ÇUKUROVA ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

YÜKSEK LİSANS TEZİ

Azer ÖNCÜ

LİNEER REGRESYONDA DEĞİŞKEN SEÇİMİ

İSTATİSTİK ANABİLİM DALI

ADANA, 2022

ÖZ

YÜKSEK LİSANS TEZİ

LİNEER REGRESYONDA DEĞİŞKEN SEÇİMİ

Azer ÖNCÜ

ÇUKUROVA ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ
İSTATİSTİK ANABİLİM DALI

Danışman : Prof. Dr. Selahattin KAÇIRANLAR
Yıl: 2022, Pages:99
Jüri : Prof. Dr. Selahattin KAÇIRANLAR
: Doç. Dr. Nimet ÖZBAY
: Dr.Öğr.Üyesi Erkut TEKELİ

Çoklu doğrusal regresyonda model için uygun bağımsız değişken alt kümesinin bulunması değişken seçim problemi olarak adlandırılır. Bağımsız değişkenlerin, bağımlı değişkene etkilerinin anlamlı olması ve bağımlı değişkeni açıklayabilmesi açısından hangi değişkenlerin modele dahil edileceği, hangi değişkenlerin model dışı edileceğine karar verilmesi önemli bir problemdir.

Bu amaçla çalışmada önce, çoklu lineer regresyon modelinde en iyi modelin seçiminde kullanılan bazı kriterler incelenmiştir. Daha sonra çoklu lineer regresyon modelinde, verilerde çoklu iç ilişki ve/veya sapan değerler olması durumunda model seçiminde kullanılan bazı kriterler ele alınmıştır.

Anahtar Kelimeler: Model seçimi, Çoklu iç ilişki, Sapan değer , Yanlı tahmin edici, Dayanıklı tahmin edici

ABSTRACT

MSc THESIS

VARIABLE SELECTION IN THE LINEAR REGRESSION MODEL

Azer ÖNCÜ

ÇUKUROVA UNIVERSITY
INSTITUTE OF NATURAL AND APPLIED SCIENCES
DEPARTMENT OF STATISTICS

Supervisor : Prof. Dr. Selahattin KAÇIRANLAR
Year: 2022, Pages:99
Jury : Prof. Dr. Selahattin KAÇIRANLAR
: Assoc. Prof. Dr.Nimet ÖZBAY
: Assist. Prof. Dr. Erkut TEKELİ

Finding the appropriate subset of independent variables for the model in multiple linear regression is called the variable selection problem. It is an important problem to decide which variables will be included in the model and which variables will be excluded from the model, in order that the effects of the independent variables on the dependent variable are significant and can explain the dependent variable.

For this purpose, some criteria used in the selection of the best model in the multiple linear regression model were firstly examined. Then, in the multiple linear regression model, in case of multicollinearity and/or outliers in the data, some criteria used in model selection were discussed.

Key Words: Model selection, Multicollinearity, Outlier, Biased estimator, Robust Estimator

GENİŞLETİLMİŞ ÖZET

Regresyon, değişkenler arasındaki ilişki ve bağıntıların incelenmesini sağlayan önemli bir istatistiksel metottur. Regresyon kuramı günümüzde birbirinden farklı birçok alanda kullanılabilmektedir. Regresyon analizi, değişkenler arasındaki bağıntının en iyi şekilde açıklandığı bir modele dayalıdır. Bir bağımlı değişkendeki toplam değişimi açıklamak amacıyla birden fazla açıklayıcı değişken kullanılarak oluşturulan regresyon modeline çoklu regresyon modeli denir. Çoklu regresyon ve model seçimi uzun bir geçmişe sahip olup, bu konudaki çalışmalar hala çok ilgi çekmeye devam etmektedir. Verinin özelliklerine göre tahmin edicinin belirlenmesi, değişen seçimi ve buna bağlı olarak doğru modelin oluşturulması önemli bir problemdir.

Çoklu lineer regresyon modelinde, y bağımlı değişkendeki toplam değişimi açıklayan en iyi regresyon modelinin seçilmesi "değişken seçimi" ya da "en iyi alt küme modelinin seçimi" olarak adlandırılır . k tane açıklayıcı değişken içeren çoklu lineer regresyon modeli için 2^k tane aday model vardır . 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 mümkün olduğunca fazla açıklayıcı değişken içermesi istenir. Çünkü modelde bulunan değişkenlerdeki bilgi içeriği, tahmin edilen bağımlı değişken değerlerini etkiler (Montgomery ve ark., 2001).

