. $\hat{\boldsymbol{y}}_i^*$ ile (4.6)'daki model için oluşturulan modelin değerleri gösterilsin. Bu durumda alt küme modeli,

$$\boldsymbol{y} = \boldsymbol{X}_p\boldsymbol{\beta}_p + \boldsymbol{\varepsilon} \tag{4.9}$$

şeklinde yazılır. Alt küme modelinde $\boldsymbol{\beta}_p$'nin en küçük kareler tahmin edicisi,

$$\hat{\boldsymbol{\beta}}_p = (\boldsymbol{X}_p'\boldsymbol{X}_p)^{-1}\boldsymbol{X}_p'\boldsymbol{y} \tag{4.10}$$

olur. Hata varyansının tahmini de

$$\hat{\sigma}^2 = \frac{y'y - \hat{\beta}_p{}' X_p' y}{n-p} = \frac{y'[I - X_p(X_p'X_p)^{-1}X_p']y}{n-p} \tag{4.11}$$

bulunur. $\hat{y}_i$ ile (4.9)'daki model için oluşturulan modelin değerleri gösterilsin.

Alt küme modelindeki $\hat{\beta}_p$ ve $\hat{\sigma}^2$ tahminlerinin özellikleri, Hocking (1972), Hocking (1976), Narula ve Ramberg(1972), Walls ve Weeks (1969) tarafından araştırılmıştır. Bu araştırma sonuçlar aşağıdaki gibi özetlenebilir:

1. $\hat{\beta}_p$'nin beklenen değeri,

$$\mathrm{E}(\hat{\beta}_p) = \beta_p + (X_p'X_p)^{-1}X_p'X_r\beta_r = \beta_p + A\beta_r \tag{4.12}$$

dir. Burada $A = (X_p'X_p)^{-1}X_p'X_r$ olup, bazen "alias matris" olarak adlandırılır. Böylece silinen değişkenlerin yerini tutan regresyon katsayıları sıfır olmadıkça yani $\beta_r = 0$ olmadıkça veya tutulan değişkenler silinen değişkenlere dik olmadıkça yani $X_p'X_r = 0$ olmadıkça $\hat{\beta}_p$, β_p'nin bir yanlı tahminidir.

2. $\hat{\beta}_p$ ve $\hat{\beta}^*$'nin varyansları sırasıyla,

$$\mathrm{V}(\hat{\beta}_p) = \sigma^2(X_p'X_p)^{-1} \tag{4.13}$$

ve

$$\mathrm{V}(\hat{\beta}^*) = \sigma^2(X'X)^{-1} \tag{4.14}$$

dir. Ayrıca $\mathrm{V}(\hat{\beta}_p^*)$ - $\mathrm{V}(\hat{\beta}_p)$ matrisi pozitif semidefinittir. Tam modeldeki parametrelerin en küçük kareler tahminlerinin varyansları, alt küme modelindeki yer alan parametrelerin varyanslarından büyüktür veya eşittir. Sonuç olarak, değişkenleri

silmek ya da modelden çıkarmak, kalan parametrelerin tahminlerinin varyanslarını arttırmaz.

3. $\hat{\beta}_p$, β_p'nin bir yanlı tahminiyken $\hat{\beta}^*$ böyle değildir. Alt küme modellerinden ve tam modelden alınan parametre tahminlerinin doğruluğunu hata kareler ortalamasına göre karşılaştırmak daha uygundur. $\hat{\theta}$, θ parametresinin bir tahminiyse $\hat{\theta}$'nın hata kareler ortalaması

$$\text{MSE}(\hat{\theta})=\text{V}(\hat{\theta})+[\text{E}(\hat{\theta})-\theta]^2 \tag{4.15}$$

olarak tanımlanır. $\hat{\beta}_p$'nın hata kareler ortalaması

$$\text{MSE}(\hat{\beta}_p)=\sigma^2(X_p'X_p)^{-1}+A\beta_r\beta_r'A' \tag{4.16}$$

dır. Eğer $\text{V}(\hat{\beta}_r^*)$-$\beta_r\beta_r'$ matrisi pozitif semidefinit ise $\text{V}(\hat{\beta}_p^*)$-$\text{MSE}(\hat{\beta}_p)$ matrisi de pozitif semidefinittir. Bu, silinen değişkenlerin regresyon katsayıları, tam modeldeki tahminlerinin standart hatalarından daha küçük olduğunda, alt küme modelindeki parametrelerin en küçük kareler tahmin edicisinin tam modelden karşılık gelen parametre tahminlerinin hata kareler ortalamasından daha küçük olduğu anlamına gelir.

4. Tam modeldeki $\hat{\sigma}_*^2$ parametresi, σ^2'nin yansız bir tahminidir. Bununla birlikte, alt küme modeli için,

$$\text{E}(\hat{\sigma}^2)=\sigma^2+\frac{\beta_r'X_r'[I-X_p(X_p'X_p)^{-1}X_p']X_r\beta_r}{n-p} \tag{4.17}$$

dir. $\hat{\sigma}^2$ genellikle, σ^2'nin yanlı tahminidir.

5. $\boldsymbol{x}' = [\boldsymbol{x}_p', \boldsymbol{x}_r']$ olsun. Bu noktadaki yanıt tahmin edilsin. Bu iş için tam model kullanılırsa oluşturulan modelin bu noktadaki tahmin edilen değeri $\boldsymbol{x}'\boldsymbol{\beta}$ ortalamalı ve $V(\hat{\boldsymbol{y}}^*) = \boldsymbol{\sigma}^2[1 + \boldsymbol{x}'(\boldsymbol{X}'\boldsymbol{X})\boldsymbol{x}]$ varyanslı olur. Bu durumda $\hat{\boldsymbol{y}}^* = \boldsymbol{x}'\hat{\boldsymbol{\beta}}^*$ dır. Bununla birlikte eğer alt model kullanılsaydı, bu noktada oluşturulan modelin beklenen değeri,

$$E(\hat{\boldsymbol{y}}) = \boldsymbol{x}_p'\boldsymbol{\beta}_p + \boldsymbol{x}_p'\boldsymbol{A}\boldsymbol{\beta}_r \tag{4.18}$$

ve bu noktada oluşturulan modelin hata kareler ortalaması,

$$MSE(\hat{\boldsymbol{y}}) = \boldsymbol{\sigma}^2[1 + \boldsymbol{x}_p'(\boldsymbol{X}_p'\boldsymbol{X}_p)^{-1}\boldsymbol{x}_p] + (\boldsymbol{x}_p'\boldsymbol{A}\boldsymbol{\beta}_r - \boldsymbol{x}_r'\boldsymbol{\beta}_r)^2 \tag{4.19}$$

olacaktı. Burada $\hat{\boldsymbol{y}} = \boldsymbol{x}_p'\hat{\boldsymbol{\beta}}_p$ dır. $\boldsymbol{x}_p'\boldsymbol{A}\boldsymbol{\beta}_r = 0$ olmadıkça (bu da genellikle $\boldsymbol{X}_p'\boldsymbol{X}_r\boldsymbol{\beta}_r = 0$ iken doğrudur.) $\hat{\boldsymbol{y}}$, $\boldsymbol{y}$'nin yanlı bir tahminidir. Ayrıca tam modeldeki varyans, alt küme modelindeki $\hat{\boldsymbol{y}}$'nın varyansından küçük değildir. Hata kareler ortalaması cinsinden $V(\hat{\boldsymbol{y}}^*) \geq MSE(\hat{\boldsymbol{y}})$ eşitsizliği, $V(\hat{\boldsymbol{\beta}}_r^*) - \boldsymbol{\beta}_r\boldsymbol{\beta}_r'$ matrisi pozitif semidefinit iken sağlanır.

Değişken seçimi ile ilgili nedenler aşağıdaki gibi özetlenebilir:

Modelden bazı değişkenler silinerek, modelde tutulan değişkenlerin parametre tahminlerinin doğruluğu geliştirilebilir. Bu oluşturulan modelde yanıt değişkenin varyansı için de doğrudur.

Modelden bazı değişkenlerin silinmesi durumunda, modelde bulunan değişkenlerin katsayılarının tahminleri ve yanıt değişkenin tahmini yanlı olur. Bununla birlikte silinen değişkenlerin modele etkileri çok küçükse yanlı tahminlerin hata kareler ortalamaları, yansız tahminlerin varyansından daha küçük olur. Yani, yanlılık miktarı varyanstaki azalmadan daha küçüktür. Modelde tutulan önemsiz değişkenler için tehlike söz konusudur. Bunlar sıfır katsayılı değişkenler veya tam modeldeki standart hatalardan daha küçük katsayıları olan değişkenlerdir. Bu tehlike

parametre tahminlerinin varyansındaki ve tahmin edilen yanıttaki artıştan kaynaklanır.

Bazı regresyon modelleri tarihsel kayıtlardan alınmış rastgele verilerden oluşmaktadır (Box ve ark., 1978). Rastgele veriler genelde aykırı ya da sapan değerleri, etkin gözlemleri, veri toplamadaki değişimlerden oluşan tutarsızlıkları ve zamana karşı bilgi-işlem sistem hatalarını içerir. Verideki bu hatalar, değişken seçimi sürecinde büyük etki yaratabilir ve doğru modeli belirleyememe problemine neden olabilir. Rastgele verideki en genel problem, kontrol edilmiş regresörleri bulmaktır. Kontrol edilmiş regresörler daha tutarlıdır. Ayrıca etkili değişkenlerdir. Regresörler, yanıtı doğru sınırlarda tutmak için kontrol edilmelidir. Verideki bu rastgele hataların etkileri en küçük kareler uyumunda önemsiz görünebilir.

4.3. Regresörlerin Bir Alt Kümesini Seçmek İçin Kriterler

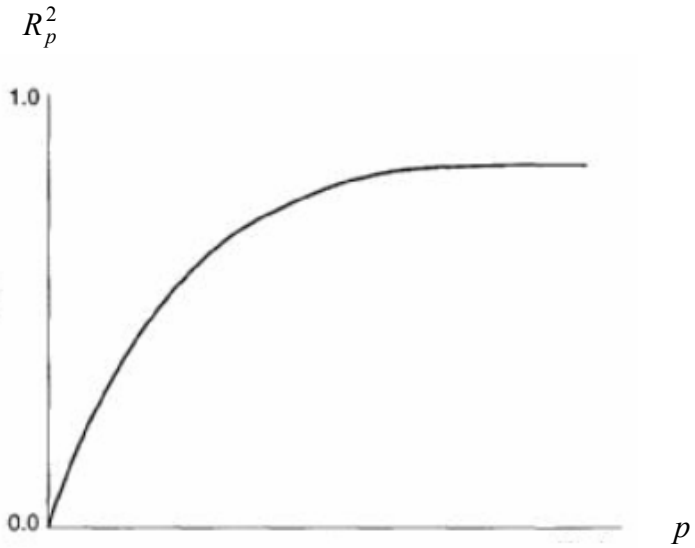
Regresyon analizinde değişken seçimi probleminin iki aşamalı çözümü vardır. Birinci aşamada alt küme modelleri üretilir. İkinci aşamada ise bir alt kümenin diğerinden daha iyi olup olmadığına karar verilir. Bu bölümde alt küme regresyon modellerini değerlendirmek ve karşılaştırmak için kriterler incelenecektir. Ayrıca değişken seçimi için hesaplama yöntemleri açıklanacaktır.

4.3.1. Çoklu Belirleyicilik Katsayısı

Regresyon modelinin uygunluğun bir ölçütü çoklu belirleyicilik katsayısıdır. Çoklu belirleyicilik katsayısı R^2 ile gösterilir. R_p^2, p terimli ve $p-1$ regresörlü bir alt küme modeli için çoklu belirleyicilik katsayısını göstersin. Modelde β_0 sabit terimi bulunmaktadır. Bu durumda R_p^2,

$$R_p^2 = \frac{SS_R(p)}{Syy} = 1 - \frac{SS_E(p)}{Syy} \tag{4.20}$$

olarak tanımlanır (Montgomery ve ark., 2001). Burada $SS_R(p)$, $SS_E(p)$ ve Syy sırasıyla p terimli alt küme modeli için regresyon kareler toplamını, hata kareler toplamını ve genel kareler toplamını göstermektedir. p'nin her değeri için R_p^2'nin $\binom{k}{p-1}$ tane değeri vardır. p artarsa R_p^2 de artar. $p = k+1$ olduğunda R_p^2 maksimum olur. Bu nedenle analizciler, yeni bir değişken gerekli olmayacak noktaya kadar modele regresörler ekleyerek bu kriteri kullanırlar. Modele yeni bir regresörün eklenmesi durumunda R_p^2'de az da olsa bir artış sağlanır. Genel yaklaşım Şekil 4.1'de gösterilmiştir.



Şekil 4.1. Modeldeki terim sayısı p değerlerine karşı çoklu belirleyicilik katsayısı R_p^2'nin grafiği (Montgomery ve ark., 2001).

Bir alt küme regresyon modeli için R^2'nin optimum değerini arama yerine R^2'nin memnun edici ve beklentileri karşılayan değeri aranmalıdır. Aitkin(1974), bu probleme bir çözüm olarak tam model için R^2'den anlamlı olarak farklı olmayan yeni bir çoklu belirleyicilik katsayısı oluşturdu. Oluşturulan yeni çoklu belirleyicilik katsayısı,

$$R_0^2 = 1-(1-R_{k+1}^2)(1+d_{\alpha,n,k}) \text{ ve } d_{\alpha,n,k} = \frac{kF_{\alpha,n,n-k-1}}{n-k-1} \quad (4.21)$$

olarak tanımlanmıştır.

Genel olarak modele dahil edilecek regresörlerin sayısına karar vermek için R^2'yi bir kriter olarak kullanmak doğru olmaz. Bununla birlikte k tane değişken için $\begin{pmatrix} k \\ p-1 \end{pmatrix}$ tane alt küme modelini karşılaştırmada R_p^2 kullanılabilir. R_p^2 değeri büyük olan modeller tercih edilir.

4.3.2. Düzeltilmiş Çoklu Belirleyicilik Katsayısı

R^2'nin yorumlanmasındaki zorluklardan kaçınmak için, bazı araştırmacılar düzeltilmiş R^2'yi kullanmayı tercih etmişlerdir (Haitovski, 1969). Düzeltilmiş R^2,

$$\overline{R}_p^2 = 1-\left(\frac{n-1}{n-p}\right)(1-R_p^2) \quad (4.22)$$

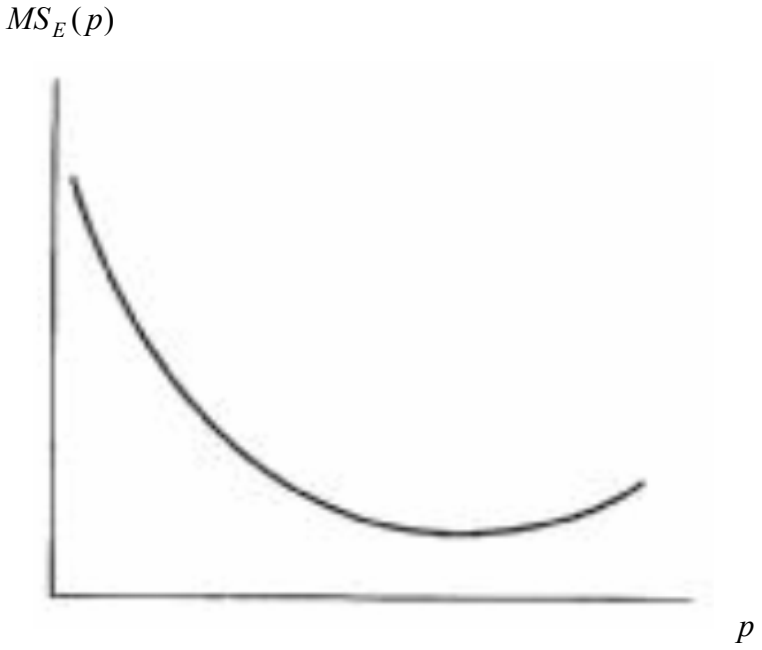
olarak tanımlanır. $\overline{R}_p^2$ istatistiği, R_p^2 istatistiğine göre daha tutarlıdır. R_p^2, modele eklenen yeni regresörlerden fazla etkilenmez (Edwards 1969; Seber 1977). Modele s tane regresör eklenirse $\overline{R}_{p+s}^2$'nin değerinin, $\overline{R}_p^2$'nin değerinden daha fazla olabilmesi için gerek ve yeter koşul modele eklenen s tane regresörün önemini test etmek için kısmi F-istatistiğinin değerinin 1'i aşmasıdır (Edwards, 1969). Sonuç olarak bir alt kümenin oluşturduğu optimum modeli seçmek için bir kriter, maksimum $\overline{R}_p^2$'ye sahip modeli seçmektir (Montgomery ve ark., 2001). Bu kritere denk olan başka bir kriter aşağıdaki açıklanmıştır.

4.3.3. Hata Kareler Ortalaması

Bir alt küme regresyon modeli için hata kareler ortalaması,

$$MS_E(p) = \frac{SS_E(p)}{n-p} \tag{4.23}$$

olarak tanımlanmıştı (Montgomery ve ark., 2001). Hata kareler ortalaması bir modeli değerlendirme kriteri olarak kullanılabilir. p değerlerine karşı $MS_E(p)$ değerlerinin grafiği Şekil 4.2'de gösterilmiştir.



Şekil 4.2. p değerlerine karşı $MS_E(p)$ değerlerinin grafiği (Montgomery ve ark., 2001).

p değeri arttıkça $SS_E(p)$ değeri daima azalır. Bununla birlikte p değeri arttıkça $MS_E(p)$ değeri önce azalır, sonra dengelenir ve daha sonra çok az artar. $MS_E(p)$'deki en son artış, modele yeni bir regresör eklendiğinde ortaya çıkan $SS_E(p)$'deki azalmanın (4.23)'deki eşitliğin paydasındaki bir serbestlik derecesi

kaybını karşılamaya yeterli olduğu zaman meydana gelir. p terimli modele yeni bir regresör eklendiğinde hata kareler toplamındaki azalma $MS_E(p)$'dekinden daha az olursa, $MS_E(p+1)$'in $MS_E(p)$'den daha büyük olmasına neden olur. $MS_E(p)$ kriteri uygulandığında p'ye karşı $MS_E(p)$'nin grafiğinde p'nin seçiminde aşağıdaki kurallar geçerlidir.

1. Minimum $MS_E(p)$.

2. p değerinde $MS_E(p)$'nin değeri, tam model için MS_E'nin değerine yaklaşık olarak eşit olsun.

3. Minimum $MS_E(p)$'nin yukarı döndüğü nokta yakınında p'nin bir değeri $MS_E(p)$'yi minimum yapan alt küme regresyon modeli $\overline{R}_p^2$'yi maksimum yapar. Bunu görmek için

$$\begin{aligned}\overline{R}_p^2 &= 1-\left(\frac{n-1}{n-p}\right)(1-R_p^2)\\ &= 1-\frac{n-1}{n-p}\frac{SS_E(p)}{S_{yy}}\\ &= 1-\frac{n-1}{S_{yy}}\frac{SS_E(p)}{n-p}\\ &= 1-\frac{n-1}{S_{yy}}MS_E(p)\end{aligned}$$

yazılır. Böylece, minimum $MS_E(p)$ ve maksimum $\overline{R}_p^2$ kriterleri eşdeğerdir.