En iyi regresyon modelinin belirlenmesinde yaygın olarak çoklu belirleyicilik katsayısı R^2 veya düzeltilmiş çoklu belirleyicilik katsayısı $\overline{R}^2$, hata kareler ortalaması (MSE), öngörü hata kareler ortalaması (MSEP) gibi kriterler kullanılır. Ayrıca, Mallows (1973)'un C_p kriteri, Golub ve arkadaşları (1979) tarafından tanımlanan genelleştirilmiş çapraz geçerlilik kriteri (GCV), Allen (1971) tarafından tanımlanmış öngörü hata kareler toplamı (PRESS) kriteri, Akaike (1974)

bilgi kriteri (AIC), Hurvich ve Tsai (1989)'un tanımladığı düzeltilmiş Akaike bilgi kriteri (AIC_c), Schwarz (1978) tarafından tanımlanmış bayesyen bilgi kriteri (BIC) gibi bir çok kriter kullanılmaktadır (Draper ve Smith,2003, Miller, 1990, Montgomery ve ark., 2001, Lindsey ve Sheather, 2010, Delaney ve Chatterjee, 1986, Linhart ve Zucchini,1986).

Tanımlanan değişken seçimi yöntemlerinin hiçbiri verilen veri seti için en iyi regresyon denklemini elde etmeyi garantilemez. Aslında genellikle tek bir en iyi regresyon denklemine nispeten birkaç iyi model bulunabilir. Regresyonda model seçimi konusunda önemli çalışmalar Cox ve Snell (1974), Hocking (1972, 1976), Myers (1990), Hocking ve LaMotte (1973) ve Thompson (1978a,b) tarafından verilmiştir.

Eşit sayıda açıklayıcı değişken içeren modellerin karşılaştırılmasında çoklu belirleyicilik katsayısı R^2 ve farklı sayıda açıklayıcı değişken içeren modellerin karşılaştırılmasında $\overline{R}^2$ kullanılır. En iyi regresyon modelinin belirlenmesinde R^2 veya $\overline{R}^2$ 'si yüksek , MSE'si düşük ve az sayıda açıklayıcı değişken içeren modeller tercih edilir (Montgomery ve ark., 2001).

Çalışmanın ikinci bölümünde çoklu lineer regresyon modeli hakkında genel bilgiler ve veride çoklu iç ilişki ve/veya sapan değerler olması durumunda parametrelerin tahmin yöntemleri verilmiştir.

Üçüncü bölümde, çoklu lineer regresyon modellerinin oluşturulmasında ve en iyi modelin belirlenmesinde kullanılan kriterler ve kullanım amaçları anlatılmıştır.

Dördüncü bölümde**,** en iyi modelin belirlenmesinde kullanılan kriterlerin yanlı tahmin edicilere uyarlanmış versiyonları incelenecektir.

TEŞEKKÜR

Bu çalışmanın hazırlanması sırasında bilgi ve tecrübeleriyle, yapıcı ve yönlendirici fikirleriyle bana yol gösteren , değerli zamanını ayırarak çalışmamın tamamlanmasını sağlayan danışman hocam Sayın Prof.Dr. Selahattin KAÇIRANLAR'a teşekkürlerimi sunarım.

İÇİNDEKİLER **SAYFA**

ÇİZELGELER DİZİNİ **SAYFA**

ŞEKİLLER DİZİNİ **SAYFA NO**

SİMGELER VE KISALTMALAR

$\hat{\beta}(k)$: Ridge Tahmin Edici

$\hat{\beta}_M$: M Tahmin Edici

MSE_p : Artık(Hata) Kareler Ortalaması

R^2 : Çoklu Belirleyicilik Katsayısı

$\hat{\alpha}_{ML}$: Dayanıklı Genelleştirilmiş Liu Tahmin Edici

$\hat{\alpha}^*_{ML}$: Dayanıklı Hemen Hemen Yansız Genelleştirilmiş Liu Tahmin Edici

RT_p : Dayanıklı T_p İstatistiği

$\hat{\beta}_{JRM}$: Dayanıklı Jacknifed Ridge Tahmin Edici

RC_p : Dayanıklı Mallows İstatistiği

$\hat{\alpha}_M(k)$: Dayanıklı Ridge Tahmin Edici

$\hat{\beta}$: En küçük kareler tahmin edici

GS_p : Genelleştirilmiş S_p İstatistiği

$\hat{\beta}(d)$: Liu Tahmin Edici

GCV_d : Liu Tahmin Edici İçin Çapraz Geçerlilik Kriteri

C_d : Liu Tahmin Edici İçin Mallows İstatistiği

$R_p(d)$: Liu Tahmin Edici İçin Model Seçim İstatistiği

C_p : Mallows İstatistiği

R_p^2 : p Terimli Model İçin Çoklu Belirleyicilik Katsayısı

$\bar{R}_p^2$: p Terimli Model İçin Düzeltilmiş Çoklu Belirleyicilik Katsayısı

GCV_k : Ridge Tahmin Edici İçin Çapraz Geçerlilik Kriteri

C_k : Ridge Tahmin Edici İçin Mallows İstatistiği

$R_p(\lambda)$: Ridge Tahmin Edici İçin Model Seçim İstatistiği

S_p : Sapan değerler Varlığında Kashid ve Kulkarni İstatistiği

$\overline{R}^2$:Düzeltilmiş Çoklu Belirleyicilik Katsayısı

$\hat{\beta}_{LM}(d)$: Dayanıklı Liu Tahmin Edici

$PRESS_d$: Liu Tahmin Edici İçin Öngörü Hata Kareler Toplamı

$PRESS_k$: Ridge Tahmin Edici İçin Öngörü Hata Kareler Toplamı

AIC (AIC_C) : Akaike Bilgi Kriteri (Düzeltilmiş Akaike Bilgi Kriteri)

AICF : Fisher Bilgi Kriterine Bağlı AIC

BIC : Bayesyen Bilgi Kriteri

EKK : En Küçük Kareler

GCV : Çapraz Geçerlilik Kriteri:

GLE : Genelleştirilmiş Liu Tahmin Edici

IRWLS : İteratif Yeniden Ağırlıklandırılmış En Küçük Kareler

ME : M Tahmin Edici

MSE : Hata Kareler Ortalaması

MSEP : Öngörü Hata Kareler Ortalaması

OLSE : En Küçük Kareler Tahmin Edici

PRESS : Öngörü Hata Kareler Toplamı

RAIC (AIC_C) : Dayanıklı Akaike Bilgi Kriteri

VIF : Varyans Şişirme Faktörü

1.GİRİŞ

Regresyon, değişkenler arasındaki ilişki ve bağıntıların incelenmesini sağlayan önemli bir istatistiksel metottur. Regresyon kuramı günümüzde birbirinden farklı birçok alanda kullanılabilmektedir. Regresyon analizi, değişkenler arasındaki bağıntının en iyi şekilde açıklandığı bir modele dayalıdır. Çok sayıda faktöre bağlı olarak değişim gösteren sosyal, psikolojik ve ekonomik olayların gerisindeki sebep-sonuç ilişkisini ortaya çıkarabilmek için kullanılan istatistiksel yaklaşımlardan biri, çoklu regresyon analizidir. Bu yaklaşımda, bir veya daha çok açıklayıcı değişken bir yanıt değişken ile birlikte seçilerek, yanıt değişkenin gerçek ölçümleri ile açıklayıcı değişkenler yardımıyla elde edilen tahmini ölçümleri arasındaki mesafeyi minimum kılan en küçük kareler (EKK) yöntemi ile regresyon katsayıları tahmin edilir. Örneklemden elde edilen bu regresyon denklemi ile değişkenler arasında var olan sebep-sonuç ilişkisini belirlemenin yanında geleceğin tahmini de daha güvenli bir şekilde yapılabilmektedir.