4.3.4. Mallows'un C_p İstatistiği

Mallows (1964, 1966 ve 1973), oluşturulan modelin değerinin hata kareler ortalamasına dayalı bir kriter önermiştir. Bu kriter,

$$E[\hat{y}_i - E(y_i)]^2 = [E(y_i) - E(\hat{y}_i)]^2 + V(\hat{y}_i) \qquad (4.24)$$

olarak tanımlanır. Dikkat edilirse $E(y_i)$, doğru regresyon eşitliğinden beklenen yanıt ve $E(\hat{y}_i)$ ise p terimli alt küme modelinden beklenen yanıttır. Böylece $E(y_i) - E(\hat{y}_i)$, i-inci veri noktasındaki yanlılıktır. Sonuç olarak (4.24)'deki eşitliğinin sağ tarafındaki iki terim sırasıyla hata kareler ortalamasının yanlılık karesi ve varyans elemanlarıdır. p terimli model için toplam yanlılığın karesi,

$$SS_B(p) = \sum_{i=1}^{n} [E(y_i) - E(\hat{y}_i)]^2 \tag{4.25}$$

ve standartlaştırılmış toplam MS_E,

$$\begin{aligned} \Gamma_p &= \frac{1}{\sigma^2}\left(\sum_{i=1}^{n} [E(y_i) - E(\hat{y}_i)]^2 + \sum_{i=1}^{n} V(\hat{y}_i) \right) \\ &= \frac{SS_B(p)}{\sigma^2} + \frac{1}{\sigma^2} \sum_{i=1}^{n} V(\hat{y}_i) \end{aligned} \tag{4.26}$$

olarak tanımlansın. Burada $\sum_{i=1}^{n} V(\hat{y}_i) = p\sigma^2$ olarak gösterilebilir. p terimli model için hata kareler toplamının beklenen değeri,

$$E[SS_E(p)] = SS_B(p) + (n-p)\sigma^2 \tag{4.27}$$

dir. $\sum_{i=1}^{n} V(\hat{y}_i)$ ve $SS_B(p)$ (4.26)'da yerine koyulduğunda,

$$\begin{aligned} \Gamma_p &= \frac{1}{\sigma^2}\left(\sum_{i=1}^{n} E[SS_E(p)] - (n-p)\sigma^2 + p\sigma^2 \right) \\ &= \frac{E[SS_E(p)]}{\sigma^2} - n + 2p \end{aligned} \tag{4.28}$$

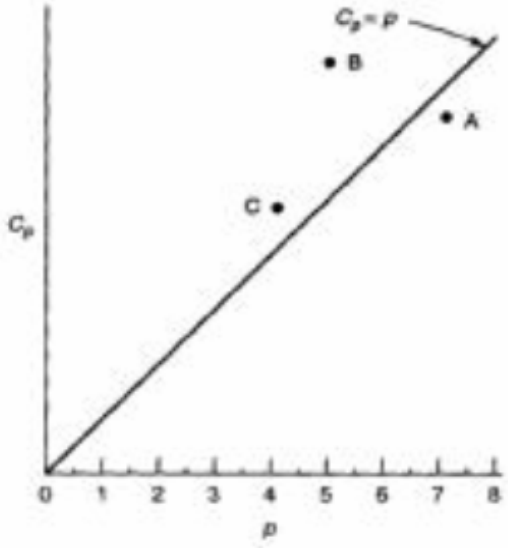
elde edilir. $\hat{\sigma}^2$, σ^2'nın iyi bir tahmini olsun. Bu durumda $E[SS_E(p)]$'yi gözlenen değer $SS_E(p)$ ile değiştirmek Γ_p'nin yeni bir tahmini oluşturur. Bu da,

$$C_p = \frac{SS_E(p)}{\hat{\sigma}^2} - n + 2p \tag{4.29}$$

sonucunu verir. p terimli model göz ardı edilebilir yanlılığa sahip ise $SS_B(p)$=0 olur. Sonuç olarak, $E[SS_B(p)] = (n-p)\sigma^2$ elde edilir. Buradan da

$$E[C_p \,|\, Bias = 0] = \frac{(n-p)\sigma^2}{\hat{\sigma}^2} - n + 2p = p \tag{4.30}$$

bulunur.



Şekil 4.3. p değerlerine karşılık C_p değerlerinin grafiği (Montgomery ve ark., 2001).

Şekil 4.3'te gösterildiği gibi , C_p kriterini kullanmak, her regresyon eşitliği için p'nin bir fonksiyonu olan C_p'nin bir grafiğini oluşturmak anlamındadır. Göz

ardı edilebilir yanlılığa sahip regresyon denklemleri için $C_p = p$ çizgisinin yakınına düşer. Yani p, C_p'nin değerini alır. Bu durum Şekil 4.3'te A noktası olarak gösterilmiştir. Belirli bir yanlılığa sahip denklemler bu çizginin üstüne düşecektir. Bu durum da Şekil 4.3'te B noktası olarak gösterilmiştir. Genellikle C_p'nin küçük değerleri istenir. Örneğin; Şekil 4.3'teki C noktası, $C_p = p$ çizgisinin üstünde olmasına rağmen, A noktasının aşağısındadır. Bu nedenle daha az toplam hata ile modeli temsil eder. Tahminin ortalama hatasını azaltmak için denklemdeki bazı yanlılıkları kabul etmek daha iyidir.

C_p'yi hesaplamak için σ^2'nin yansız bir tahminine gereksinim vardır. Bu amaç için tam denklem için hata kareler ortalaması kullanılır. Bu, tam model için $C_p = p = k+1$ olmasını zorlar.

σ^2'nin bir tahmini olarak tam modeldeki $MS_E(k+1)$'yi kullanmak, tam modelin göz ardı edilebilir yanlılığa sahip olduğunu varsayar. Tam modelin fazla sayıda regresöre sahip olmasının anlamlı olarak modele katkısı yoktur. Yani Sıfır regresyon katsayılarının olması durumu. Öyleyse $MS_E(k+1)$ çoğu kez σ^2'yi tekrar tekrar tahmin edecektir. Sonuç olarak C_p'nin değerleri küçük olacaktır.

4.4. Regresyon ve Model Değerlendirme Ölçütlerinin Kullanımı

Görüldüğü gibi altküme regresyon modellerini değerlendirmek için kullanılan bir çok kriter vardır. Model seçimi için kullanılacak kriter kesinlikle modelin kastedilen kullanımıyla ilgili olmalıdır. Regresyonun; (1) veri tanımlama, (2) kestirim ve tahmin, (3) parametre tahmini ve (4) kontrol olmak üzere bir çok olası kullanımı vardır.

Amaç, verilmiş bir yöntem için iyi bir tanımlama elde etmekse veya karmaşık bir sistemin modelini elde etmekse, hata kareler toplamı küçük olan regresyon denklemleri için bir araştırma gösterilmiştir. Bütün k tane aday regresörleri kullanarak SS_E minimum yapıldığında, SS_E sonuçlarında küçük artmalar

olabileceğinden daima bazı değişkenlerin elenmesi önerilir. y'deki toplam değişim açıklanırken, birkaç regresörlü sistemin uygun olduğu söylenir.

Çoğu kez regresyon denklemleri, gözlemlerin ön tahmini veya yanıtın ortalamasının tahmini için kullanılır. Genel olarak, kestirimin hata kareler ortalamasının minimum yapıldığı regresörler seçilir. Bu da az etkili regresörlerin modelden silineceği anlamına gelir. Birisi, bir alt küme üretme yöntemi tarafından oluşturulan aday denklemleri değerlendirmek için

$$PRESS_p = \sum_{i}^{n} [y_i - \hat{y}_{(i)}]^2 = \sum_{i}^{n} \left(\frac{e_i}{1 - h_{ii}} \right)^2 \tag{4.31}$$

şeklinde tanımlanan $PRESS_p$ istatistiğini kullanabilir (Chatterjee ve ark., 2000). $PRESS_p$'nin küçük değerine dayalı bir alt küme regresyon modeli seçilebilir. $PRESS_p$, özellikle tahmin problemi için sezgisel başvurmaya sahip olduğundan, hata kareler toplamının basit bir fonksiyonu değildir. Bu kritere dayalı değişken seçimi için bir algoritma geliştirilmesi kolay değildir. $PRESS_p$ istatistiği alternatif modelleri ayırt etmede kullanışlıdır (Montgomery ve ark., 2001).

Parametre tahminiyle ilgileniliyorsa hem değişken silmenin sonucundaki yanlılıklar, hem de tahmin edilen katsayıların varyansları göz önünde bulundurulmalıdır. Regresörler yüksek içilişkili olduğunda, regresyon katsayılarının en küçük kareler tahminleri son derece zayıf olur. Regresyon modeli kontrol için kullanıldığında, parametrelerin doğru tahminleri çok önemlidir.

4.5. Değişken Seçimi İçin Hesaplama Teknikleri

En son denklemde kullanmak için değişkenlerin alt kümesini bulmada, aday regresörlerin çeşitli kombinasyonlarıyla model oluşturma dikkate alınmalıdır. Bu bölümde, alt küme regresyon modellerini üretmek için bir çok hesaplama yöntemi incelenecek ve bu modellerin değerlendirilmesi için kriterler açıklanacaktır.

4.5.1 Olası Bütün Regresyonlar

Bu yöntem, araştırmacının sabit terimli model (Bu model regresör içermemektedir.), bir-aday regresör içeren model, iki-aday regresör içeren model, ..., k-aday regresör içeren model gibi bütün regresyon denklemlerini oluşturmasını gerektirir. Oluşturulan bu aday modeller kritere göre değerlendirilir ve "en iyi" regresyon modeli seçilir (Draper ve Smith, 1998). β_0 sabit teriminin bütün denklemlere dahil edildiğini varsayalım. k tane regresör için 2^k tane toplam aday regresyon denklemi vardır. Örneğin, $k = 4$ için $2^4 = 16$ aday regresyon denklem vardır. $k = 10$ için $2^{10} = 1024$ aday regresyon denklemi vardır (Gunst ve Mason, 1980). Aday regresörlerin sayısı arttıkça incelenmesi gereken denklemlerin sayısı da hızlıca artmaktadır. Yeterli bilgisayar kodlarının gelişiminden önce, birkaç regresörden fazla regresör içeren problemler için bütün regresyonları üretmek pratik değildi. Yüksek hızlı bilgisayarların ortaya çıkmasıyla bütün regresyonlar için etkili algoritmaların gelişimi sağlandı. Çoklu lineer regresyon modelinde olası bütün regresyon modellerini kullanarak değişken seçimi örneği, Örnek 4.1'de verilmiştir.

Örnek 4.1. Gözlenen veriler Tablo 4.1'de gösterilmiştir.

Tablo 4.1. Örnek 4.1. için Hald Çimento verisi (Montgomery ve ark., 2001).

Gözlem i	y_i	x_{i1}	x_{i2}	x_{i3}	x_{i4}
1	78.5	7	26	6	60
2	74.3	1	29	15	52
3	104.3	11	56	8	20
4	87.6	7	31	8	47
5	95.9	7	52	6	33
6	109.2	11	55	9	22
7	102.7	3	71	17	6
8	72.5	1	31	22	44
9	93.1	2	54	18	22
10	115.9	21	47	4	26
11	83.8	1	40	23	34
12	113.3	11	66	9	12
13	109.4	10	68	8	12

Bu veri, değişken seçimine yaklaşım olarak bütün aday regresyon modellerini açıklamak için kullanılacaktır. Veride dört tane regresör bulunmaktadır. Yani $k = 4$'tür. Aday regresyon modellerinin sayısı $2^4 = 16$'dır. 16 aday regresyon modellerinin oluşumunun sonuçları Tablo.4.2'de gösterilmiştir. $R^2{}_p$, $\overline{R}^2{}_p$, $MS_E(p)$ ve C_p istatistiklerinin değerleri de bu tabloda verilmiştir.

Tablo.4.2. Örnek 4.1 için bütün aday regresyon modellerinin özeti (Montgomery ve ark., 2001).

Modeldeki Regresör Sayısı	p	Modeldeki Regresörler	$SS_E(p)$	$R^2{}_p$	$\overline{R}^2{}_p$	$MS_E(p)$	C_p
Yok	1	Yok	2715.7635	0	0	226.3136	442.92
1	2	x_1	1265.6867	0.53395	0.49158	115.0624	202.55
1	2	x_2	906.3363	0.66627	0.63593	82.3942	142.49
1	2	x_3	1939.4005	0.28587	0.22095	176.3092	315.16
1	2	x_4	883.8669	0.67459	0.64495	80.3515	138.73
2	3	$x_1\ x_2$	57.904	0.97868	0.97441	5.7904	2.68
2	3	$x_1\ x_3$	1227.0721	0.54817	0.45780	122.7073	198.10
2	3	$x_1\ x_4$	74.7621	0.97247	0.96697	7.4762	5.50
2	3	$x_2\ x_3$	415.4427	0.84703	0.81644	41.5443	62.44
2	3	$x_2\ x_4$	868.8801	0.68006	0.61607	86.8880	138.23
2	3	$x_3\ x_4$	175.7380	0.93529	0.92235	17.5738	22.37
3	4	$x_1\ x_2\ x_3$	48.1106	0.98228	0.97638	5.3456	3.04
3	4	$x_1\ x_2\ x_4$	47.9727	0.98234	0.97645	5.3303	3.02
3	4	$x_1\ x_3\ x_4$	50.8361	0.98128	0.97504	5.6485	3.50
3	4	$x_2\ x_3\ x_4$	73.8145	0.97282	0.96376	8.2017	7.34
4	5	$x_1\ x_2\ x_3\ x_4$	47.8636	0.98238	0.97356	5.9829	5.00

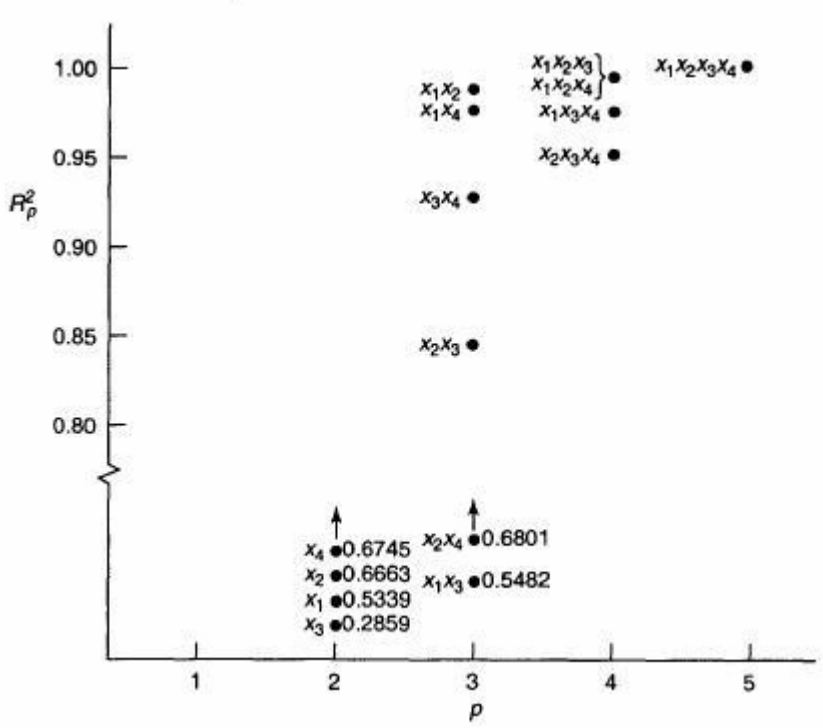
Tablo 4.3, regresyon katsayılarının en küçük kareler tahminini göstermektedir. Örneğin, model sadece x_2'yi içerdiğinde, x_2'nin etkisinin en küçük kareler tahmini 0.789 dur. Eğer x_4 modele eklenirse, x_2'nin etkisi 0.311 olup yüzde 50 civarında bir azalma vardır. Ayrıca x_3'ün eklenmesi x_2'nin etkisini -0.923 olarak değiştirir. Bu durumda her bir regresyon katsayısının en küçük kareler tahmini,

modeldeki diğer regresörlere bağlıdır. Modele değişkenler eklendiğinde veya modelden değişkenler çıkarıldığında Hald çimento verisinde gözlenen regresyon katsayılarındaki büyük değişim, dört regresör arasında gerçek korelasyon olduğunu belirtir. Bu çoklu iç ilişki probleminin bulunduğunu belirtir.

Tablo 4.3. Örnek 4.1 için bütün aday regresyon modellerindeki parametrelerin en küçük kareler yöntemiyle elde edilen tahminleri (Montgomery ve ark., 2001).

Modeldeki Değişkenler	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
x_1	81.479	1.869			
x_2	57.424		0.789		
x_3	110.203			-1.256	
x_4	117.568				-0.738
$x_1\ x_2$	52.577	1.468	0.662		
$x_1\ x_3$	72.349	2.312		0.494	
$x_1\ x_4$	103.097	1.440			-0.614
$x_2\ x_3$	72075		0.731	-1.008	
$x_2\ x_4$	94.160		0.311		-0.457
$x_3\ x_4$	131.282			-1.200	-0.724
$x_1\ x_2\ x_3$	48.194	1.696	0.657	0.250	
$x_1\ x_2\ x_4$	71.648	1.452	0.416		-0.237
$x_1\ x_3\ x_4$	203.642		-0.923	-1.448	-1.557
$x_2\ x_3\ x_4$	111.684	1.052		-0.410	-0.643
$x_1\ x_2\ x_3\ x_4$	62.405	1.551	0.510	0.102	-0.144

Alt küme modellerini değerlendirmek için R_p^2 kriterini göz önünde bulunduralım. p değerlerine karşı $R^2{}_p$ değerlerinin grafiği Şekil 4.4'de gösterilmiştir.



Şekil 4.4. p değerlerine karşı $R^2{}_p$ değerlerinin grafiği (Montgomery ve ark., 2001).

Şekil 4.4 incelendiğinde, modeldeki iki regresörden sonra eklenen değişkenlerin R^2 değerine katkısının az olduğu görülür. (x_1, x_2) ve (x_1, x_4) iki regresörlü modellerin her ikisinin de R^2 değerleri aynı olup bu kriterin ifadesiyle son regresyon denklemi olarak hangi model seçilirse seçilsin, küçük farklılık yaratacaktır. Draper ve Smith (1998), (x_1, x_4)'lü modeli kullanmayı önermişlerdir. Çünkü, x_4 en iyi bir-regresörlü modeli sağlamaktadır. (4.21)'deki eşitlikte $\alpha = 0.05$ alınırsa R_0^2 değeri,

$$R_0^2 = 1 - \left(1 - R_5^2\right)\left(1 + \frac{4F_{0.05,4,8}}{8}\right)$$

$$= 1 - 0.01762\left[1 + \frac{4(3.84)}{8}\right]$$

$$= 0.94855\,.$$

bulunur. $R_p^2 > R_0^2 = 0.94855$ olduğundan herhangi bir regresyon modeli için R^2 değeri R_{k+1}^2 değerinden anlamlı olarak farklı değildir. Tablo 4.2'deki birçok model bu kriteri sağlamaktadır. Son modeli seçme işi hala belli değildir.

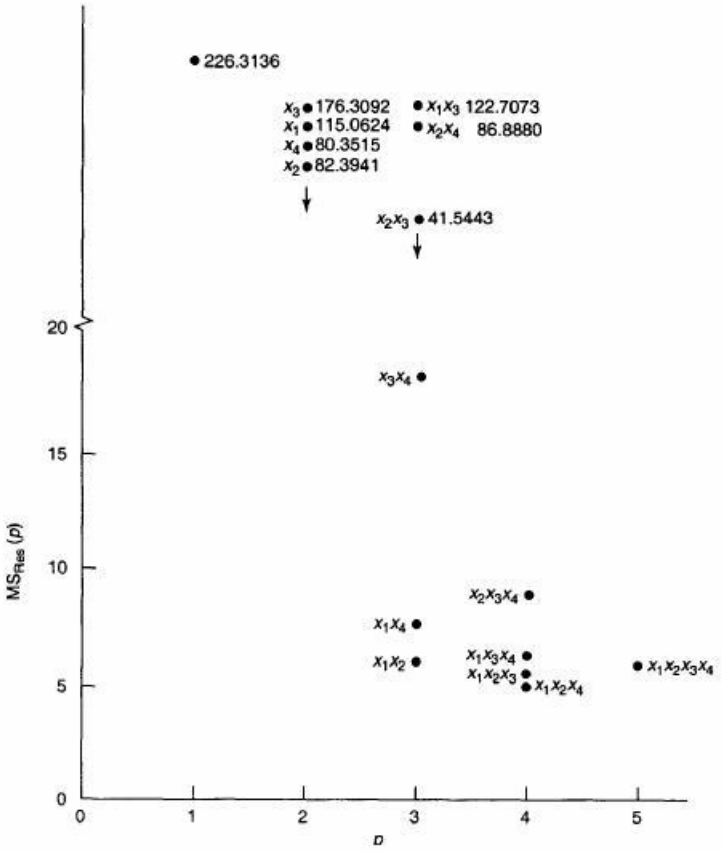
x_i ile x_j arasındaki ve x_i ile y arasındaki ikili korelasyonu hesaplamak yararlı olacaktır. Bu basit korelasyonlar Tablo.4.4'te gösterilmiştir. Dikkat edilirse (x_1, x_3) ve (x_2, x_4) regresör çiftleri yüksek korelasyona sahiptir. Bu değerler sırasıyla $r_{13} = -0.824$ ve $r_{24} = -0.973$ şeklindedir.

Tablo 4.4. Örnek 4.1'deki Hald çimento verisi için basit korelesyon matrisi (Montgomery ve ark., 2001).

	x_1	x_2	x_3	x_4	y
x_1	1.0				
x_2	0.229	1.0			
x_3	-0.824	-0.139	1.0		
x_4	-0.245	-0.973	0.030	1.0	
y	0.731	0.816	-0.535	-0.821	1.0

Sonuç olarak x_1 ve x_2 veya x_1 ve x_4 modelde mevcut iken modele fazladan regresör eklemek anlamlı değildir. Ayrıca Tablo 4.4'teki değerlere bakıldığında değişkenler arasında yüksek ilişki olduğu görülür.

p değerlerine karşı $MS_E(p)$ değerlerinin grafiği Şekil 4.5'te gösterilmiştir. En düşük hata kareler ortalamasına sahip model (x_1, x_2, x_4) olup bunun hata kareler ortalaması $MS_E(4) = 5.3303$ dür. Dikkat edilirse, beklendiği gibi $MS_E(p)$'yi minimum yapan model R_p^2'yi maksimum yapmaktadır. Bununla birlikte, (x_1, x_2, x_3) ve (x_1, x_3, x_4) üçlü modellerinin ve (x_1, x_2) ve (x_1, x_4) ikili modellerinin hata kareler ortalamaları karşılaştırılabilir. Eğer (x_1, x_2) veya (x_1, x_4) modelde bulunursa, modele fazladan regresör eklendiğinde hata kareler ortalamalarındaki azalma daha az olur. (x_1, x_2), (x_1, x_4)'den daha uygun bir altküme modelidir. Çünkü hata kareler ortalaması daha küçüktür.



Şekil 4.5. p değerlerine karşı $MS_E(p)$ değerlerinin grafiği (Montgomery ve ark., 2001).

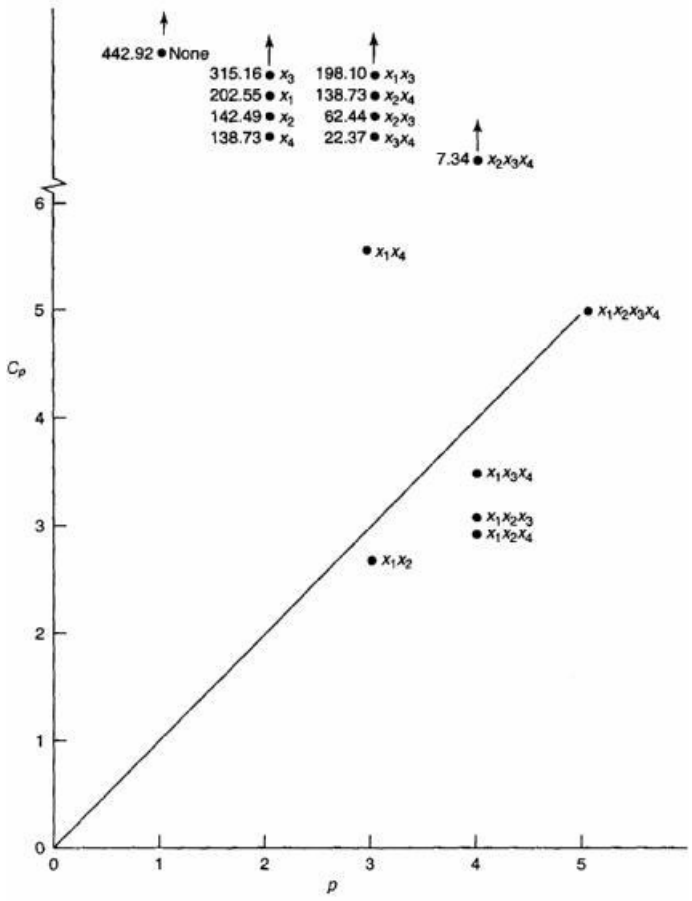
p değerlerine karşı C_p grafiği Şekil 4.6'da gösterilmiştir. Hesaplamaları açıklamak için $\hat{\sigma} = 5.9829$ olarak alalım. Bu değer tam model için MS_E'nin değeridir. (x_1, x_4) modeli için C_3'ü hesaplayalım. C_3,

$$C_3 = \frac{SS_E(3)}{\hat{\sigma}^2} - n + 2p$$

$$= \frac{74.7621}{5.9829} - 13 + 2(3)$$

$$= 5.50$$

bulunur.



Şekil 4.6. p değerlerine karşı C_p grafiği (Montgomery ve ark., 2001).

Tablo 4.5. Hald çimento verisi için iki modelin karşılaştırılması.

	$\hat{y} = 52.58 + 1.468x_1 + 0.662x_2$ [a]			$\hat{y} = 71.65 + 1.452x_1 + 0.416x_2 - 0.237x_4$ [b]		
Gözlem i	e_i	h_{ii}	$[e_i/(1-h_{ii})]^2$	e_i	h_{ii}	$[e_i/(1-h_{ii})]^2$
1	-1.5740	0.25119	4.4184	0.0617	0.52058	0.0166
2	-1.0491	0.26189	2.0202	1.4327	0.27670	3.9235
3	-1.5147	0.11890	2.9553	-1.8910	0.13315	4.7588
4	-1.6585	0.24225	4.7905	-1.8016	0.24431	5.6837
5	-1.3925	0.08362	2.3091	0.2562	0.35733	0.1589
6	4.0475	0.11512	20.9221	3.8982	0.11737	19.5061
7	-1.3031	0.36180	4.1627	-1.4287	0.36341	5.0369
8	-2.0754	0.24119	7.4806	-3.0919	0.34522	22.2977
9	1.8245	0.17195	4.9404	1.2818	0.20881	2.6247
10	1.3625	0.55002	9.1683	0.3539	0.65244	1.0368
11	3.2643	0.18402	16.0037	2.0977	0.32105	9.5458
12	0.8628	0.19666	1.1535	1.0556	0.20040	1.7428
13	-2.8934	0.21420	13.5579	-2.2247	0.25923	9.0194
$PRESS_{x_1,x_2} = \overline{93.8827}$				$PRESS_{x_1,x_2,x_4} = \overline{85.3516}$		
[a] $R^2_{\text{Prediction}} = 0.9654$ $VIF_1 = 1.06, VIF_2 = 1.06.$				[b] $R^2_{\text{Prediction}} = 0.9684$ $VIF_1 = 1.07, VIF_2 = 18.78, VIF_3 = 18.94$		

Şekil 4.6'daki grafik incelendiğinde uygun olarak dört tane model bulunur. Bunlar (x_1, x_2), (x_1, x_2, x_3), (x_1, x_2, x_4) ve (x_1, x_3, x_4)'tür. Ek faktörler dikkate alınmaksızın (x_1, x_2)'yi son model olarak seçmek uygun olur. Çünkü bu model, en küçük C_p değerine sahiptir.

Bu örnek, bütün aday regresyonla model oluşturma ile bir tutulan hesaplama yöntemlerini açıklamıştır. Dikkat edilirse en iyi regresyon modelinin kesin olarak seçimi yoktur. Çoğu kez farklı kriterlerin farklı denklemler önerdiği görülür. Örneğin, en küçük C_p'li denklem (x_1, x_2) ve en küçük MS_E'li denklem (x_1, x_2, x_4)'tür. Bütün son aday modellere, sapan ya da aykırı değerleri, kaldıraç etkisi yapan etkin gözlem değerlerini ve çoklu iç ilişki bilgilerini içeren uygunluk testleri yapılmalıdır. Tablo 4.5, (x_1, x_2) ve (x_1, x_2, x_4) modellerini *PRESS* ve varyans şişirme faktörlerine (*VIFs*) göre incelemiştir. Her iki modelin *PRESS* değerleri birbirine çok benzerdir. En düşük MS_E'li denklem için hata kareler toplamının yaklaşık olarak iki katıdır. *PRESS*'den hesaplanan tahmin için R^2, her iki modelde de benzerdir. Bununla birlikte (x_1, x_2, x_4)'deki çok büyük *VIF*'lerden

de görüldüğü gibi x_2 ve x_4 yüksek derecede çoklu iç ilişkiye sahiptir. Her iki model de denk $PRESS$ istatistiklerine sahip olduğundan çoklu iç ilişkisi olmayan (x_1, x_2)'li model tavsiye edilir.

4.5.2. Adımsal Regresyon Yöntemleri

Tüm aday regresyon modellerini hesaplayıp değerlendirmek zor olabileceğinden, sadece az sayıdaki alt küme regresyon modellerini değerlendirmek için her seferinde bir tane regresör ekleyerek veya çıkararak yapılan çeşitli yöntemler geliştirilmiştir. Bu yöntemler adımsal türdeki yöntemlere ilişkindir. Bunlar üç ana grupta sınıflandırılabilir:

1. İleriye doğru seçim yöntemi
2. Geriye doğru ayıklama yöntemi
3 1-inci ve 2-inci yöntemin birleşimi olan adımsal regresyondur.

4.5.2.1. İleriye Doğru Seçim Yöntemi

İleriye doğru seçim yöntemi, sabit terim dışında modelde hiç regresör olmadığı varsayımıyla başlar. Her defasında modele sadece bir regresör eklenerek bir optimal alt küme bulunmaya çalışır. y yanıt değişken ile en büyük basit korelasyona sahip değişken, modele dahil edilecek ilk regresör olarak seçilir. Bu regresörün x_1 olduğunu varsayalım. Bu regresör aynı zamanda regresyonun önemliliği testi için F-istatistiğinin en büyük değerini üretecek olan regresördür. F-istatistiği önceden seçilmiş olan bir F değerini yani F_{IN}'i veya F-girileni aşarsa bu regresör modele dahil edilir. İlk regresör yani x_1'in, y üzerindeki etkisi için düzenlemeler yapıldıktan sonra y ile en yüksek korelasyona sahip olan değişken ikinci regresör olarak seçilir. Bu korelasyonlar kısmi korelasyonlar gibidir. Bunlar $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1$ oluşturulan regresyon modelinin rezidüleri ile x_1 üzerinden diğer aday regresörler

$\hat{x}_j = \hat{\alpha}_{0j} + \hat{\alpha}_{1j} x_1 \quad j = 2, 3, ..., k$ tarafından oluşturan rezidüler arasındaki basit korelasyonlardır.

İkinci adımda y ile en büyük kısmi korelasyona sahip değişkenin x_2 olduğunu varsayalım. Bu demektir ki en büyük kısmi F -istatistiği,

$$F = \frac{SS_R(x_2 / x_1)}{MS_E(x_1, x_2)} \tag{4.32}$$

dır. Eğer bu F değeri, F_{IN} değerini aşarsa x_2 modele eklenir. Genel olarak, her adımda y ile en yüksek kısmi korelasyona sahip regresör veya diğer regresörler modele dahil iken verilen en büyük kısmi F -istatistiğine sahip regresörün modele eklenebilmesi için o regresörün kısmi F -istatistiğinin değeri önceden seçilip girilen F_{IN} seviyesini aşması gerekir. Bu yöntem, ya belirli bir adımdaki kısmi F -istatistiği, F_{IN} değerini aşmadığı zaman ya da son aday regresör modele eklendiğinde sonlandırılır.

Çoklu lineer regresyon modelinde ileriye doğru seçim yöntemi kullanılarak en iyi modelin belirlenmesi örneği, Örnek 4.2'de verilmiştir.

Örnek 4.2. İleriye doğru seçim yöntemini Örnek 4.1'de verilen Hald çimento verisine uygulansın (Montgomery ve ark., 2001). Verinin SAS bilgisayar programıyla yapılan analizi sonunda elde edilen sonuçlar Şekil 4.7'de gösterilmiştir.

The REG Procedure
Dependent Variable: y

Forward Selection: Step 1 Variable x4 Entered: R-Square = 0.6473 and C(p) = 145.4119

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	1620.83817	1620.83817	18.35	0.0016
Error	10	883.29100	88.32910		
Corrected Total	11	2504.12917			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	117.36967	6.03895	33365	377.74	< .0001
x4	−0.73369	0.17128	1620.83817	18.35	0.0016

Bounds on condition number: 1, 1

Forward Selection: Step 2 Variable x1 Entered: R-Square = 0.9704 and C(p) = 6.8727

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	2430.01276	1215.00638	147.54	< .0001
Error	9	74.11641	8.23516		
Corrected Total	11	2504.12917			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	103.28516	2.32787	16212	1968.60	< .0001
x1	1.44219	0.14549	809.17459	98.26	< .0001
x4	−0.61850	0.05357	1097.63523	133.29	< .0001

Bounds on condition number: 1.0494, 4.1975

Forward Selection: Step 3 Variable x2 Entered: R-Square = 0.9835 and C(p) = 3.1716

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	2462.83753	820.94584	159.05	< .0001
Error	8	41.29164	5.16145		
Corrected Total	11	2504.12917			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	67.01422	14.50041	110.24162	21.36	0.0017
x1	1.46153	0.11544	827.35355	160.29	< .0001
x2	0.48585	0.19266	32.82477	6.36	0.0357
x4	−0.18871	0.17563	5.95877	1.15	0.3139

Bounds on condition number: 17.994, 110.63

No other variable met the 0.0500 significance level for entry into the model.

Summary of Forward Selection

Step	Variable Entered	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	x4	1	0.6473	0.6473	145.412	18.35	0.0016
2	x1	2	0.3231	0.9704	6.8727	98.26	< .0001
3	x2	3	0.0131	0.9835	3.1716	6.36	0.0357

Şekil 4.7. Hald çimento verisinin SAS bilgisayar programıyla yapılan analiz sonucu (İleriye Doğru Seçim Yöntemi) (Montgomery ve ark., 2001).

Bu programda, kullanıcı kesim değeri F_{IN}'ni birinci tip hata oranı α'yı seçerek belirlemiştir. Böylece y ile en yüksek kısmi korelasyona sahip regresörün kısmi F-istatistiği, $F_{\alpha,1,n-1}$ değerini aşarsa bu regresör modele eklenir. Bu örnekte F_{IN}'i belirlemek için $\alpha = 0.10$ kullanıldı. Bazı bilgisayar kodları ya da programları F_{IN} için seçilen sayısal değerin 2 ile 4 arasında olmasını gerektirir.

Tablo 4.4'ten, y ile en yüksek kısmi korelasyona sahip regresörün x_4 olduğu görülür. Bu değer $r_{4y} = -0.821$ dir. Hesaplanan kısmi F-istatistiği değeri

$F = 22.80$'dir. Bu değer $F_{\alpha,1,n-1}$ değerini yani $F_{0.10,1,11}$ değerini aştığı için x_4 modele eklenir. İkinci adımda: y ile en yüksek kısmi korelasyona sahip regresör x_1'dir. Yani x_4 modelde iken en büyük kısmi F -istatistiğine sahip regresör x_1'dir. Bu regresör için kısmi F -istatistiği, $F = \frac{SS_R(x_1 / x_4)}{MS_E(x_1, x_4)} = \frac{809.1048}{7.4762} = 108.22$ olarak hesaplanır. Burada $F = 108.22$ değeri, $F_{IN} = F_{0.10,1,10} = 3.29$ değerini aştığı için x_1 modele eklenir. Üçüncü adımda: x_2, y ile en yüksek kısmi korelasyona sahiptir. Bunun kısmi F -istatistiğinin değeri, $F = \frac{SS_R(x_2 / x_1, x_4)}{MS_E(x_1, x_2, x_4)} = \frac{27.7894}{5.3303} = 5.03$ olarak hesaplanır. Bu değer, $F_{IN} = F_{0.10,1,9} = 3.36$ değerini aştığı için x_2 değişkeni de modele dahil edilir ya da modele eklenir. Bu noktada geriye kalan regresör sadece x_3 olup, bunun kısmi F -istatistiğinin değeri, $F_{IN} = F_{0.10,1,9} = 3.36$ değerini aşmadığı için ileriye doğru seçim yöntemi sonlandırılır. İşlemler sonucunda oluşturulan model $\hat{y} = 71.6483 + 1.4519x_1 + 0.4161x_2 - 0.2365x_4$ şeklindedir.

4.5.2.2. Geriye Doğru Ayıklama Yöntemi

İleriye doğru seçim yönteminde, hiçbir regresör olmadan işe başlar ve uygun bir model elde edilene kadar modele değişkenler eklenir. Geriye doğru ayıklama yönteminde ise ileriye doğru seçim yöntemine zıt yönde çalışılarak iyi bir model bulmaya çalışır. Geriye doğru ayıklama yönteminde k tane regresörün tümünü içeren bir modelle işe başlanır. Daha sonra modele dahil edilecek son regresörmüş gibi her regresör için kısmi F -istatistikleri hesaplanır. Bu kısmi F -istatistiklerinden en küçüğü, önceden seçilmiş olan bir F_{OUT} ya da F -çıkarılan değeri ile karşılaştırılır. Örneğin, en küçük kısmi F değeri, F_{OUT}'dan daha küçük ise o regresör modelden çıkarılır. Bu durumda $k-1$ regresörlü bir regresyon modeli oluştu. Bu yeni model için kısmi F -istatistikleri hesaplanır ve işlem tekrar edilir. Geriye doğru ayıklama yönteminde en küçük kısmi F değeri önceden seçilmiş olan kesim değerinden ya da F_{OUT}'dan az olmadığında sona erer.

Geriye doğru ayıklama yöntemi çok iyi bir değişken seçme yöntemidir. Bu yöntem, bütün aday regresörleri içermenin etkilerini görmek için araştırmacılar tarafından tercih edilir.

Çoklu lineer regresyon modelinde geriye doğru ayıklama yöntemi kullanılarak en iyi modelin belirlenmesi örneği, Örnek 4.3'te verilmiştir.

Örnek 4.3. Geriye doğru ayıklama yöntemi, Örnek 4.1'de verilen Hald çimento verisi kullanılarak açıklanacaktır. Verinin SAS bilgisayar programıyla yapılan analizi sonunda elde edilen sonuçlar Şekil 4.9'da gösterilmiştir.

The REG Procedure
Dependent Variable: y

Backward Elimination: Step 0 All Variables Entered: R-Square = 0.9839 and C(p) = 5.0000

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	2463.82566	615.95642	106.98	< .0001
Error	7	40.30350	5.75764		
Corrected Total	11	2504.12917			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	37.87077	71.99629	1.59306	0.28	0.6151
x1	1.77030	0.75524	31.63523	5.49	0.0515
x2	0.78462	0.74934	6.31260	1.10	0.3299
x3	0.31636	0.76365	0.98814	0.17	0.6911
x4	0.10303	0.72823	0.11525	0.02	0.8915

Bounds on condition number: 277.33, 2431.4

Backward Elimination: Step 1 Variable x4 Removed: R-Square = 0.9839 and C(p) = 3.0200

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	2463.71042	821.23681	162.55	< .0001
Error	8	40.41875	5.05234		
Corrected Total	11	2504.12917			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	48.04042	3.80647	804.75231	159.28	< .0001
x1	1.66782	0.20019	350.68522	69.41	< .0001
x2	0.67884	0.04653	1075.38866	212.85	< .0001
x3	0.21188	0.18221	6.83166	1.35	0.2784

Bounds on condition number: 3.2382, 22.304

Backward Elimination: Step 2 Variable x3 Removed: R-Square = 0.9811 and C(p) = 2.2066

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	2456.87876	1228.43938	233.99	< .0001
Error	9	47.25041	5.25005		
Corrected Total	11	2504.12917			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	51.62412	2.27741	2697.66345	513.84	< .0001
x1	1.47601	0.11563	855.48424	162.95	< .0001
x2	0.68673	0.04692	1124.50123	214.19	< .0001

Bounds on condition number: 1.0397, 4.1587

All variables left in the model are significant at the 0.0500 level.