Lineer regresyon analizinde; EKK yönteminin uygulanması için belli varsayımların sağlanması gerekmektedir. Bu varsayımlardan özellikle, açıklayıcı değişkenlerin lineer bağımsız olması ve tüm gözlemlerin kestirilmiş regresyon doğrusu üzerinde aynı etkiye sahip olması çok önemlidir. Bu iki temel varsayım genellikle uygulamalarda sağlanamamaktadır. Bu durumda çoklu iç ilişki (multicollinearity) ve sapan değerlerin (outliers) varlığı problemi olarak adlandırılan iki problem ile karşılaşılır.

Açıklayıcı değişkenler arasında lineer bağıntı olması çoklu iç ilişki problemine neden olmaktadır. Bu durumda EKK tahmin edici (OLSE) hala yansızdır fakat varyansı çok büyüktür. Bu problemin etkisini minimuma indirgemek için önerilen yöntemler genel olarak ek verinin toplanması, modelin yeniden belirlenmesi ve yanlı tahmin edicilerin kullanılmasıdır.

Diğer taraftan, çoklu iç ilişkinin nedeni model seçiminden kaynaklanabilir. Bu durumda ya açıklayıcı değişkenler yeniden tanımlanır ya da ilişkili açıklayıcı değişkenlerden biri çıkarılır. Fakat açıklayıcı değişkenlerin birinin çıkartılması modelin etkinliğini azaltabilir. Çünkü çoklu iç ilişki olsa bile değişkenler birbirini tam temsil etmeyebilir .

Çoklu iç ilişkinin etkisini azaltmak için kullanılan diğer bir yöntem yanlı tahmin ediciler tanımlamaktır. Yanlı tahmin edicilerde temel amaç OLSE'nin yansızlığından vazgeçip küçük bir yanlılık terimi ekleyerek tahmin edicinin varyansını minimize etmektir. Böylece yanlı tahmin edicilerin performansı, tahmin edicinin varyansının ve yanlılığının karesinin toplamı olan hata kareler ortalaması (MSE) ile ölçülmektedir. Yanlı tahmin edicilerle ilgili literatürde birçok çalışma bulunmaktadır. Bu çalışmada da bunlardan bazılarının kullanımı ile model seçimi konusu ele alınacaktır.

Diğer yandan sapan değerlerin OLSE üzerinde önemli derecede olumsuz etkileri vardır. Bu sorunu ortadan kaldırmak amacıyla regresyon tanısı (regression diagnostic) ve dayanıklı tahmin ediciler (robust estimators) olmak üzere temel iki yaklaşım kullanılır. Regresyon tanısında amaç sapan değerleri kesin olarak belirlemek ve sapan değer olarak saptanan gözlemler düzeltildikten ya da veriden çıkarıldıktan sonra OLSE uygulamaktır. Dayanıklı regresyonda ise sapan değerlerden etkilenmeyen dayanıklı tahmin ediciler üretilmesi amaçlanmaktadır. Veri kümesinde sapan değerler bulunması durumunda OLSE tutarlı ve güvenilir sonuçlar vermemektedir. Bu tahmin edici için tek bir sapan değer bile tahminleri olumsuz olarak etkileyebilir. Bu nedenle OLSE'ye alternatif olarak daha dayanıklı tahmin ediciler geliştirilmeye çalışılmıştır.

Literatürde, dayanıklı tahmin ediciler içerisinde yer alan en geniş tahmin edici sınıflarından birisi de M-tahmin edicidir (M-estimator, ME). ME'de temel yaklaşım; hataların karelerini minimum yapmak yerine, hataların bir başka fonksiyonunu minimize etmektir. ME için literatürde yapılan bazı çalışmalar:

Tukey (1970), Andrews (1974), Mallows (1975), Dutter (1977), Holland ve Welsch (1977), Huber (1981) ve Hampel ve ark. (1986) şeklindedir.