Summary of Backward Elimination

Step	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	x4	3	0.0000	0.9839	3.0200	0.02	0.8915
2	x3	2	0.0027	0.9811	2.2066	1.35	0.2784

Şekil 4.8. Hald çimento verisinin SAS bilgisayar programıyla yapılan analiz sonucu (Geriye Doğru Ayıklama Yöntemi) (Montgomery ve ark., 2001).

Bu örnekte $\alpha = 0.10$ alınarak kesim değeri F_{OUT} ya da F-çıkarılan belirlenmiştir. Böylece, $F_{0.10,1,n-p}$'den daha az kısmi F-istatistiğine sahip regresör modelden atılacak ya da modelden çıkarılacaktır. Adım 0: oluşturulan tam modelin sonuçlarını göstermektedir. En küçük kısmi F değeri $F = 0.02$ olup bu değer x_3'e aittir. Böylece $F = 0.02 < F_{OUT} = F_{0.10,1,8} = 3.46$ olduğu için x_3 modelden çıkarılır. Şekil 4.9'da Adım 1'de (x_1, x_2, x_4) üç-değişkenli modelin oluşumunun sonuçları görülmektedir. Bu modeldeki en küçük kısmi F değeri $F = 1.86$ olup x_4'e aittir. $F = 1.86 < F_{OUT} = F_{0.10,1,9} = 3.36$ olduğu için x_4 modelden çıkarılır. Adım 2'de (x_1, x_2) iki değişkenli modelin oluşumunun sonuçları görülmektedir. Bu modeldeki en küçük kısmi F-istatistiği $F = 146.52$ olup x_1'e ait değerdir. Bu değer $F_{OUT} = F_{0.10,1,10} = 3.29$ değerini aştığı için daha fazla regresör modelden çıkarılamaz. Bu yüzden geriye doğru ayıklama yöntemi son modeli oluşturarak sonlandırılır. Oluşturulan model, $\hat{y} = 52.5773 + 1.4683x_1 + 0.6623x_2$ şeklindedir. Dikkat edilirse, bu model ileriye doğru seçim yöntemiyle bulunan modelden farklıdır. Ayrıca bu model, bütün aday regresyon modelleri tarafından en iyi olarak tanıtılan modeldir.

4.5.2.3. Adımsal Regresyon Yöntemi

Yukarıda tanımlan iki yöntem de en iyi regresyon modelinin belirlenmesi için değişik yaklaşımlar içerirler. Bunlara ek olarak Efroymson (1960) da değişken seçimi için adımsal regresyon yöntemini önermiştir. Bu yöntemde ileriye doğru seçim yönteminin bir düzenlemesinden oluşur. Bu yöntemde modele daha önce eklenen regresörler kısmi F-istatistikleriyle yeniden değerlendirilir. Modele daha önceden eklenen bir regresör daha sonraki adımlarda modelden çıkarılabilir. Bir değişken için kısmi F-istatistiğinin değeri, F_{OUT} ya da F-çıkarılan'ın değerinden daha az ise o değişken modelden atılır. Adımsal regresyon iki kesim değerine gereksinim duyar. Bunlar F_{IN} ya da F-girilen ve F_{OUT} ya da F-çıkarılan

değerleridir. Çoklu lineer regresyon modelinde adımsal regresyon yöntemi kullanılarak en iyi modelin belirlenmesi örneği, Örnek 4.4'de verilmiştir.

Örnek 4.4. Adımsal regresyon yöntemi, Örnek 4.1'de verilen Hald çimento verisi kullanılarak açıklanacaktır. Verinin SAS bilgisayar programıyla yapılan analizi sonunda elde edilen sonuçlar Şekil 4.9'da gösterilmiştir.

The REG Procedure
Dependent Variable: y

Stepwise Selection: Step 1 Variable x4 Entered: R-Square = 0.6473 and C(p) = 145.4119

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	1620.83817	1620.83817	18.35	0.0016
Error	10	883.29100	88.32910		
Corrected Total	11	2504.12917			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	117.36967	6.05895	33365	377.74	< .0001
x4	−0.73369	0.17128	1620.83817	18.35	0.0016

Bounds on condition number: 1, 1

Stepwise selection: Step 2 Variable x1 Entered: R-Square = 0.9704 and C(p) = 6.8727

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	2430.0.276	1215.00638	147.54	< .0001
Error	9	74.1:641	8.23516		
Corrected Total	11	2504.12917			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	103.28516	2.32787	16212	1968.60	< .0001
x1	1.44219	0.14549	809.17459	98.26	< .0001
x4	−0.61850	0.05357	1097.63523	133.29	< .0001

Bounds on condition number: 1.0494, 4.1975

Stepwise Selection: Step 3 Variable x2 Entered: R-Square = 0.9835 and C(p) = 3.1716

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	2462.83753	820.94584	159.05	< .0001
Error	8	41.29164	5.16145		
Corrected Total	11	2504.12917			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	67.01422	14.50041	110.24162	21.36	0.0017
x1	1.46153	0.11544	827.35355	160.29	< .0001
x2	0.48585	0.19266	32.82477	6.36	0.0357
x4	−0.18871	0.17563	5.95877	1.15	0.3139

Bounds on condition number: 17.994, 110.63

Stepwise Selection: Step 4 Variable x4 Removed: R-Square = 0.9811 and C(p) = 2.2066

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	2456.87876	1228.43938	233.99	< .0001
Error	9	47.25041	5.25005		
Corrected Total	11	2504.12917			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	51.62412	2.27741	2697.66345	513.84	< .0001
x1	1.47601	0.11563	855.48424	162.95	< .0001
x2	0.68673	0.04692	1124.50123	214.19	< .0001

Bounds on condition number: 1.0397, 4.1587

All variables left in the model are significant at the 0.0500 level.
No other variable met the 0.0500 significance level for entry into the model.

Summary of Stepwise Selection

Step	Variable Entered	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	x4		1	0.6473	0.6473	145.412	18.35	0.0016
2	x1		2	0.3231	0.9704	6.8727	98.26	< .0001
3	x2		3	0.0131	0.9835	3.1716	6.36	0.0357
4		x4	2	0.0024	0.9811	2.2066	1.15	0.3139

Şekil 4.9. Hald çimento verisinin SAS bilgisayar programıyla yapılan analiz sonucu (Adımsal Regresyon Yöntemi) (Montgomery ve ark., 2001).

Bu örnekte bir regresörü modele girmek ve modelden çıkarmak için $\alpha = 0.05$ alınmıştır. Adım 1: Adımsal regresyon yöntemi, modelde hiç regresör yokken başlar. Yani sabit modelden başlar. Modele önce x_4'ü eklemeye çalışır. Kısmi F-istatistiğinin değeri, $F_{IN} = F_{0.05,1,11} = 4.84$ değerini aştığı için x_4 modele eklenir. Adım 2: x_1 modele eklenir. x_4 için kısmi F-istatistiğinin değeri, $F_{OUT} = F_{0.05,1,10} = 4.96$ değerinden daha az ise x_4 modelden çıkarılır ya da silinir. Kısmi F-istatistiğinin değeri, $F = 159.30$ olduğundan x_4 modelde bırakılır. Adım 3: Adımsal regresyon yöntemi modele x_2'yi ekler. Bu durumda x_1 ve x_4 için kısmi F-istatistiği $F_{OUT} = F_{0.05,1,9} = 5.12$ değeriyle karşılaştırılır. x_4 için kısmi F-istatistiğinin değeri $F = 1.86$ ve $F_{OUT} = 3.36$ olduğundan x_4 modelden çıkarılabilir ya da silinebilir. Adım 4: modelden x_4'ün çıkarılmasıyla ilgili sonuçları göstermektedir. Bu noktada geriye kalan tek regresör x_3'tür. x_3'ün kısmi F-istatistiğinin değeri, F_{IN} değerini aşmadığından x_3 modele eklenemez. Bu durumda işlemler sonlandırılır. Oluşturulan model, $\hat{y} = 52.5773 + 1.4683x_1 + 0.6623x_2$ şeklindedir. Dikkat edilirse, bu model geriye doğru ayıklama yöntemiyle bulunan modelle aynıdır.

5. ÇOKLU LİNEER REGRESYON MODELİ İÇİN BİR GENETİK ALGORİTMA

Çoklu lineer regresyon modelindeki açıklayıcı değişken sayısı k'nın fazla olması durumunda örneğin $k = 10$ olsun. Aday model sayısı $2^k = 2^{10} = 1024$ olacaktır (Gunst ve Mason, 1980). Bu durumda ne geleneksel yöntemler, ne de adımsal yöntemler kullanılamamaktadır (Wasserman ve Sudjianto 1994; Bozdogan 2003). Çoklu lineer regresyon modelindeki açıklayıcı değişken sayısının fazla olması durumunda çoklu lineer regresyon modelinin oluşturulmasında Genetik Algoritma kullanılabilir (Wallet ve ark., 1996).

5.1. Genetik Algoritmalar Hakkında Genel Bilgiler

Genetik algoritmalar (GA); evrim, gelişim ya da değişim hesaplamalarının bir parçasıdır. Genetik algoritmalar, Darwin'in evrim teorisinden esinlenerek oluşturulur. Genetik algoritmalar, yapay zekanın çok hızlı gelişen bir alanıdır (Goldberg, 1989). Genetik algoritmalar, genellikle bir problemin çözümünü kolaylaştırmak için kullanılır. Bir problemin çözümünde genetik algoritmaların kullanılması ilk defa John Holland tarafından ortaya atılmıştır. Daha sonra kendisi, öğrencileri ve meslektaşları tarafından geliştirilmiştir. John Holland, bu çalışmalar sonucunda 1975 yılında "Doğal ve Yapay Sistemlerde Adaptasyon / Uyum" (Adaption in Natural and Artificial Systems) adlı kitabı yazmıştır. 1992 yılında John Koza, belirli işleri yapabilmek veya yerine getirmek amacıyla, program geliştirmek için genetik algoritmayı kullandı. Bu yöntemini de "Genetik Programlama (GP)" olarak adlandırdı. Genetik programlamada LISP (LISt Processing) programlama dili kullanıldı. Bunun nedeni LISP programlama dilinin, genetik algoritmalarda da kullanılan soyağacı (parse tree) yapısını daha kolay ve etkin işleyebilmesidir. Her problemin çözümünde problemin yapısına göre bir genetik algoritma oluşturulabilir (Michalewicz, 1992). Genetik algoritma, bir problemin çözümü için bir yöntem değildir. Bununla birlikte genetik algoritma bir problemin çözümünü elde etmek için izlenen yol olarak ifade edilebilir (Bauer, 1994).

Genetik algoritmalar, popülasyon (erişkinlerin ya da yetişkinlerin bir kitlesi) denilen ve kromozomlar ile gösterilen çözümlerin bir kümesi ile başlatılır. Bir popülasyondan çözümler alınır. Bu çözümler daha sonra yeni bir popülasyon oluşturmak için kullanılır. Bu işlem yeni popülasyonun eski popülasyondan daha iyi olacağı varsayımından hareketle yapılır. Yeni çözümleri (nesilleri) oluşturmak için seçilen çözümler uygunluk ya da uyumluluk değerlerine göre seçilir (Goldberg, 1989).

Genetik algoritmada kullanılan işlemler yeni bir populasyonu oluşturmak için kullanılır. Bu işlemler tamamen uygunluk fonksiyonuna bağlı olarak gerçekleşir.

Genetik algoritmayla ilgili bazı kavramlar aşağıda açıklanmıştır.

Üreme: Üreme işlemi belli bir seçme kriterine göre bireylerin seçilip yeni kuşağın oluşturulması işlemidir. Seçme kriterleri uyumluluğu esas alarak birbiriyle uyumlu olan bireyleri seçer. Daha sonra çaprazlama ve mutasyon uygulanacak olan bireylerden daha uyumlu yeni bireylerin ortaya çıkması olasıdır. Bireylerin tamamı uyumluluğa göre seçilebilir veya bir kısmı rasgele seçilerek yeni kuşağa aktarılabilir.

Çaprazlama: Kromozomların nasıl temsil edileceğine karar verildikten sonra çaprazlama yapılabilir. Çaprazlamada ebeveynlerden bazı genleri alınır ve yeni bireyler oluşturulur.

Kromozom 1	11011 | 00100110110
Kromozom 2	11011 | 11000011110
Birey 1	11011 | 11000011110
Birey 2	11011 | 00100110110

Çaprazlama yapılacak konum rasgele seçilir (|). Oluşan yeni birey ebeveynlerin bazı özelliklerini almış ve bir bakıma ikisinin kopyası olmuştur. Çaprazlama işlemi başka şekillerde de yapılabilir. Örneğin birden fazla çaprazlama noktası seçilebilir. Daha iyi performans almak amacıyla değişik çaprazlamalar kullanılabilir.

Mutasyon: Çaprazlama gerçekleştikten sonra mutasyon gerçekleştirilir. Mutasyon oluşan yeni çözümlerin önceki çözümü kopyalamasını önlemek ve sonuca daha hızlı ulaşmak amacıyla yapılır. Mutasyon oluşan yeni bireyin (eğer ikili düzende ifade edilmişse) bir bitini rasgele değiştirir.

Orjinal Birey 1	1101111000011110
Orjinal Birey 2	1101100100110110
Değişmiş Birey 1	1100111000011110
Değişmiş Birey 2	1101101100110110

Elitizm: Üreme, çaprazlama ve mutasyon işlemleri sonrasında kuşakta bulunan en iyi uyumluluğa sahip birey sonraki kuşağa aktarılamayabilir. Bunu önlemek için bu işlemlerden sonra oluşan yeni kuşağa bir önceki kuşağın en iyi (elit) bireyi, yeni kuşaktaki herhangi bir birey ile değiştirilir. Buna elitizm adı verilir.

Genetik algoritmanın çaprazlama olasılığı ve mutasyon olasılığı olmak üzere iki parametresi vardır:

Çaprazlama olasılığı çaprazlamanın hangi sıklıkta yapılacağını belirtir. Eğer hiç çaprazlama yapılmazsa (bu durumda çaprazlama olasılığı %0 dır) yeni bireyler eski bireylerin aynısı olur. Bu yeni kuşağın eskisiyle aynı olacağı anlamına gelmez. Eğer bu oran %100 olursa yeni bireyler tamamıyla çaprazlama ile elde edilir. Çaprazlama eski bireylerden iyi taraflar alınarak elde edilen yeni bireylerin daha iyi olması umuduyla yapılır.

Mutasyon olasılığı ise mutasyonun hangi sıklıkta yapılacağını belirtir. Mutasyon olmaz ise yeni birey çaprazlama veya kopyalama sonrasında olduğu gibi kalır. Eğer mutasyon olur ise yeni bireyin bir kısmı değiştirilmiş olur. Eğer bu oran %100 olursa kuşak içindeki bireyler tamamen değişir, %0 olursa hiç değişmeden kalır.

Genetik algoritmada kullanılan başka parametreler de vardır. Bunların en önemlilerinden birisi de populasyon büyüklüğüdür. Bu parametre populasyon içinde (yalnızca bir kuşakta) kaç adet kromozom yani birey olduğunu söyler. Eğer kromozom sayısı az olursa GA çözüm aranan uzayın ancak bir kısmını gezebilir ve çaprazlama için fazla bir seçeneği yoktur. Kromozom sayısı çok fazla olursa GA çok yavaş çalışır. Araştırmalar belli bir noktadan sonra populasyon sayısını artırmanın bir yararı olmadığını göstermiştir.

Yeni bireyler uyumluluğa göre veya rasgele olarak seçilebilir. Yeni bireylerin tamamen rasgele seçilme durumunda yakınsama zorlaşabilir. Bu durumda çözüme ulaşmamız zorlaşabilir. Bu sorunların üstesinden gelmek için belli bir oranda uyumluluk seçimi belli bir oranda da rastgele seçim yapılabilir. Bu orana Kuşak Farkı

(Generation Gap) denir. Kuşak farkı %100 olduğunda yeni bireylerin tamamı uyumluluğa göre seçilir.

Genel olarak bir genetik algoritmanın adımları aşağıda açıklanmıştır (Bozdogan, 2003).

1. Başlangıç: n adet kromozom içeren populasyonun oluşturulması (problemin uygun bir çözümü).

2. Uyumluluk: her x kromozomu için uyumluluğun $f(x)$ değerlendirilmesi.

3. Yeni populasyon: Yeni populasyon oluşuncaya kadar aşağıdaki adımlar tekrarlanır.

3.1. Seçim: İki ebeveyn kromozomun uyumluluğuna göre seçimi (daha iyi uyum seçilme şansını artırır.)

3.2. Çaprazlama: Yeni bir fert oluşturmak için ebeveynlerin bir çaprazlama olasılığına göre çaprazlanması. Eğer çaprazlama yapılmazsa yeni fert anne veya babanın kopyası olacaktır.

3.3. Mutasyon: Yeni ferdin mutasyon olasılığına göre kromozom içindeki konumu (lokus) değiştirilir.

3.4. Ekleme: Yeni bireyin yeni populasyona eklenmesi.

4. Değiştirme: Algoritmanın yeniden çalıştırılmasında oluşan yeni populasyonun kullanılması.

5. Test: Eğer sonuç tatmin ediyorsa algoritmanın sona erdirilmesi ve son populasyonun çözüm olarak sunulması.

6. Döngü: 2. adıma geri dönülmesi.

Bu adımlar, aşağıda kısaca açıklanmıştır.

Adım-1. Bu adıma populasyonda bulunacak birey sayısını belirleyerek başlanır. Kullanılacak sayı için bir standart yoktur. Genel olarak önerilen 100-300 aralığında bir büyüklüktür. Büyüklük seçiminde yapılan işlemlerin karmaşıklığı ve aramanın derinliği önemlidir. Populasyon bu işlemden sonra rasgele oluşturulur.

Adım-2. Kromozomların ne kadar iyi olduğunu bulan fonksiyona uygunluk fonksiyonu denir. Bu fonksiyon işletilerek kromozomların uygunluklarının bulunmasına ise evaluation adı verilir. Bu fonksiyon genetik algoritmanın en önemli kısmını oluşturmaktadır. Genetik algoritmada probleme özel çalışan tek kısım bu fonksiyondur. Uygunluk fonksiyonu, kromozomları problemin

parametreleri haline getirir. Onların bir bakıma şifresini çözmektedir. Bu işleme de decoding denir. Sonra bu parametrelere göre hesaplamayı yaparak kromozomların uygunluğunu bulur. Bu fonksiyon ne kadar hassas ve verimli ise Genetik algoritma da o kadar başarılı sonuçlar verir.

Adım-3. Kromozomların eşlenmesi kromozomların uygunluk değerlerine göre yapılır. Bu seçimi yapmak için rulet tekerleği seçimi (roulette wheel selection), turnuva seçimi (Tournament Selection) gibi seçme yöntemleri vardır. Örnek olarak bu çalışmada kullanılan rulet tekerleği seçimi aşağıda açıklanmıştır.

1- Tüm bireylerin uygunluk değerleri bir tabloya yazılır. Genetik algoritmada başarı olasılığı için gereklidir.

2- Bu değerler toplanır.

3- Tüm bireylerin uygunluk değerleri toplama bölünerek [0,1] aralığında sayılar elde edilir. Bu sayılar bireylerin seçilme olasılıklarıdır. Sayıların hepsi bir tabloda tutulur.

4- Seçilme olasılıklarının tuttulduğu tablodaki sayılar birbirine eklenerek rastgele bir sayıya kadar ilerlenir. Bu sayıya ulaşıldığında ya da geçildiğinde son eklenen sayının ait olduğu çözüm seçilmiş olur. Bu yönteme rulet tekerleği seçimi ismi, bir daireyi, çözümlerin uygunluklarına göre dilimleyip çevirdiğimizde olacakların benzeşimi olduğu için verilmiştir. Rulet tekerleği seçimi çözümlerin uygunluk değerlerinin negatif olmamasını gerektirir. Çünkü olasılıklar negatif olursa bu çözümlerin seçilme şansı yoktur. Çoğunluğunun uygunluk değeri negatif olan bir populasyonda yeni nesiller belli noktalara takılıp kalabilir. Gen takası (crossover) genetik algoritmanın en önemli aracı olarak kabul edilir. Basitçe olay iki ebeveyn kromozomun arasında belirlenen parçaların takasıdır. Genetik algoritmalar bu olayın benzeşimini temelde: Tek noktalı (Single (one) point crossover) ve Çok noktalı (Multi point crossover) şeklindedir. Gen takası adı verilen iki yolla yapar. Genetik algoritmalarda ikilik dizi (binary string) çok kullanılır. Doğadaki genlerin benzeşimi birer ‘bit’ olarak gösterilir. İkilik dizilerde bir gen takası aşağıdaki gibi gerçekleşebilir.

	Önce	Sonra
Fert A	010 111	010 001
Fert B	101 001	101 111

Crossover noktası

Gen takası populasyonda çeşitliliği sağlar. İyi özelliklerin bir araya gelmesini kolaylaştırarak en iyiye yaklaşmayı sağlar. Değiştirme kromozomun bir parçasının dışarıdan değiştirilmesi şeklinde tanımlanır. Değiştirme görünüşte genetik algoritmanın dayanak noktasıdır, ancak etkisi bir çözüm üzerindedir. Bu da yalnız başına başarılı olmasını zorlaştırır. İkilik dizilerde değiştirme rasgele bir bit'in değiştirilmesiyle sağlanabilir. Çok düşük bir değiştirme olasılığı populasyonda bazı özelliklerin kaybolmasına neden olabilir. Örnek bir problemde genetik algoritmayı uygulayarak bu adıma geldik. Populasyonda cros-over veya mutasyon işlemini uygularken düşük seçilme olasılıklı bir birey alınırsa veya yüksek seçilme olasılıklı bir birey elenirse bu problemde en iyi sonuçların bulunmasına engel olur. Ancak yüksek bir değiştirme olasılığı da eldeki çözümleri bozarak sonuca ulaşmayı zorlaştırır. Gen takası ve değiştirmenin olasılıkları için kesin bir sayı yoktur. Değiştirme (mutasyon) olasılığı 0.01-0.001, gen takası (cross-over) olasılığı 0.5-1.0 aralığında tavsiye edilir.

Adım-4. Eski kromozomlar çıkartılarak sabit büyüklükte bir populasyon sağlanır.

Adım-5. Tüm kromozomlar yeniden hesaplanarak yeni populasyonun başarısı bulunur.

Adım-6. Genetik algoritma defalarca çalıştırılarak çok sayıda populasyon oluşturulup hesaplanır.

Adım-7. Populasyonların hesaplanması sırasında en iyi bireyler saklandığı için o ana kadar bulunmuş en iyi çözüm, çözümdür.

5.2. Çoklu Lineer Regresyon Modelinde En İyi Modelin Oluşturulmasında Genetik Algoritmanın Kullanılması ve Bilgi Karmaşıklık Kriteri

Çoklu lineer regresyon analizinde, lojistik regresyon analizinde ya da sıralı lojistik regresyon analizinde olduğu gibi regresyon tipi modellerde model oluşturma ve hesaplama ve açıklayıcı değişkenlerin uygun alt küme seçimi veri madenciliğinde

merkezi bir problemdir (Lanning ve Bozdogan, 2003). Çoğu kez açıklayıcı değişkenlerin bir alt kümesi verildiğinde bir miktarsal,ikili veya sıralı düzeyde yanıt değişken çalışır.Böyle durumlarda açıklayıcı değişkenlerin hangilerinin yanıt değişkendeki toplam değişimi açıklamak için ve çok fazla sayıdaki regresyon katsayılarını yorumlamak için kullanılacağını belirlemek önemlidir.

Çok fazla sayıda açıklayıcı değişken olması durumunda ve araştırmacının bu değişkenler arasındaki tam ilişki hakkında bilgi sahibi olmaması durumunda en iyi regresyon modelini seçme problemi açık/aşikar olmayan bir alıştırmadır.

Bir çok durumda mümkün / olası modellerin sayısı oldukça fazladır. (Örneğin, 20'den fazla açıklayıcı değişken olması durumunda mümkün/olası modellerin sayısı milyonu geçmektedir. Bu durumda zaman ve maliyet bakımından alt kümelerin kombinasyonlarının mümkün / olası modellerin hesaplanması gerçekçi değildir.

Bu nedenle,sayısal optimizasyon tekniklerine ve model seçimi stratejilerine gereksinim vardır.Sayısal teknikler kullanılarak alt küme seçim probleminin çözümü genel olarak iki bileşeni gerektirir. Bunlar:

1. Çözüm uzayının etkin olarak araştırılması için bir algoritma.
2. En iyi modelin seçimine kılavuzluk edecek yarışan modellerin karşılaştırılması için bir kriter veya ölçüm.

İstatistiksel analiz için bir çok istatistiksel paket program,en iyi alt küme modelini seçmek için geriye doğru ayıklama ve ileriye doğru seçim gibi adımsal seçim yöntemlerini sağlar / içerir. Bununla birlikte,regresyon analizinde geriye doğru ayıklama ve ileriye doğru seçim adımsal yöntemlerinin her ikisi de k değişkeninin bir kümesinden açıklayıcı (predictor) değişkenlerin en iyi alt kümesini daima bulmaz. Geriye doğru ayıklama ve ileriye doğru seçim adımsal yöntemleri hakkında en önemli kritikler(eleştiriler):

1. Algoritmaya hangi değişkenlerin modele dahil edileceği veya modelden çıkarılacağı sıralaması için teorik düzenlemelerin bulunmaması / olmaması (Boyce ve ark., 1974, s.19; Wilkinson, 1989, s.177-178).
2. Analizde madole dahil edilecek ve modelden çıkarılacak değişkenlerin belirtilen bir prior olasılıklarının seçimiyle ilgili herhangi bir teorik düzenlemelerin bulunmaması/ olmamasıdır.

3. Diğer bir kritik (eleştiri), adımsal arama arasıra (nadiren) genel en iyi modeli bulur veya hatta özel bir hacimdeki en iyi alt kümeyi bulur (Mantel 1970, Hocking 1976, Hocking 1983, Moses 1986).
4. Son olarak ve en önemlisi yerel araştırmaya başvurulduğundan adımsal seçim geniş çözüm uzayının küçük bir alanından oldukça sınırlı bir örneklemi sağlar. Adımsal seçim en iyisinde sadece yeterli modeli oluşturur (Sokal ve Rohlf, 1981, s.668).

Regresyon analizinde mevcut problemlerin yukarıdaki açıklamalarına dayalı olarak bu çalışmanın amacı çoklu regresyon modellerinde alt küme seçim için bilgi tabanlı model seçim kriterine ve genetik algoritmaya (GA) dayalı hesaplama bakımından uygulanabilir akıllı veri madenciliğini tanıtmak ve geliştirmektir.Buradaki yaklaşım aynı zamanda üç yönlü hibrid olarak lojistik regresyon ve sıralı lojistik regresyon modellerine genişletilebilir. Sıralı lojistik regresyon modellerinde en iyi açıklayıcı değişkenlerin alt küme seçimi için Lanning ve Bozdoğan (2003) incelenebilir.

Bir genetik algoritma geniş sayıda mümkün/olası çözümlerin bulunduğu problemin çözümüne uygulanabilen ve biyolojik değişim/dönüşüm ve doğal seçime dayalı stokastik (rastgele) arama algoritmasıdır. Genetik algoritmalar mühendislik,ekonomi,oyun teorisi (Holland, 1992), hesaplama bilimleri (Forrest, 1993), pazarlama (Bauer, 1994) ve biyoloji (Sumida ve ark., 1990) gibi geniş bir alanda kullanılabilir. Geleneksel optimizasyon yaklaşımından farklı olarak genetik algoritma amaç fonksiyonunun gradyanlarını hesaplamaya gereksinim duymaz ve bir yerel optimuma sınırlanmaz (Goldberg, 1989). Bir genetik algoritma bir ikili string halindeki kodların bir dizisi olarak bilgilendirir. İkili stringler verilen probleme farklı çözümleri gösterir. Bu stringler bir kromozom üzerindeki genler tarafından kodlanan genetik bilgiye analog modellerdir. Bir string problemi çözmek için özel yeteneği için uyum / uygunluk değerlerine göre hesaplanabilir. Uyum değerleri tabanında stringler, her bir çalıştırmadan sonra ve analizde problemin çözümü için kullanılır ya da atılır. Bir çok çalıştırmadan sonra en iyi çözüm belirlenir / tespit edilir. Herhangi bir genetik algoritmadaki zorluk, her bir çözümü hesaplamak için temel olarak uygun bir uyum fonksiyonunun seçimidir.

Çoklu regresyon analizine göre uyum değeri en iyi alt kümenin araştırılmasında alt küme modellerin karşılaştırılması için bir alt küme seçim kriteridir. Bu bilgisel model seçimi kriteri kullanılarak kolaylıkla belirlenebilir.

5.3. ICOMP Bilgi Karmaşıklık Kriteri Kullanılarak Genetik Algoritmanın Uygulanması

Genel olarak istatistiksel modelleme ve model hesaplama problemlerinde model karmaşıklığı kavramı önemli bir rol oynar. Karmaşıklık bağlantı yapıları olarak tasarımlar ve model bileşenlerinin etkileşimlerini içerir. "Genel" model karmaşıklığının bir ölçümü olmaksızın model davranışını tahmin etmek ve modelin kalitesini değerlendirmek zordur. Bu detaylı istatistiksel analize ve verilen sonlu bir örneklem için yarışan modellerin tümü arasında en iyi modeli seçmek için hesaplamalara gereksinim duyar. Bu bölümde istatistiksel sonuç çıkarmaya uygun yeni yaklaşımların oluşumuna yardımcı olmak için istatistiksel modellemede "genel" model karmaşıklığının bir ölçümünün bilgi-teorik düşünceleri geliştirilecek ve gösterilecektir.

Yakın zamanda Akaike'nin (1973) orijinal AIC'kine dayalı bir çok model-seçimi prosedürü önerilmiştir (Sclove, 1987). Model seçiminde AIC'in bu formu,

$$AIC(k) = -2logL\left(\hat{\theta}_k\right) + 2m(k) \tag{5.1}$$

biçimindedir. Burada $L\left(\hat{\theta}_k\right)$, maksimum yapılan likelihood fonksiyon; $\hat{\theta}_k$, M_k modeli altında θ_k parametre vektörünün maksimum likelihood tahmin ve, $m(k)$, M_k model olduğunda bağımsız parametrelerin sayısıdır.

AIC'de uzlaşma,maksimum yapılan log likelihood yani $-2logL\left(\hat{\theta}_k\right)$ (uyumun eksikliği bileşeni) ve $m(k)$ arasında yer alır. Model içinde tahmin edilen serbest parametrelerin sayısı (penaltı bileşeni) karmaşıklığın bir ölçümüdür aynı zamanda maksimum likelihood tahmin edici kullanıldığında uyumun eksikliğindeki yanlılıktır.

AIC'in kullanılmasında Akaike (1987)'ye göre parametre tahminlerinin doğruluğu bir evresel kriter ile ölçülür. Bu,

Doğruluk Ölçümü = E [Oluşturulan Modelin Likelihoodu] (5.2)

biçimindedir. Burada E beklenen değeri göstermektedir. Çünkü AIC, eksi iki çarpı loglikelihoodun beklenen değerinin bir yansız tahminidir.

AIC'dekine benzer şekildeki incelemelerden hareketle işlemler yapılmıştır.Bununla birlikte yeni prosedür van Emden (1971)'in bilgi-tabanlı kovaryans karmaşıklık indeksinin bir genelleştirilmesi yoluyla bir elemanın veya rasgele vektörlerin yapısal karmaşıklığı üzerinde ICOMP'a dayandırılmıştır.

Bir genel çok değişkenli lineer ya da lineer olmayan model,

İstatistiksel model = Sinyal + Gürültü (5.3)

bağıntısıyla tanımlanır.

ICOMP,

Kayıp=Uyumun eksikliği+Tutumluluğun eksikliği+Karmaşıklığın ön birleşimi (5.4)

şeklindeki bir kayıp fonksiyonu tahmin etmek için bilgi teorisinin ekleme özellikleri kullanılarak değişik yollarla tasarlanır. Buradaki yapılan çalışmada işlemler Rissanen (1976)'daki tahminde ve model tanımlama problemlerinde son tahmin kriterine (final estimation criterion (FEC)), aynı zamanda Akaike'nin (1973) AIC'kine ve Bozdogan (1987)'deki AIC'nin analitik genişlemelerine benzer şekilde incelenmiştir.

ICOMP'un oluşturulması ve geliştirilmesi orijinal olarak Van Emden (1971) tarafından tanımlanan kovaryans karmaşıklık indeksinin bir genelleştirmesine dayalıdır. Direk olarak serbest parametrelerin saysını cezalandırma yerine ICOMP modelin kovaryans karmaşıklığını cezalandırır. ICOMP,

$$ICOMP = -2logL\left(\hat{\theta}\right) + 2C\left(\hat{\Sigma}_{Model}\right) \tag{5.5}$$

ile tanımlanır (Bozdogan, 2003). Burada, $L\left(\hat{\theta}_k\right)$ maksimum yapılan likelihood fonksiyonudur. $\hat{\theta}_k$, M_k modeli altında θ_k parametre vektörünün maksimum likelihood tahminidir. C, bir gerçel-değerli karmaşıklık ölçümüdür. $\widehat{Cov}\left(\hat{\theta}\right) = \hat{\Sigma}_{Model}$ modelin parametre vektörünün tahmin edilen kovaryans matrisini göstermektedir.

ICOMP'un değişik formları ve düzenlemeleri olduğundan (5.5)'deki eşitliğe dayalı olarak bu çalışmada ICOMP'un en genel formu gösterilecektir. ICOMP'un en genel formu ICOMP(IFIM) dir. ICOMP(IFIM), maksimum likelihood tahminlerin iyi-bilinen asimptotik optimallik özelliğini açıklar ve bir modelin inverse-fisher bilgi matrisinin (IFIM) bilgi tabanlı karmaşıklığını kullanır. Bu, Cramér-Rao alt sınır matrisi olarak bilinir (Cramér 1946, Rao 1945, Rao 1947, Rao 1948).

ICOMP(IFIM)'i elde etmeden önce karmaşıklık kavramını anlamaya çalışmak için bazı ön bilgileri verelim ve daha sonra bir sistemin karmaşıklığının tanımını verelim.

5.3.1. Karmaşıklık Kriteri ve Bir Sistemin Karmaşıklığı

Karmaşıklık, istatistiksel modellerin bir genel özelliğidir ve modellerin olasılık tanımlarından / özelliklerinden, yapısından veya özel içeriğinden çoğunlukla bağımsızdır. Literatürde, karmaşıklık kavramı bir çok değişik içerikte kullanılmıştır. Van Emden (1971)'e göre tasarım anlaşılması zor olduğundan genel olarak istatistikte karmaşıklığın tek bir tanımı yoktur. Karmaşıklığın bir çok yönü vardır ve "Kolmogorov karmaşıklığı" (Cover ve ark., 1989), "Shannon Karmaşıklığı" (Rissanen 1987, 1989) gibi bir çok adlar altında tanımlanır. Bilgi teorik kodlama teorisinde Rissanen (1986, 1987, 1989), modellerin sınıfları tarafından ortaya çıkarılabilen veriler için en kısa kod uzunluğu cinsinden karmaşıklığı tanımlayan Kolmogorov (1983)'dekine benzer şekilde karmaşıklığı tanımlamıştır ve onu Stokastik Karmaşıklık (SC) olarak adlandırmıştır. Wallace ve Freemen (1987), Wallace ve Dowe (1993) ve

Baxter (1996) karmaşıklığı Minimum Mesaj Uzunluğu (Minimum Message Length-MML) cinsinden tanımlamıştır. Minimum mesaj uzunluğu, veriyi kapsayan bir mesajı sıkıştırma yeteneğine göre modellerin hesaplanmasına dayalıdır.

Karmaşıklığın anlaşılması ve verinin ışığında belirsizliğini çalışmak için (tümevarımsal) sonuç çıkarmak genel model oluşturma teorisinde çok gereklidir. İstatistiksel modeller ve yöntemler tam olarak tümdengelimli değildir. Çünkü insanlar çoğu zaman belirsizlik durumunda sonuç çıkarırlar. Tümevarımsal sonuç çıkarma, bir hipotezden veya model uzayından bir parametreyi veya bir modeli seçme problemidir. Çalışılan veriyi en iyi açıklar (Baxter, 1996). Akaike (1994)'de incelendiği gibi belirsizlik altında sonuç çıkarma Pierce (1955) tarafından çalışılmıştır. Pierce bunu kaçırma mantığı ya da kısaca kaçırma olarak adlandırdı. Kaçırma, sonuç çıkarmanın bir yoludur, genel prensipler ve yeni gerçekleri elde etmek için gözlenen gerçekleri kullanır. Hepsinin bir belirsizlik derecesi vardır. Kaçırma nümerik fonksiyonları kullanarak yerini alır ve bilgi teorik model seçim kriteri gibi büyüklükleri ölçer. Pierce bilimsel çalışmanın orijinal bölümünün çoğunun kaçırma aşamasıyla veya uygun hipotezlerin seçim aşamasıyla ilgili olduğu konusunda ısrar etmiştir. Bu nedenle karmaşıklığın tasarımı yardımıyla kaçırmaya dayalı sonuç çıkarma için bir sistematik prosedür geliştirmek öğrenme ve değişim/evrimleşme işlemini anlamak öncelikle yapılması gereken işlemdir (Von Neumann, 1966). Bu çerçevede istatistiksel modelleme ve model oluşturma, kaçırma bilimidir. Bu nedenle karmaşıklığın çalışılması uygun hipotezlerin model seçimi veya veri madenciliği işi içinde modeller için oldukça önemlidir.

Aşağıda bir istatistiksel olarak tanımlanmış ölçümden hareketle karmaşıklığın basit sistem teorik tanımı verilmiştir (Bozdogan, 2003).

Tanım 5.1. Herhangi bir tip sistemin karmaşıklığı tüm sistem ve bu sistemin bir basit sayılabilir bileşeni veya kısmı arasında karşılıklı bağımsızlığın derecesinin bir ölçümüdür.

Karmaşıklığın bu tanımının bir modeldeki parametreleri tahmininin sayısı anlamına gelen ve literatürde sıkça kullanılandan farklı olduğuna dikkat ediniz. Amaç

için bir modelin karmaşıklığı modelin bileşenlerinin etkileşimleri cinsinden daha doğal olarak tanımlanır ve gereksinim duyulan bilgi tam olarak tanımlandığı yolla modeli oluşturmaktır. Bu nedenle karmaşıklığın tasarımı bir gerçek dünya sistemi kapsamı içinde ortaya çıkan bir istatistiksel modeli incelemeyle daha iyi açıklanabilir. Örneğin sistem fiziksel, biyolojik, sosyal, davaranış bilimine ait, ekonomik ve buna benzer olabilir, incelenen sistem yanıtları rasgele olmaktadır.

Tanım 5.1' de tanımlanan karmaşıklık olarak tüm sistemin, S olsun, bileşenlerinden ne kadar farklı olduğuyla ilgilenilir. C, bir sistemin karmaşıklığının gerçel-değerli ölçümünü göstersin,bu durumda $C(S)$, tüm sistem ile ayrıştırılan bileşenleri arasındaki farkın miktarını ölçecektir. Bilgi teorik yorum kullanılarak bu miktar eldeki olasılık modelinin ortak dağılımı ile modelin marjinal dağılımlarının çarpımı arasındaki ayrıştırma bilgisi olarak tanımlanır. Ayrıştırma bilgisi dağılımlar aynı ise sıfırdır ve diğer durumda pozitiftir (Van Emden 1971)

Yani, karmaşıklık kavramını bir sabit indeks cinsinde değerlendirmek için etkileşimler bir matematiksel tanımdaki etkileşimleri ifade etmeliyiz. Bu bilgi teorisinin cazibesiyle başarılabilir, çünkü bu eklenebilirlik ve özellikleri kısıtlama ve bağımlılıkları ölçmeye izin gibi geleneksel prosedürlere göre bazı önemli analitik avantajlar taşır (Van Emden 1971; Bozdoğan 1990).