Diğer taraftan regresyon analizinde veri kümesinde çoklu iç ilişki ve sapan değer probleminin aynı anda bulunması çok sık karşılaşılan bir problemdir. Bu durumda hem çoklu iç ilişkinin etkisini hem de sapan gözlemlerin etkisini azaltmak için yanlı ve dayanıklı tahmin edicilerin birlikte bulunduğu dayanıklı tahmin ediciler tanımlanmıştır. Bu çalışmalar genellikle iki yöntem üzerinde ilerlemiştir: Birincisi hatalara farklı ağırlıklar vererek dayanıklı ağırlıklı yanlı tahmin edicilerin tanımlanmasıdır. Bu yöntem doğrultusunda yapılan bazı çalışmalar şu şekildedir: Holland (1973), Askin ve Montgomery (1980), Pfaffenberger ve Dielman (1985, 1990). Diğer yöntemde ise yanlı tahmin edicilerden bazıları OLSE'nin bir formu şeklinde yazıldığından ve OLSE sapan değerlere karşı hassas olduğundan OLSE yerine dayanıklı tahmin ediciler kullanılarak, (y-yönündeki sapan değerler için) dayanıklı yanlı tahmin edicilerin tanımlanmasıdır. Bu yöntem doğrultusunda yapılan bazı çalışmalar şu şekildedir: Silvapulle (1991), Arslan ve Billor (1996), Arslan ve Billor (2000), Jadkav ve Kashid (2011) ,Kan ve ark. (2013), Ertaş ve ark.(2015,2017).

2. ÇOKLU LİNEER REGRESYON MODELİ

$$y = X\beta + \varepsilon \tag{2.1.}$$

çoklu lineer regresyon modeli verilmiş olsun. Burada y, $n\times 1$ tipinde yanıt değişkenlerin vektörü, X, $n\times p$ tipinde stokastik olmayan açıklayıcı değişkenlerin gözlenen matrisi, β, $p\times 1$ tipinde bilinmeyen regresyon parametrelerinin vektörü, ε, $\varepsilon \sim N(0,\sigma^2 I_n)$ olan $n\times 1$ tipinde rasgele hataların vektörüdür. Çoklu lineer regresyonda amaç, iki veya daha fazla açıklayıcı değişken ile yanıt değişken arasında doğrusal bağlantının varlığını araştırmaktır.

2.1. Çoklu İç İlişki

Model (2.1.)'deki çoklu lineer regresyon modelini düşünelim ve,

$$\sum_{j=1}^{p} a_j x_j = 0 \tag{2.2.}$$

şeklinde tamamı sıfır olmayan $a_1, a_2, ..., a_p$ sabit kümesi var ise $x_1, x_2, ..., x_p$ vektörleri lineer bağımlıdır ve doğrusal bağlantılı olduğu görülmektedir. Herhangi bir X sütunlarının kümesi için, bu ifade sağlanıyorsa, tam çoklu iç ilişki olduğu söylenir (Silvey, 1969).

Eğer, X'in sütunlarının bir takım alt kümeleri için yaklaşık olarak doğrusal bir ilişki varsa, yaklaşık çoklu iç ilişki vardır. $X'X$ matrisi tekile yakın fakat, tekil olmayan bir matristir $\left(\left|X'X \neq 0\right|\right)$. Bu durumda, $X'X$ matrisinin tersi alınabilmektedir. Fakat, OLSE yönteminde parametrelerin standart hataları çok daha büyük elde edilecektir.

2.1.1. Çoklu İç İlişkinin Nedenleri

1. Örnekleme Yöntemleri: Açıklayıcı değişkenlerin aldığı değerlerin kısıtlı bir uzayından örneklem alınması çoklu iç ilişkiye sebep olur.

2. Model veya Kitledeki Kısıtlamalar: Model veya kitledeki kısıtlamalar genellikle çoklu iç ilişkiye sebep olur.

3. Modelin Kurulması: X açıklayıcı değişkenlerin aralığı küçük iken, herhangi bir regresyon modeline polinomal ifadelerin dahil edilmesi çoklu iç ilişkiye sebep olabilir.

4. Modelin Aşırı Tanımlanması Durumu (Overdefined): Açıklayıcı değişken sayısının gözlemlerden fazla olduğu modeller aşırı tanımlanmış model olarak tanımlanır. Daha çok tıbbi ve deneysel bilimlerde karşılaşılan aşırı tanımlanmış modeller, açıklayıcı değişkenler ile alakalı verilerin elde edilemediği zamanlarda ortaya çıkmaktadır. Bu gibi durumlarda, bazı açıklayıcı değişkenleri modelden çıkararak, modeli daha az sayıda açıklayıcı değişkenler ile yeniden oluşturmak gerekmektedir.