5.4. Çoklu Lineer Regresyon Modeli İçin Bilgi Kriteri

Çoklu lineer regresyon modeli matris formunda,

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon} \tag{5.6}$$

şeklinde ifade edilebilir. Burada $\boldsymbol{y}$, $n\times 1$ tipinde bir vektörü, $\boldsymbol{X}$, $n\times p$ tipinde bir matrisi, $\boldsymbol{\beta}$, $p\times 1$ tipinde bir vektörü ve $\boldsymbol{\varepsilon}$, $n\times 1$ tipinde bir vektörü göstermektedir. $\varepsilon \sim N(0,\sigma^2 I)$ normal dağılımına sahiptir. $i = 1, 2,..., n$ için $\varepsilon_i \sim N(0,\sigma^2)$ normal dağılımına sahiptir (Montgomery ve ark., 2001). Bu durumda gözlemler için regresyon modelinin yoğunluk fonksiyonu,

$$f(y_i|x_i,\beta,\sigma^2)=(2\pi\sigma^2)^{-\frac{1}{2}}\exp\left[-\frac{(y_i-x_i'\beta)^2}{2\sigma^2}\right] \tag{5.7}$$

olarak yazılabilir. Yani rastgele gözlemlerin vektörü $\boldsymbol{y}$, $X\beta$ ortalama vektörlü ve $\sigma^2 I_n$ kovaryans matrisli çok değişkenli normal dağılıma sahiptir. Bu durumda örneklemin likelihood fonksiyonu,

$$L(\beta,\sigma^2|y,X)=(2\pi\sigma^2)^{-\frac{n}{2}}\exp\left[-\frac{(y-X\beta)'(y-X\beta)}{2\sigma^2}\right] \tag{5.8}$$

olur. Log likelihood fonksiyonu ise

$$l(\beta,\sigma^2)=-\frac{n}{2}\log(2\pi)-\frac{n}{2}\log\sigma^2-\frac{(y-X\beta)'(y-X\beta)}{2\sigma^2} \tag{5.9}$$

olur. Magnus ve Neudecker (1999)'un matris differansiyel analizi kullanılarak, (β,σ^2) parametrelerinin $(\hat{\beta},\hat{\sigma}^2)$ maksimum likelihood tahminleri,

$$\hat{\beta}=(X'X)^{-1}X'y \tag{5.10}$$

ve

$$\hat{\sigma}^2=\frac{(y-X\hat{\beta})'(y-X\hat{\beta})}{n} \tag{5.11}$$

olarak elde edilir.

Tahmin edilen regresyon katsayılarının maksimum likelihood kovaryans matrisi,

$$\hat{C}ov(\hat{\beta})=\hat{\sigma}^2(X'X)^{-1} \tag{5.12}$$

dir. Burada X matrisine merkezileştirme ve ölçeklendirme uygulanmamıştır.

Fisher Bilgi Matrisi IFIM,

$$\hat{Cov}(\hat{\beta},\hat{\sigma}^2)=\hat{F}^{-1}=\begin{bmatrix}\hat{\sigma}^2(X'X)^{-1} & 0\\ 0 & \dfrac{2\hat{\sigma}^4}{n}\end{bmatrix} \tag{5.13}$$

olarak tanımlanır (Bozdoğan, 2003).

5.4.1. Karmaşıklık Ölçülerine Dayalı ICOMP Değeri

$C_0(\Sigma)$ ve $C_1(\Sigma)$ sırasıyla,

$$C_0(\Sigma)=\frac{1}{2}\sum_{j=1}^{p}\log(\sigma_{jj})-\frac{1}{2}\log|\Sigma| \tag{5.14}$$

ve

$$C_1(\Sigma)=\frac{p}{2}\log\left[\frac{tr(\Sigma)}{p}\right]-\frac{1}{2}\log|\Sigma| \tag{5.15}$$

olarak tanımlansın. Burada $\sigma_{jj}=\sigma_j^2$ olup, Σ varyans kovaryans matrisinin j-inci köşeğen elemanıdır. p, Σ varyans kovaryans matrisinin boyutudur.

$C_0(\Sigma)$ karmaşıklık tanımına dayalı olarak $ICOMP(\mathrm{Re}\,g)_{C_0}$ değeri,

$$\begin{aligned}ICOMP(\mathrm{Re}\,g)_{C_0} &= -2\log L(\hat{\theta})+2C_0(\hat{Cov}(\hat{\beta}))\\ &= n\log(2\pi)+n\log(\hat{\sigma}^2)+n+2\left[\frac{1}{2}\sum_{j=1}^{q}\log\left(\hat{\sigma}_{jj}(\hat{\beta})\right)-\frac{1}{2}\log\left|\hat{Cov}(\hat{\beta})\right|\right]\\ &= n\log(2\pi)+n\log(\hat{\sigma}^2)+n+2\left[\frac{1}{2}\sum_{j=1}^{q}\log\left(\hat{\sigma}_{jj}(\hat{\beta})\right)-\frac{1}{2}\sum_{j=1}^{q}\log(\lambda_j)\right]\end{aligned} \tag{5.16}$$

olarak tanımlanır (Bozdoğan, 2003). Benzer biçimde $C_1(\Sigma)$ karmaşıklık tanımına dayalı olarak $ICOMP(\mathrm{Re}\,g)_{C_1}$ değeri,

$$ICOMP(\mathrm{Re}\,g)_{C_1} = -2\log L(\hat{\theta}) + 2C_1(\hat{C}ov(\hat{\beta}))$$

$$= n\log(2\pi) + n\log(\hat{\sigma}^2) + n + 2\left[\frac{q}{2}\log\left(\frac{tr(\hat{C}ov(\hat{\beta}))}{q}\right) - \frac{1}{2}\log\left|\hat{C}ov(\hat{\beta})\right|\right]$$

$$= n\log(2\pi) + n\log(\hat{\sigma}^2) + n + 2\left[\frac{q}{2}\log\left(\frac{\overline{\lambda}_a}{\overline{\lambda}_g}\right)\right] \qquad (5.17)$$

olarak tanımlanır (Bozdoğan, 2003).

IFIM 'in tahmini kullanıldığında $ICOMP(IFIM)$,

$$ICOMP(IFIM)_{regresyon} = -2\log L(\hat{\theta}_M) + 2C_1(\hat{F}^{-1}(\hat{\theta}_M))$$

$$= n\log(2\pi) + n\log(\hat{\sigma}^2) + n + C_1(\hat{F}^{-1}(\hat{\theta}_M)) \qquad (5.18)$$

olarak tanımlanır. Burada $C_1(\hat{F}^{-1}(\hat{\theta}_M)) = (q+1)\log\left[\frac{tr\hat{\sigma}^2(X'X)^{-1} + \frac{2\hat{\sigma}^4}{n}}{q+1}\right] - \log\left|\hat{\sigma}^2(X'X)^{-1}\right| - \log\left(\frac{2\hat{\sigma}^4}{n}\right)$ dir.

5.5. Çoklu Lineer Regresyon Modeli İçin Bir Genetik Algoritma

Öncelikle Genetik algoritmada kullanılan gösterimleri açıklayalım. Genetik algoritmanın amacı ve işlevi fazla sayıda açıklayıcı değişken içeren çoklu lineer regresyon modellerinde en iyi modeli bulmaktır. Genetik algoritmada bir model, o modeldeki parametrelerin var olup olmadığına bakılarak ikili sistemde gösterilir. Örneğin:

$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$ modeli ikilik sistemde 1111 biçiminde

$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$ modeli ikilik sistemde 1110 biçiminde

$y = \beta_0 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$ modeli ikilik sistemde 1011 biçiminde

$y = \beta_0 + \beta_1 X_1 + \beta_3 X_3 + \varepsilon$ modeli ikilik sistemde 1101 biçiminde

$y = \beta_0 + \beta_1 X_1 + \varepsilon$ modeli ikilik sistemde 1100 biçiminde

$y = \beta_0 + \beta_2 X_2 + \varepsilon$ modeli ikilik sistemde 1010 biçiminde

$y = \beta_0 + \beta_3 X_3 + \varepsilon$ modeli ikilik sistemde 1001 biçiminde

$y = \beta_0 + \varepsilon$ modeli ikilik sistemde 1000 biçiminde gösterilir.

Regresyon modellerinin 1110, 1011, 1101, 1100, 1010, 1001 ve 1000 gösterimlerinin her birine bir kromozom denir.

Çoklu lineer regresyonda model seçimi problemi için Genetik algoritmaların buradaki uygulaması, Goldberg (1989) tarafından yapılan çalışmaya benzer şekilde gerçekleştirilmiştir. Çoklu lineer regresyon modeli matris formunda,

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon} \tag{5.19}$$

şeklinde ifade edilebilir. Burada $\boldsymbol{y}$, $n \times 1$ tipinde bir vektörü, $\boldsymbol{X}$, $n \times p$ tipinde bir matrisi, $\boldsymbol{\beta}$, $p \times 1$ tipinde bir vektörü ve $\boldsymbol{\varepsilon}$, $n \times 1$ tipinde bir vektörü göstermektedir.

Regresyon modellerinin alt kümesinde model seçim problemi için bir genetik algoritma aşağıdaki adımlar izlenerek oluşturulabilir (Goldberg 1989, Michalewicz 1992). Goldberg tarafından oluşturulan genetik algoritma "Basit Genetik Algoritma" (Simple Genetic Algorithm) olarak adlandırılır. Basit Genetik Algoritma aşağıdaki bileşenleri içerir.

5.5.1. Çoklu Lineer Regresyon Modelleri İçin Bir Genetik Kodlama Şeması

Çoklu lineer regresyon modeli bir ikili string (0 ve 1 karakterlerinin bir dizisi) olarak kodlanır. Çoklu lineer regresyon modelini temsil eden ikili stringin uzunluğu (ikili stringte bulunan 0 ve 1 karakterlerinin sayısı) aynıdır. İkili stringteki her bir konum, açıklayıcı değişkenin modelde bulunmasına (1) veya bulunmamasına (0) bağlı

olarak 1 veya 0 karakterlerinden biriyle doldurulur. Her bir ikili string, açıklayıcı değişkenlerin farklı kombinasyonlarını gösteren bir ikili kodlama içerir. Örneğin, çoklu lineer regresyon modelinde beş açıklayıcı değişken ($k = 5$) bulunsun. Bu çoklu lineer regresyon modeli sabit terim içersin. Böyle bir çoklu lineer regresyon modelin için ikili string kodlaması Tablo 5.1'deki biçimde olabilir (Bozdoğan, 2003).

Tablo 5.1. Beş açıklayıcı değişken bulunan ve sabit terim içeren çoklu lineer regresyon modeli için ikili string gösterimi.

	x_1	x_2	x_3	x_4	x_5
β_0	β_1	β_2	β_3	β_4	β_5
1	0	1	0	1	1

Tablo 5.1 deki ikili string gösterimine göre çoklu lineer regresyon modeli: sabit terim, x_2, x_4 ve x_5 açıklayıcı değişkenlerini içermektedir; bununla birlikte x_1, x_3 açıklayıcı değişkenlerini içermemektedir.

5.5.2. Çoklu Lineer Regresyon Modeli İçin Genetik Algoritmada Kullanılacak Başlangıç Popülasyonunun Oluşturulması

Popülasyon hacmi (bir kitledeki erişkinlerin ya da yetişkinlerin sayısı) ya da diğer bir ifadeyle erişkinler veya yetişkinler için model sayısı N , genetik algoritmanın önemli bir parametresidir. Popülasyon hacmi, bir oluşumda ya da nesilde, bir kitlede kaç kromozom olduğunu belirtir. Çok az sayıda kromozom varsa, genetik algoritma çaprazlama (crossover) yapmak için oldukça az olanağa sahiptir. Bu durumda araştırma uzayının sadece küçük bir kısmı açıklanır. Bununla birlikte, çok fazla sayıda kromozom varsa genetik algoritma çaprazlama (crossover) yapmak için oldukça fazla olanağa sahiptir. Bu durumda araştırma uzayının büyük bir kısmı açıklanır. Bu durumda ise genetik algoritma yavaşlar. Araştırmalar, temel olarak probleme ve çözümlemeye bağlı olarak bazı kısıtlamalardan sonra kitle hacmini arttırmanın kullanışlı olmayacağını göstermiştir. Çünkü bu problemi daha hızlı çözmeyi sağlamaz.

İlk önce N tane erişkin ya da yetişkin model alınır. Bu durumda N, genetik algoritmayı başlatmak için erişkin ya da yetişkin modellerin sayısını gösterir. N rasgele olarak değil, araştırmacının isteğine bağlı olarak seçilir. Buradaki genetik algoritma esnek olduğundan N herhangi bir sayı olarak seçilebilir.

5.5.3. Herhangi Bir Çoklu Lineer Regresyon Modelinin Performansının Değerlendirilmesi için Bir Uyum Fonksiyonu

Çoklu lineer regresyon analizi için genel olarak genetik algoritmada kullanılan uyum fonksiyonu için bir model seçim kriterleri kullanılmalıdır. Bu çalışmada karmaşık bilgi kriteri *ICOMP* (Information COMPlexity) kriteri kullanılacaktır. Analizciler ya da araştırmacılar gereksinimlerine veya önceliklerine dayalı olarak herhangi bir uygun model seçim kriterini seçebilirler (Bozdoğan, 2003).

5.5.4. Oluşturulan Çoklu Lineer Regresyon Modellerini Seçmek için Bir Mekanizma

Bu adım eşleştirme ya da çiftleştirme havuzunda (mating pool) birleştirme için modellerin $ICOMP(IFIM)$ değerlerine dayalı olarak modelleri seçme işleminden oluşur. Burada $IFIM$ (Inverse Fisher Information Matrix), modellerin inverse Fisher bilgi matrislerini göstermektedir. Popülasyonda ya da erişkinlerin veya yetişkinlerin oluşturduğu N tane modelde olası altküme modellerinden her biri için $ICOMP(IFIM)$ değerleri hesaplandıktan sonra popülasyondaki en yüksek kriter değerinden her bir model için kriter değeri çıkarılır. Diğer bir ifadeyle $i = 1, 2, ..., N$ için

$$\Delta ICOMP_{(i)}(IFIM) = ICOMP(IFIM)_{mak} - ICOMP(IFIM)_{(i)} \qquad (5.20)$$

değeri hesaplanır. Burada N popülasyon hacmidir ya da erişkinlerin veya yetişkinlerin oluşturduğu modellerin sayısıdır.

Bir sonraki adımda bu farkların ya da $\Delta ICOMP(IFIM)$ değerlerinin ortalaması hesaplanır. Yani,

$$\overline{\Delta ICOMP(IFIM)} = \frac{1}{N}\sum_{i=1}^{N} \Delta ICOMP_{(i)}(IFIM) \tag{5.21}$$

hesaplanır. Daha sonra da her bir modelin fark değerinin ortalama fark değerine oranı hesaplanır. Yani,

$$\frac{\Delta ICOMP_{(i)}(IFIM)}{\overline{\Delta ICOMP(IFIM)}} \tag{5.22}$$

hesaplanır.

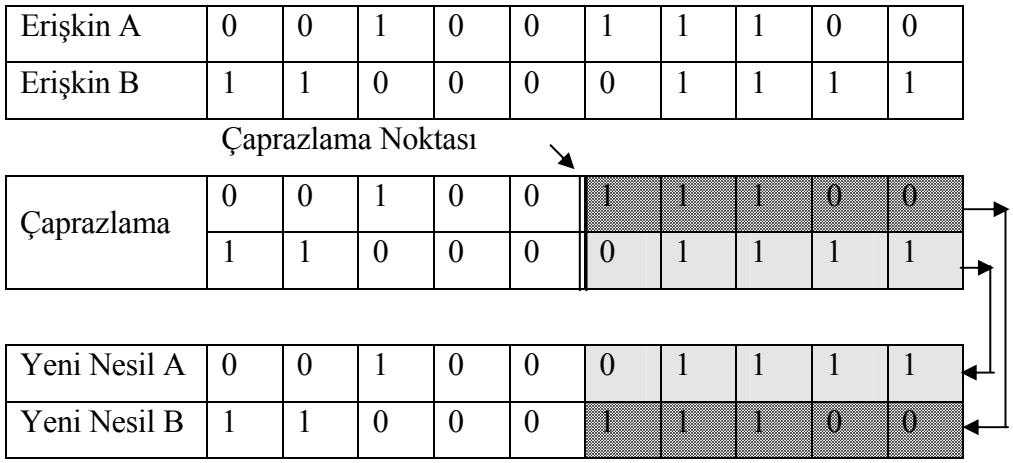
Bu değer hangi erişkinlerin veya yetişkinlerin eşleştirme ya da çiftleştirme havuzuna alınacağına karar vermede kullanılır. Bir modelin eşleştirme ya da çiftleştirme havuzuna alınmasının şansı bu değerle orantılıdır. Yeni nesillerin oluşturulması için erişkinlerin veya yetişkinlerin seçilmesi işlemine yeni nesil modellerin sayısının başlangıçtaki erişkin veya yetişkin modellerin sayısına N (popülasyon hacmi) eşit olana kadar devam edilir. Bu işlem oransal seçim (proportional selection) ya da uyum (fitting) olarak adlandırılır. Aynı zamanda $ICOMP$ ile uyum ya da sıra seçimi (rank selection) vardır (Bearse ve Bozdogan, 2002).

5.5.5. Yeni Nesil Çoklu Lineer Regresyon Modellerini Üretmek İçin Erişkin Modellerin Eşleştirilmesini Yapmak Amacıyla Bir Yeniden Üretim İşlemi

Erişkin modellerin çiftleştirilmesi ya da eşleştirilmesi, bir çaprazlama işlemi olarak hazırlanır. Çaprazlama için seçilen bir erişkin model, $P_ç$ çaprazlama olasılığı veya çaprazlama hızı ile kontrol edilir. $P_ç$ çaprazlama olasılığının sıfır (0) olması, çiftleştirme işlemine alınan elemanların bir sonraki oluşuma taşınması ve herhangi bir

yeni nesilin oluşturulmaması anlamındadır. Çaprazlama olasılığının bir (1) olması, çiftleştirmenin işleminin ya da çaprazlamanın her zaman çiftleştirme işlemi için seçilen herhangi iki erişkin model arasında olacağı yani bir sonraki oluşumun yeni nesil modellerden oluşacağı diğer bir ifadeyle bir önceki oluşumdan herhangi bir model olmayacağı anlamındadır.

Çaprazlama işleminde çaprazlama noktası olarak erişkin modellerin ikili stringlerinin her bir çifti boyunca rasgele bir konum seçilir. Erişkin modellerin herhangi bir çifti için ikili stringleri çaprazlama noktasından iki parçaya ayrılır. Çaprazlama işleminde, çaprazlama noktası olarak adlandırılan bu noktanın sağındaki iki ikili stringin bölümleri iki yeni nesil ikili string oluşturmak amacıyla erişkin modellerin ikili stringleri arasında karşılıklı yer değiştirilir. Bu işlem Şekil 5.1 de gösterilmiştir.



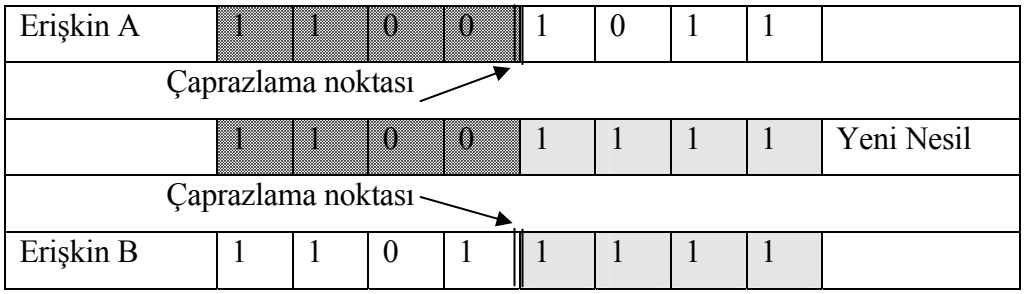
Şekil 5.1. Verilen erişkin iki modelin ikili string çifti için çaprazlama yoluyla çiftleştirme işlemine bir örnek (Bozdogan, 2003)

Bu durumda her bir erişkin model için on yerleşke kullanılmaktadır. Her bir erişkin modelin ikili stringinin uzunluğu boyunca rasgele seçilen bir nokta, çaprazlama noktası olarak seçilir. Erişkin modellerin ikili stringi bu noktaya göre ikiye ayrılır ve iki yeni nesil model üretilir. Üretilen bu iki yeni nesil model daha sonra yeni nesil modeller oluşturmak için erişkin modeller kümesine eklenir. Çaprazlama işleminin

değişik tipleri vardır. Çaprazlama işleminde bunlardan biri seçilebilir. Bu çalışmada tüm pratik amaçlar için yeterli olacağı düşünülen üç tür çaprazlama işlemi verilecektir.