2.1.2. Çoklu İç İlişkinin Belirlenmesi

Çoklu iç ilişkinin belirlenmesi için bir çok yöntem geliştirilmiştir. Bu yöntemlerin bazıları şunlardır:

1. Açıklayıcı Değişkenler Arasındaki Korelasyonların Analiz Edilmesi: Açıklayıcı değişkenler arasındaki korelasyon katsayılarının yüksek derecede olması çoklu iç ilişkinin varlığını ifade eder.

2. Varyans Şişirme Faktörü (VIF): Parametre tahminlerinin ve varyanslarının çoklu iç ilişki sebebi ile asıl değerlerinden ne kadar uzaklaştığını ifade eder.

$$VIF_j = \frac{1}{1-R_j^2} \quad , \quad 0 \le R_j^2 \le 1 \qquad (2.3.)$$

Her bir tahmin edicinin çoklu belirleyicilik katsayısı olan R_j^2 , X_j 'nin geriye kalan $(p-1)$ açıklayıcı değişkenle aralarında lineer bir ilişki söz konusuysa 1'e yaklaşır ve VIF_j büyür. $VIF_j > 10$ ise, çoklu iç ilişkiden söz edilebilir.

3. X'X Matrisinin Özdeğerlerinin Analizi: Çoklu iç ilişki durumunda $X'X$ korelasyon matrisinin özdeğerlerinden en az biri küçük olacaktır. X'in sütunlarındaki yaklaşık lineer bağımlılık, özdeğerlerin küçük olmasından kaynaklı olduğunu gösterir. $\lambda_{\max}$ ve $\lambda_{\min}$, $X'X$ matrisinin maksimum ve minimum özdeğerlerini ifade eder ve koşul sayısı

$$K = \frac{\lambda_{\max}}{\lambda_{\min}} \qquad (2.4.)$$

şeklinde tanımlanır.

K değerinin 100'den büyük olması kuvvetli çoklu iç ilişki olduğunu gösterir. Bunun haricinde çoklu iç ilişkinin belirlenmesinde bir diğer ölçüt koşul indeksidir.

$$\sqrt{K} = \sqrt{\frac{\lambda_{\max}}{\lambda_{\min}}} \qquad (2.5.)$$

şeklinde hesaplanır.

Koşul indeksi, 10 ile 30 arasında bir değer elde edilmesi durumunda, orta dereceden biraz daha yüksek derecede, 30'dan büyük bir değer elde edilmesi durumunda ise kuvvetli derecede çoklu iç ilişkinin olduğu kabul edilir.

4. ($\mathbf{X}'\mathbf{X}$ Matrisinin Determinantı): Korelasyon matrisinin determinantının değer aralığı $0 \le |X'X| \le 1$ 'dir. $|X'X| = 1$ ise açıklayıcı değişkenler ortogonaldir ve çoklu iç ilişkiden söz edilemez. Diğer türlü, $|X'X| = 0$ durumunda açıklayıcı değişkenler arasında tam çoklu iç ilişki söz konusudur. Eğer, $|X'X|$ değeri sıfıra yaklaşırsa tama yakın çoklu iç ilişkiden söz edilebilir .

Çoklu iç ilişkinin belirlenmesinde yararlanılan bu farklı yöntemlerin birlikte değerlendirilmesi daha uygun olacaktır.

2.1.3. Çoklu İç İlişkinin Sonuçları

1. Çoklu iç ilişki EKK tahmin edicilerin olduğundan daha yüksek değere sahip olmasına sebep olur. Bu da EKK tahmincisinin parametre vektörü ile tahmin vektörü arasındaki uzaklığın karesinin büyük olmasını açıklamaktadır.
2. Parametre tahmin edicilerinin standart hataları büyük olduğunda, t değerinin sıfırdan önemli ölçüde farklı olmadığı görülür. Yani t değeri küçük bulunur. Bunun sonucunda t-testi anlamsız bulunur ve H_0 hipotezi kabul edilir. Bu gibi durumlar, tam çoklu iç ilişki varlığında ortaya çıkar. Ayrıca, tam çoklu iç ilişki durumunda parametreler tahmin edilemez, R^2 değeri yüksek çıkar ve F-testi anlamlı olur.