5.5.5.1. Tek Nokta Çaprazlama

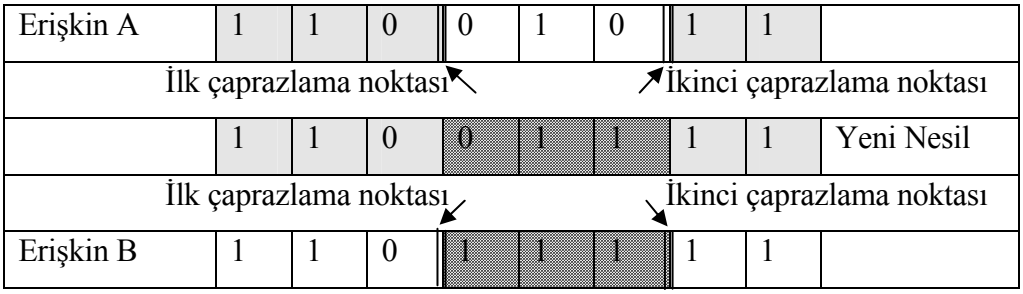
İki erişkin modelin (Erişkin A ve Erişkin B) ikili stringlerini kullanarak yeni nesil modelin ikili stringini oluşturmak amacıyla bir çaprazlama noktası seçilir. Yeni nesil modelin ikili stringi oluşturulurken kromozomun (değişkenin) ikili stringinin başlangıcından çaprazlama noktasına kadar olan bölüm ilk erişkinden (Erişkin A), geri kalan bölüm ikinci erişkinden (Erişkin B) kopyalanır. Bu işlem Şekil 5.2 de gösterilmiştir.



Şekil 5.2. Verilen erişkin iki modelin ikili string çifti için tek nokta çaprazlama yoluyla çiftleştirme işlemine bir örnek (Bozdogan, 2003)

5.5.5.2. İki Nokta Çaprazlama

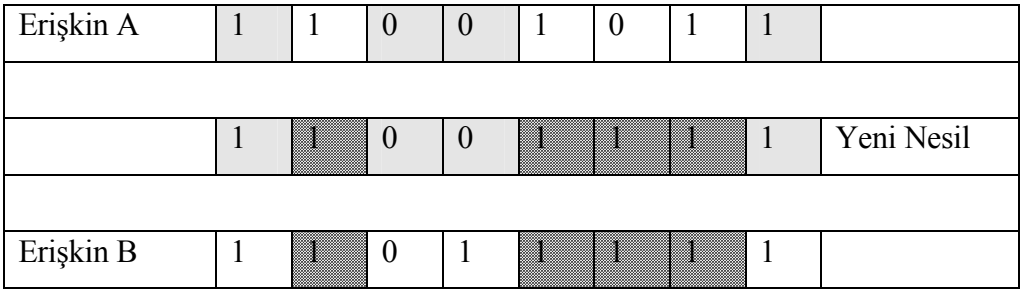
İki erişkin modelin (Erişkin A ve Erişkin B) ikili stringlerini kullanarak yeni nesil modelin ikili stringini oluşturmak amacıyla iki çaprazlama noktası seçilir. Yeni nesil modelin ikili stringi oluşturulurken kromozomun (değişkenin) ikili stringinin başlangıcından ilk çaprazlama noktasına kadar olan bölüm ilk erişkinden (Erişkin A), ilk çaprazlama noktasından ikinci çaprazlama noktasına kadar olan bölüm ikinci erişkinden (Erişkin B) ve geri kalan bölüm yine ilk erişkinden (Erişkin A) kopyalanır. Bu işlem Şekil 5.3 te gösterilmiştir.



Şekil 5.3. Verilen erişkin iki modelin ikili string çifti için iki nokta çaprazlama yoluyla çiftleştirme işlemine bir örnek (Bozdogan, 2003)

5.5.5.3. Düzgün Çaprazlama

Birinci erişkinden (Erişkin A) ve ikinci erişkinden (Erişkin B) bölümler rasgele kopyalanır. Bu işlem Şekil 5.4'te gösterilmiştir.



Şekil 5.4. Verilen erişkin iki modelin ikili string çifti için düzgün çaprazlama yoluyla çiftleştirme işlemine bir örnek (Bozdogan, 2003)

Bu çalışmada kullanılan algoritmada, yukarıdaki çaprazlama seçeneklerinden herhangi biri alınabilir. Ayrıca algoritmada seçicilik kuralı olarak adlandırılan seçme seçeneği vardır. Seçicilik kuralıyla en azından bir en iyi çözümün herhangi bir değişiklik olmaksızın yeni nesile kopyalanır. Böylece en iyi çözüm, algoritmanın çalıştırılması sonucunda elde edilir.

5.5.6. Yeni Nesil Modellerin Birleşimini Değiştirmek İçin Değişmenin Etkisi

Modellerin değişimi, değişkenlerin yeni kombinasyonlarının oluşturulması anlamındadır. Böylece en iyi modeli araştırma işlemi, kısıtlı bir alan yerine uyum yerleşkesinin diğer bir alanına atlayabilir. Bir değişim oranı veya olasılığı belirtilerek, 0 dan 1 e veya 1 den 0 a değişebilen bir konum rasgele seçilerek değişime izin verilebilir. Diğer bir ifadeyle rasgele seçilen bir açıklayıcı değişken modele eklenebilir ya da modelden çıkarılabilir.

Belirli bir çaprazlama türüne ve değişim oranına bağlı olarak ikinci nesil modeller, yeni nesil modellerin ve erişkin modellerin bir karmasıdır. İkinci nesildeki modeller daha sonra üçüncü nesili üretmek için kullanılır. Bu işlem bir oluşumdan sonra araştırmacı ya da analizci tarafından kontrol edilen belirtilmiş sayıda oluşumları üretmek için kullanılır.

Genetik algoritmanın ana hatları özet olarak aşağıda adımlar halinde verilmiştir (Bozdoğan, 2003).

1. Adım: [Algoritmaya başlangıç] N kromozomun (problem için uygun çözümler) rasgele bir popülasyonunu üret. Diğer bir ifadeyle N tane erişkin model üret.
2. Adım: [Uyumun kontrol edilmesi] Model seçim kriterlerinden birini kullanarak popülasyondaki her bir kromozomun uyumunu hesapla. Diğer bir ifadeyle erişkin modellerin uyumunu hesapla.
3. Adım: [Yeni popülasyonun oluşturulması] Aşağıdaki adımları takip ederek yeni popülasyon tamamlanana kadar yeni bir popülasyon oluştur.

3.1. [Seçim] Uyumlarına (örneğin $ICOMP$ değeri) göre bir popülasyondan iki erişkin model (kromozom) seç. (İyi uyum: seçilmek için daha büyük şans)

3.2. [Çaprazlama] Bir çaprazlama olasılığıyla yeni nesil model oluşturmak için erişkin modelleri (kromozomları) çaprazla. Çaprazlama yapılmazsa yeni nesil model, erişkin modellerin tam bir kopyası olur. Üç tür çaprazlama seçeneği vardır.

3.3. [Değişim] Bir değişim olasılığı ile her bir yerleşkede (kromozomun konumu) yeni nesilleri değiştir.

3.4. [Kabul etme] Yeni bir popülasyondan yeni bir nesil yerleştir.

4. Adım: [Yeni nesilin popülasyona yerleştirilmesi] Algoritmanın bir adım ileri çalıştırılması amacıyla yeni nesil modeli kullan ve kullanılan model için seçim kriterinin minimumuna bak.
5. Adım [Deneme] Eğer model seçim kriterine dayalı olarak son koşul sağlanmışsa dur ve halihazırdaki popülasyondan iyi çözümü ver.
6. Adım [Döngü] İkinci adıma git.

Çoklu lineer regresyon modelinde genetik algoritma uygulanarak ve bilgi kriteri kullanılarak en iyi modelin belirlenmesi örneği, Örnek 5.1'de verilmiştir.

Örnek 5.1. Bu örnekte çoklu lineer regresyon modelinde genetik algoritma kullanılacaktır (Bozdogan, 2003). Bu örnekte 13 tane açıklayıcı değişken ya da regresör bulunmaktadır. Yani $k = 13$'tür. Bunlar: $x_1 = $ yaş (yıl), $x_2 = $ Ağırlık (lbs), $x_3 = $ Boy (inch), $x_4 = $ Boyun çevresi (cm), $x_5 = $ Göğüs çevresi (cm), $x_6 = $ Karın 2 çevresi (cm), $x_7 = $ Kalça çevresi (cm), $x_8 = $ Uyluk/pazı çevresi (cm), $x_9 = $ Diz çevresi (cm), $x_{10} = $ Ayak bileği çevresi (cm), $x_{11} = $ Pazı (genişletilmiş) çevresi (cm), $x_{12} = $ Önkol çevresi (cm), $x_{13} = $ Kol bileği çevresi (cm) şeklindedir. Bu açıklayıcı değişkenleri kullanarak Siri (1956) $y = $ Vücut yağ yüzdesi yanıt değişkenini açıklamaya çalışmıştır. Bozdogan (2003), $y = $ Vücut yağ yüzdesi yanıt değişkenindeki toplam değişimin regresyon modelini oluşturmak için regresörlerin en iyi alt kümesini, uygunluk fonksiyonunu *ICOMP* alarak ve Genetik Algoritmayı kullanarak belirlemiştir.

Veriler, sualtı ağırlıkları ve çeşitli vücut çevresi ölçüleri belirlenmiş $n = 252$ insanın vücut yağı yüzdelerinin tahminlerinden oluşmaktadır. Bu örnek GA ile çoklu regresyon analizi kullanılarak yaklaşımımızın çok yönlülüğü ve yararlılığını açıklamada iyi bir örnektir Bozdogan (2003).

Vücut yağının tam / hatasız olarak ölçümü zahmetli ve masraflı olduğundan vücut yağını tahmin etmede zahmetli ve masraflı olmayan kolay yöntemlerin kullanılması arzu edilir. Sağlık kitabı okuyucularının bilinen yöntemleri kullanarak vücut yağ yüzdelerini tahmin ettikleri ve kendi sağlıklarına kısmen de olsa değer biçtiklerini ileri sürmektedir. Okuyucular pergel ile belirledikleri çeşitli deri kıvrım

ölçümlerini ve yaşlarını kullanarak Bailey (1994)'deki tablodan vücut yağını tahmin edebilmektedirler.

Bireysel vücut yağı yüzdesi, vücut yoğunluğu belirlendiğinde tahmin edilebilir. Siri (1956), vücut yapısının yağsız / sıska / zayıf vücut dokusu ve yağ dokusu olmak üzere iki bileşenden oluştuğunu varsaymaktadır.

D = Vücut Yoğunluğu (gm/cm^3), A = Yağsız / zayıf vücut dokusunun oranı, B = Yağ dokusunun oranı ($A+B=1$), a = Yağsız / zayıf vücut dokusunun yoğunluğu (gm/cm^3) ve b = Yağ dokusunun yoğunluğu (gm/cm^3) şeklinde alınırsa Vücut Yoğunluğu,

$$D = 1/\left[(A/a)+(B/b)\right] \tag{5.23}$$

olur. Buradan B çözülürse,

$$B = (1/D)*\left[ab/(a-b)\right]-\left[b/(a-b)\right] \tag{5.24}$$

olur. $a=1.10\ \ gm/cm^3$ ve $b=0.90\ \ gm/cm^3$ alınırsa (Katch ve McAdle 1977, s.111) ya da Wilmore (1976)'daki tahminler kullanılarak Siri (1956)'nin denklemi,

$$Vücut\ Yağ\ Yüzdesi(100*B) = 495/D - 450 \tag{5.25}$$

şeklinde yazılabilir.

Artık vücut yoğunluğu ve hacim çeşitli biçimlerde hatasız / tam olarak ölçülebilir. Sualtı ağırlık tartma tekniği vücut hacmini, havadaki vücut ağırlığı ile su altında ölçülen ağırlık arasındaki fark olarak ölçmektedir. Diğer bir ifadeyle, vücut hacmi su yoğunluğu için düzeltilmiş uygun sıcaklık ile sudaki ağırlık kaybına eşittir (Katch ve McArdcle 1977). Bu teknik kullanılarak, D = Vücut Yoğunluğu,

$$\text{Vücut Yoğunluğu} = WA/\left[(WA-WW)/c.f.-LV\right] \tag{5.26}$$

yazılabilir. Burada $WA =$ Havadaki ağırlık (kg), $WW =$ Sudaki ağırlık (kg), $c.f =$ su düzeltme faktörü, $LV =$ Rezidü / Hata Akciğer Hacmi (litre) dir (Katch ve McArdcle, 1977). Vücut hacmini hesaplamak için diğer yöntemler Behnke ve Wilmore (1974) tarafından verilmiştir.

Bu örnek için öncelikle bütün olası alt küme regresyon modelleri değerlendirilir. Daha sonra ise en küçük / minimum $ICOMP(IFIM)$ değerlerinin derecesine göre en iyi 15 alt küme modeli seçilecektir.

Bütün olası alt küme seçim yöntemi ile bulunan en iyi 15 tane regresyon modeli Tablo 5.2'de verilmiştir.

Tablo 5.2. Vücut yağı verileri için bütün olası modeller arasından en küçük $ICOMP(IFIM)$ değerlerine göre seçilmiş on beş en iyi model (Bozdogan, 2003)

Sıra Numarası	Değişkenler	*ICOMP*
1	1, 4, 6, 7, 8, 12, 13	1473.9065
2	1, 4, 6, 7, 8, 9, 12, 13	1474.5525
3	1, 3, 4, 6, 7, 8, 12, 13	1474.6751
4	1, 4, 6, 7, 8, 10, 12, 13	1475.1721
5	1, 4, 6, 7, 8, 11, 12, 13	1475.2089
6	1, 4, 6, 7, 8, 9, 10, 12, 13	1475.5406
7	1, 3, 4, 6, 7, 8, 9, 12, 13	1475.6024
8	1, 3, 4, 6, 7, 8, 10, 12, 13	1475.7067
9	1, 4, 6, 7, 8, 9, 11, 12, 13	1475.8208
10	3, 4, 6, 7, 12, 13	1475.9539
11	1, 3, 4, 6, 7, 8, 11, 12, 13	1476.0138
12	1, 4, 5, 6, 7, 8, 12, 13	1476.0362
13	4, 6, 7, 12, 13	1476.1600
14	1, 4, 6, 7, 8, 10, 11, 12, 13	1476.3913
15	1, 3, 4, 6, 7, 8, 9, 10, 12, 13	1476.4430

Vücut yağı verileri için Matlab programında hazırlanan GA programının çalıştırılmasında kullanılan parametreleri Tablo 5.3'de verilmiştir.

Tablo 5.3. Vücut yağı verileri için Matlab programında hazırlanan GA programının çalıştırılmasında kullanılan parametreler (Bozdogan, 2003).

Çalıştırmaların sayısı=100	
Nesillerin sayısı	=30
Populasyon Boyutu	=20
Uygunluk Değeri	$= ICOMP(IFIM)$
Çaprazlama olasılığı	=0.5
	(Düzenli çaprazlama kullanıldı.)
Elitizm	Evet
Mutasyon olasılığı	=0.01

Genetik Algoritmanın 100 kez çalıştırılmasından sonra vücut yağı veri kümesi için ilk 10 sıradaki en iyi açıklayıcı değişkenlerin alt kümesi Tablo 5.4'te gösterilmiştir.

Tablo 5.4. Genetik Algoritmanın 100 kez çalıştırılmasından sonra vücut yağı veri kümesi için ilk 10 sıradaki en iyi açıklayıcı değişkenlerin alt kümesi (Bozdogan, 2003).

Genetik Algoritma Sıralaması	Kromozom (Değişkenler)	İkili Gösterimi	$ICOMP(IFIM)$
1	(1) 1, 4, 6, 7, 8, 12, 13	01001011100011	1473.9065
2	(2) 1, 4, 6, 7, 8, 9, 12, 13	01001011110011	1474.5525
3	(3) 1, 3, 4, 6, 7, 8, 12, 13	01011011100011	1474.6751
4	(4) 1, 4, 6, 7, 8, 10, 12, 13	01001011101011	1475.1721
5	(7) 1, 3, 4, 6, 7, 8, 9, 12, 13	01011011110011	1475.6024
6	(8) 1, 3, 4, 6, 7, 8, 10, 12, 13	01011011101011	1475.7067
7	(9) 1, 4, 6, 7, 8, 9, 11, 12, 13	01001011110111	1475.8208
8	(11) 1, 3, 4, 6, 7, 8, 11, 12, 13	01011011100111	1476.0138
9	(13) 4, 6, 7, 12, 13	00001011000011	1476.1600
10	(15) 1, 3, 4, 6, 7, 8, 9, 10, 12, 13	01011011111011	1476.4430

Vücut yağı verisi için en iyi alt küme modelinin uyumunun özeti Tablo 5.5'de gösterilmiştir.

Tablo 5.5. En iyi alt küme modelinin uyumunun özeti (Bozdogan, 2003).

R Kare	= 0.741266
Düzeltilmiş R Kare	= 0.733844
Hata Kareler Ortalamasının Karekökü	= 4.317462
Ortalama Yanıt	= 19.15079
Gözlemler (Veya Ağırlıklar Toplamı)	= 252

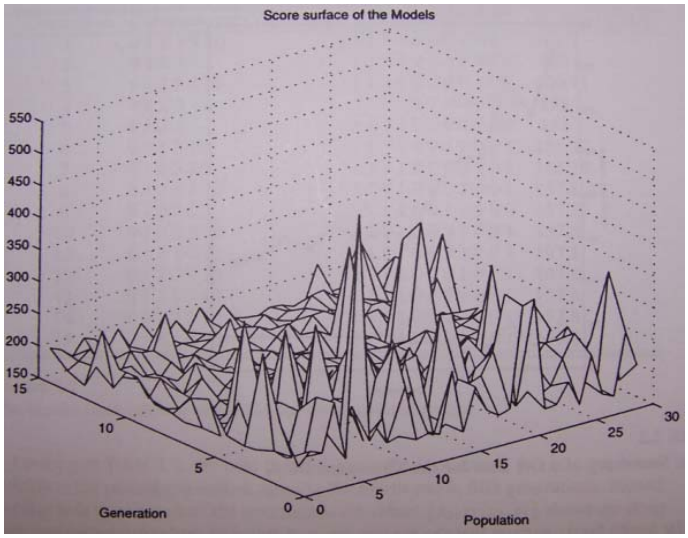
Vücut yağ verisi için en iyi alt küme Genetik Algoritma modelinin parametre tahminleri Tablo 5.6'da verilmiştir.

Tablo 5.6. En iyi alt küme Genetik Algoritma modelinin parametre tahminleri (Bozdogan, 2003).

Terim	Katsayı Tahmini	Standart Hata	t-Oranı	Olasılık>
Sabit	-0.63164	6.498054	-0.10	0.9226
x_1 = yaş (yıl)	0.0838616	0.029956	2.80	0.0055
x_4 = Boyun çevresi (cm)	-0.634546	0.213624	-2.97	0.0033
x_6 = Karın 2 çevresi (cm)	0.8808665	0.066639	13.22	<0.0001
x_7 = Kalça çevresi (cm)	-0.359215	0.118802	-3.02	0.0028
x_8 = Uyluk/pazı çevresi (cm)	0.2826235	0.129812	2.18	0.0304
x_{12} = Önkol çevresi (cm)	0.4529919	0.185745	2.44	0.0155
x_{13} = Kol bileği çevresi (cm)	-1.935856	0.481505	-4.02	<0.0001

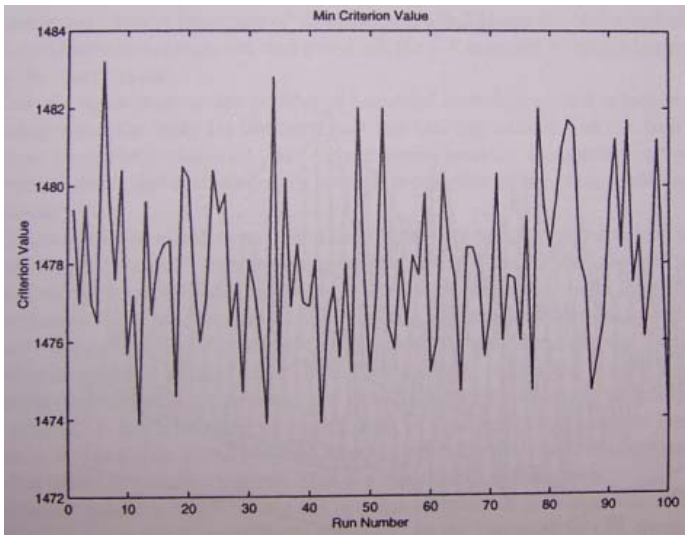
Bu kümeden bir model seçilirse, en iyi alt küme; $x_1 =$ yaş (yıl), $x_4 =$ Boyun çevresi (cm), $x_6 =$ Karın 2 çevresi (cm), $x_7 =$ Kalça çevresi (cm), $x_8 =$ Uyluk/pazı çevresi (cm), $x_{12} =$ Önkol çevresi (cm), $x_{13} =$ Kol bileği çevresi (cm) değişkenlerini içeren ve $ICOMP(IFIM)$=1473.9 değerine sahip olan sıralamadaki ilk modelidir. Gerçekten de bu model, bütün olası alt küme seçiminden elde edilen en iyi modele karşılık gelir. Dikkat edilirse GA seçimleri, bütün olası alt kümelerin sonuçlarından elde edilen zirvedeki yedi en iyi alt kümeye karşılık gelmektedir. GA daha çok sayıdaki alt modelleri ayıklayarak/budayarak en uygun / optimal ya da en uyguna yakın alt küme regresyon modellerini elde edebilen üstün yetenekte istatistiksel bir model seçim aracıdır.

Vücut yağ verisi için ICOMP ile hesaplanan tüm modellerin oluşturduğu yapının üç boyutlu grafiği Şekil 5.5'te gösterilmiştir.



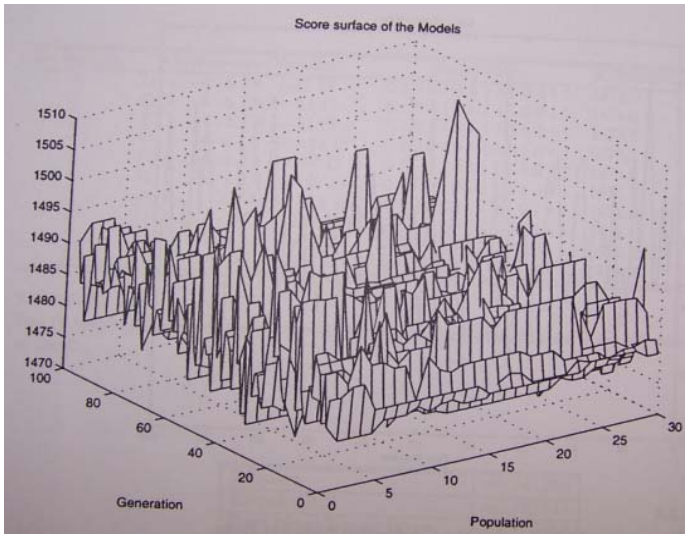
Şekil 5.5. ICOMP ile hesaplanan tüm modellerin oluşturduğu yapının üç boyutlu grafiği (Bozdogan, 2003).

Vücut yağ veririsi için Genetik Algoritmanın 100 kez çalışmasının bir özeti Şekil 5.6'da gösterilmiştir.



Şekil 5.6. Vücut yağ veririsi için Genetik Algoritmanın 100 kez çalışmasının bir özeti (Bozdogan, 2004)

Vücut yağ verisi için Genetik Algoritmanın 100 kez çalıştırılması sonucunda ICOMP(IFIM) ile hesaplanan tüm modellerin oluşturduğu yapının üç boyutlu grafiği Şekil 5.7'de gösterilmiştir.



Şekil 5.7. Vücut yağ verisi için Genetik Algoritmanın 100 kez çalıştırılması sonucunda ICOMP(IFIM) ile hesaplanan tüm modellerin oluşturduğu yapının üç boyutlu grafiği (Bozdogan, 2003)

En iyi alt küme modelinin özeti ve GA ile elde edilen en iyi alt küme modelinin parametre tahminleri Tablo 5.5 ve Tablo 5.6'de verilmiştir. Şekil 5.6 vücut yağı verileri için GA'nın 100 kez çalıştırılmasının özetini ve Şekil5.7 vücut yağı verileri için GA'nın 100 kez çalıştırılması sonucunda bilgi karmaşıklık kriteri ile ölçülen bütün modellerin üç boyutlu grafiğini göstermektedir.

Vücut yağı veriler kümesine ileriye doğru adımsal regresyon analizi uygulanırsa, tam doymuş model en iyi model olarak belirlenebilir. Yani, adımsal yöntemler modeldeki kestiricilerin önemini ayırt edememektedir.Çünkü,adımsal seçim yönteminde kullanılan p-değeri keyfi (isteğe bağlı) olup, F-değeri model arama uzayında en iyi modeli bulmaya yeltenmez.

Bu nedenle derin araştırmacılar çok uygun olmayan bu gibi yöntemleri kullanmayı bırakacaklardır.

6. SONUÇ VE ÖNERİLER

Bu çalışmada bir yanıt (bağımlı) değişkendeki toplam değişimi açıklamak için birden fazla regresör (açıklayıcı) değişken kullanılması durumunda oluşturulan regresyon modeli incelenmiştir.

Çoklu lineer regresyon modelinde en önemli problemlerden biri, k tane açıklayıcı değişken ya da regresör için y yanıt değişkendeki toplam değişimi açıklayan en iyi regresyon modelinin belirlenmesidir.

Bu çalışmada çoklu lineer regresyon modelinde "değişken seçimi" ya da "en iyi alt küme modelinin seçimi" olarak adlandırılan problem incelenmiştir. Bu amaca yönelik olarak:

1. Klasik yöntem kullanılarak model seçimi yöntemi ($1 \le k \le 5$),
2. İleriye doğru seçim, geriye doğru ayıklama ve adımsal regresyon yöntemi kullanılarak model seçimi yöntemleri ($6 \le k \le 10$),
3. Genetik algoritma kullanılarak model seçimi yöntemi ($k \ge 11$),

incelenmiştir.

Tüm bu incelemelerde veride kaldıraç etkisi yapan etkin gözlem değerlerinin bulunması durumu, veride sapan ya da aykırı değer bulunması durumu ve verinin çoklu iç ilişkili olması durumu gözardı edilmiştir.

Özellikle çoklu lineer regresyon modelinde en iyi modelin belirlenmesi problemi çalışılırken genetik algoritmanın uygulanması ve bilgi kriterinin kullanılması durumunda veride kaldıraç etkisi yapan etkin gözlem değerlerinin bulunması, veride sapan ya da aykırı değer bulunması ve verinin çoklu iç ilişkili olması durumları da hesaba katılarak ya ayrı ayrı, veya ikişerli olarak ya da tümü birden incelenmelidir.

KAYNAKLAR

AITKIN, M. A. (1974). Simultaneous inference and the choice of variable subsets, Technometrics, 16, 221-227.

AKAIKE, H. (1973). İnformation theory and extension of the maximum likelihood principle. In B.N Petrov and F. Csáki (Eds.), Second international symposium on information theory, Académiai Kiadó, Budapest, 267-281.

AKAIKE, H. (1987). Factor analysis and AIC. Psychometrika, 52, 317-332.

AKAIKE, H. (1994). Implications of informational point of view on the development of statistical science. In H.Bozdogan (Ed.), Engineering & Scientific applications of informational modeling, Volume 3, pp. 27-38. Proceeding of the first US / Japan conference on the frontiers of statistical modeling: An informational approach. Kluwer Academic Publishers, the Netherlands, Dordrecht.

BAILEY, C. (1994). Smart Exercise: Burning Fat, Getting Fit, Houghton-Mifflin Co., Boston, pp. 179-186.

BAUER, R. J. JR. (1994). Genetic Algorithm and Investment Strategies.John Wiley & Sons, New York.

BAXTER, R. A. (1996). Minimum Message Length Inference: Theory and Applications. Unpublished Ph. D. Thesis, Department of Computer Science , Monash University, Clayton, Victoria, Australia.

BEARSE, P. M. and BOZDOGAN, H. (2002). Multivariate regressions, Genetic Algorithms, and Information Complexity: A three Way Hybrid. In Measurement and Multivariate Analysis, S. Nishisato, Y. Baba, H. Bozdogan, and K. Kanefuji (Eds.), Springer, Tokyo, 2002, 269-278.

BEHNKE, A. R. and WILMORE, J. H. (1974). Evaluation and Regulation of Body Build and Composition, Prentice-Hall, Englewood Cliffs, NJ.

BERK, K.N. (1978). Comparing subset regression procedures, Technometrics, 20(1): 1-6.

BOX, G. E. P., HUNTER, W. G. and HUNTER, J. S. (1978). Statistics for Experimenters, Wiley, New York.

BOYCE, D. E., FAHRİ, A. and WEISCHEDEL, R. (1974). Optimal Subset Selection: Multiple Regression, Interdepence, and optimal Network Algorithms.Springer-Verlag, New York.

BOZDOGAN, H. (1987). Model selection and Akaike's Information Criterion (AIC): The general theory and its analytical extensions. Psychometrika, 52(3), 345-370.

BOZDOGAN, H. (1990). On the information-based measure of covariance complexity and its application to the evaluation of multivariate linear models. Communicatios in Statistics, Theory and Methods, 19, 221-278.

BOZDOGAN, H. (2000). Akaike's information criterion and recent developments in information complexity. Journal of Mathematical Psychology, 44, 62-91.

BOZDOGAN, H. and UENO, M. (2000). A unified approach to information theoretic and Bayesian model selection criteria. Invited paper presented in the Technical Session Track C on: Information Theoretic Methods and Bayesian Modelling at the 6th World Meeting of the International Society for Bayesian Analysis (ISBA). May 28-June 1, 2000. Hersonissos-Heraklion, Crete.

BOZDOGAN, H. (2003). Intelligent statistical data mining with information complexity and genetic algorithms. Statisticals Data Mining and Knowledge Discovery. Joint International Summer School JISS-2003, Vol. II, July 23th – 30th, 2003. Universidade de Lisboa Lisbon, Portugal.

BOZDOGAN, H. (2004). Statisticals Modeling and Model Evaluation: A new Informational Approach. To appear.

CHATTERJEE, S., HADI, A.S. and PRICE, B. (2000). Regression analysis by example, 3rd edition, John Wiley & Sons, New York.

COVER, T. M., GACS, P. and GRAY, R. M. (1989). Kolmogorov's contibutions to information theory and algorithmic complexity. Ann. Prob., 17, 840-865.

COX, D. R. and SNELL, E. J. (1974). The choice of variables in observational studies, Appl. Statist., 23, 51-59.

CRÁMER, H. (1946). Mathematical Methods of Statistics Princeton University Press, Princeton, NJ.

DRAPER, N. R. and SMITH, H. (1989). Applied regression analysis, 3rd edition, John Wiley & Sons, New York.

EDWARDS, J. B. (1969). The relation between the F-test and R^2, Am. Statist., 23, 28.

EFROYMSON, M. A. (1960). Multiple regression analysis in A. Ralston and H. S Wilf (Eds.), Mathematical Methods for Digital Computers, Wiley, New York.

FORREST, S. (1993). Genetic Algorithms in Search Optimization, and Machine Learning, Addison-Wesley, New York.

GOLDGERG, D. E. (1989). Genetic Algorithms in Search, Optimization, and Machine Learning, Addison Wesley, New York.

GRAYBILL, F. A. (1976). Theory and Application of the Linear Model, Duxbury, North Scituate, Mass.

GUNST, R. F. and MASON, R. L. (1980). Regression analysis and its applications, Marcel Dekker, New York.

HAITOVSKİ, Y. (1969). A note on the maximization of $\overline{R}^2$, Am. Statist., 23(1), 20-21.

HOCKING, R. R. (1972). Criteria for selection of a subset regression: Which one should be used, Technometrics, 14, 967-970.

HOCKING, R. R. and LAMOTTE, L. R. (1973). Using the SELECT program for choosing subset regressions, in W. O. Thompson, and F. B. Cady (Eds.), Proceedings of the University of Kentucky Conference on Regression with a Large Number of Predictor Variables, Department of Statistics, University of Kentucky, Lexington.

HOCKING, R. R. (1976). The analysis and selection of variables in linear regression, Biometrics, 32, 1-49, 1044.

HOCKING, R. R. (1983). Development in linear regression methodology: 1959-1982, Technometrics, 25, 219-230.

HOLLAND, J. (1992). Genetic Algoriths. Scientific American, 66-72.

KATCH, F. and MCARDLE, W. (1977). Nutrition, Weight Control, and Exercise, Houghton Mifflin Co., Boston.

KOLMOGOROV, A. N. (1983). Combinatorial foundations of information theory and the calculus of probabilities. Russian Math Surveys, 38, 29-40.

LANNING, M. J. and BOZDOGAN, H. (2003). Ordinal Logistic Modeling Using ICOMP as a Goodness-of-Fit Criteria. In Statistical Data Mining and Knowledge Discovery, H. Bozdogan (Ed.), Chapman & Hall / CRC, Boca Raton, FL.

MAGNUS, J. R. and NEUDECKER, H. (1999). Matrix Differential Calculus, 2nd Edition, John Wiley & Sons, New York.

MALLOWS, C. L. (1964). Choosing variables in a linear regression : A graphical aid, presented at the Central Regional Meeting of the Institute of Mathematical Statistics, Manhattan, Kansas.

MALLOWS, C. L. (1966).). Choosing a subset regression, presented at the Joint Statistical Meetings, Los Angeles.

MALLOWS, C. L. (1973). Some comments on C_p, Technometrics, 15, 661-675.

MANTEL, N. (1970). Why stepdown procedures in variables selection, Technometrics, 12, 591-612.

MICHALEWICZ, Z. (1992). Genetic Algorithms + Data Structures = Evolution Programs, Springer-Verlag, New York.

MILLER, A. J. (1990). Subset selection in regression, London: Chapman and Hall.

MONTGOMERY, D. C., PECK, E.A. and VINING, G. G. (2001). Introduction to Linear Regression Analysis, 3rd Edition, John Wiley & Sons, New York.

MOSES, L. E. (1986). Think and Explain with Statistics, Addison-Wesley, Reading, MA.

MYERS, R. H. (1990). Classical and Modern Regression with Applications, 2nd ed., PWS-Kent Publishers, Boston.

NARULA, S. and RAMBERG, J. S. (1972). Letter to the Editor, Am. Statist., 26, 42.

RAO, C. R. (1945). Information and accuracy attainable in the estimation of statistical parameters. Bull. Calcutta Math. Soc., 37, 81.

RAO, C. R. (1947). Minimum variance and the estimation of several parameters. Proc. Cam. Phil. Soc., 43, 280.

RAO, C. R. (1948). Sufficient statistics and minimum variance estimates. Proc. Cam. Phil. Soc., 45, 213.

RISSANEN, J. (1976). Minmax entropy estimation of models for vector processes. In system identification: R. K Mehra and D. G Lainiotis (Eds.), Academic Pres, New York, 97-119.

RISSANEN, J. (1978). Modeling by shortest data description. Automatica, 14, 465-471.

RISSANEN, J. (1986). Stochastic complexity and modeling. Ann. Statist., 14, 1080-1100.

RISSANEN, J. (1987). Stochastic complexity. (With discussion), J. of the Royal Statist. Soc., Series B, 49, 223-239.

RISSANEN, J. (1989). Stochastic complexity in Statistical Inquiry. World scientific Publishing Company, Teaneck, NJ.

SCLOVE, S. L. (1987). Application of model-selection criteria to some problems in multivariate analysis. Psychometrika, 52, 333-343.

SCHWARZ, G. (1978). Estimating the dimension of a model. Annals of Statistics. 6, 461-464.

SEARLE, S. R. (1971). Linear Models, Wiley, New York.

SEBER, G. A. F. (1977)., Linear Regression Analysis Wiley, New York.

SIRI, W. E. (1956). Gross composition of the body.In Advances in Biological and Medical Physics, Vol. IV, J. H. Lawrance and C. A. Tobias (Eds.), Academic Press, New York.

SOKAL, R. R. and ROHLF, F. J. (1981). Biometry, 2nd ed., W. H Freeman and Company, New York.

SUMIDA, B. H., HOUSTON, A. I., MCNAMARA, J. M. and HAMILTON, W. D. (1990). Genetic Algorithms and evolution. J. Theoretical Biology, 147,59-84.

THOMPSON, M. L. (1978a). Selection of variables in multiple regression: Part I. A review and evaluation,Int. Statist. Rev., 46, 1-19.

THOMPSON, M. L. (1978b). Selection of variables in multiple regression: Part II. Chosen procedures, computations and examples, Int. Statist. Rev., 46, 129-146.

VAN EMDEN, M. H. (1971). An analysis of Complexity. Mathematical Centre Tracts, Amsterdam, 35.

VON NEUMANN, J. (1986). Theory of Self-Reproducing Automata. In A. W. Burks (Ed.), University of illinois Press, :Urbana.

WALLACE, C. S. and FREEMAN, P. R. (1987). Estimation and inference by compact coding. (With discussion). J. Royal Statist. Soc., Series B, 49, 240-265.

WALLACE, C. S. and DOWE, D. L. (1993). MML estimation of the von Mises concentration parameter. Technical Report 93 / 193, Department of Computer Science, Monash University, Clayton 3168, Australia.

WALLET, B. C., MARCHETTE, D. J., SOLKA, J. L. and WEGMAN, E. J. (1996). A genetic algorithm for best subset selection in linear regression, Proceedings of the 28th Symposium on the Interface.

WALLS, R. E., AND WEEKS, D. L. (1969). A note on the variance of a predicted response in regression, Am. Statist., 23, 24-26.

WASSERMAN, G. S. and SUDJIANTO, A. (1994). All subsets regression using a genetic algorithm, Computers and Industrial Engineering, 27(1): 489-492.

WILKINSON, L. (1989). SYSTAT: The System for Statistics, SYSTAT, Evanston, IL.

WILMORE, J. (1976). Athletic Training and Physical Fitness: Physiological Principles of the Conditioning Process, Allyn and Bacon, Inc., Boston.

ÖZGEÇMİŞ

1980 yılında Osmaniye' nin Bahçe ilçesinde doğdum. İlkokula Düziçi Cumhuriyet İlkokulunda 1987 yılında başladım. İlkokul eğitimimi Düziçi Üzümlü İlkokulunda 1992 yılında tamamladım. Ortaokul öğrenimimi Adana Anadolu Lisesi'nde 1996 yılında tamamladım. Lise öğrenimimi de Adana Anadolu Lisesinde 1999 yılında tamamladım. 1999 yılında Ankara Üniversitesi Fen Fakültesi Matematik bölümüne girdim. 2003 yılında buradan mezun oldum. Aynı yıl Çukurova Üniversitesi Fen Bilimleri Enstitüsü İstatistik Anabilim Dalı'nda yüksek lisans öğrenimime başladım. Halen eğitimime Çukurova Üniversitesi Fen Bilimleri Enstitüsü İstatistik Anabilim Dalı'nda yüksek lisans öğrencisi olarak devam etmekteyim.