2.1.4. Çoklu İç İlişki Probleminin Çözüm Yöntemleri

1. Ek Bilginin Toplanması:

Çoklu iç ilişki probleminde ek verilerin toplanması ve bunların analize dahil edilmesi bu problemi büyük oranda ortadan kaldırır. Ne yazık ki, bazı kısıtlamalar sebebiyle veri toplamak her zaman mümkün olmayabilir.

2. Modelin Yeniden Belirlenmesi (Değişken Seçimi):

Çoklu iç ilişki probleminin kaynağı, model ve buna bağlı olarak açıklayıcı değişken seçimidir. Açıklayıcı değişkenlerin tekrardan tanımlanması veya modelden çıkarılması çoklu iç ilişki problemini ortadan kaldırmada iyi bir metot olacaktır. Fakat, modelden açıklayıcı değişken çıkarılması sonucunda katsayı tahminleri gerçek değerinin dışında tahmin edilebilir. Bu da modelin etkinliğini azaltabilir.

3. Değişken Dönüştürme Yöntemi:

Çoklu iç ilişki problemini ortadan kaldıran ya da şiddetini azaltan bu yöntem daha çok zaman serilerinde kullanılır.

4. Birbirleriyle bağlantılı olan iki açıklayıcı değişkenin toplanarak tek bir açıklayıcı değişken olarak alınmasıyla çoklu iç ilişki problemi çözülebilir.

5. Son olarak ridge regresyon yöntemi, Liu tahmin ediciler ve bunlara bağlı olarak tanımlanan yanlı tahmin yöntemleri kullanılır.

2.2. Sapan Değerler

Çoklu lineer regresyon analizinde, veri kümesinde sapan değerlerin bulunması durumunda OLSE güvenilir sonuçlar vermemektedir. Bu nedenden dolayı sapan gözlemlerin belirlenmesi modelin uygunluğu, güvenirliği ve kararlılığı için gereklidir. En genel manada verilerin homojen çoğunluğu tarafından önerilen modele uyumsuzluk gösteren gözlem veya gözlemlere sapan değer(ler) denir.

Sapan değerler regresyon analizi sonuçları üzerinde yaptıkları etkilere bağlı olarak bağımlı değişken yönünde (*y*- yönünde) sapan değerler, bağımsız değişkenler yönünde (*X*-yönünde) sapan değerler, hem bağımlı hem de bağımsız değişkenler yönünde sapan değerler ve etkili gözlemler (influential observations) olmak üzere dört grupta incelenir.

- Lineer regresyonda, regresyon doğrusunun uzağında olan bir başka ifade ile rezidü (artık-hata) değeri büyük olan gözlemlere *y*-yönünde sapan değerler denir. Regresyonda, *y*-yönünde sapan değerlere aykırı değer de denir.
- *X*-uzayında veri kümesinden uzakta bulunan noktalar, *X*-yönünde sapan değerler (high leverage points) olarak adlandırılır. *X*-yönündeki sapan değerler ikiye ayrılırlar:
  (i) Kötü kaldıraç (Bad leverage) noktası: Regresyon doğrusunun eğimini çok fazla değiştiren *X*- yönündeki sapan değerlere kötü kaldıraç noktası denir.
  (ii) İyi kaldıraç (good leverage) noktası: Regresyon katsayılarının doğruluğunu arttıran *X*- yönündeki noktalara iyi kaldıraç noktası denir.
- Hem *X*-uzayındaki hem de *y*-uzayındaki veri kümesinden uzakta bulunan noktalara hem bağımlı hem de bağımsız değişkenler yönünde sapan değerler denir.
- Veri kümesindeki diğer gözlemlerle karşılaştırıldığında tek tek ya da hep beraber kestirilmiş regresyon denklemine etki eden gözlemlere etkili gözlemler (influential observations) denir.

Sapan değerlerin OLSE üzerindeki olumsuz etkilerini ortadan kaldırmak amacıyla Regresyon Tanısı (Regression Diagnostic) ve Dayanıklı Regresyon

(Robust Regression) olmak üzere iki temel yaklaşım kullanılır. Regresyon tanısında amaç etkili gözlemleri kesin olarak belirlemek ve sapan değer olarak saptanan gözlemler düzeltildikten ya da veriden çıkarıldıktan sonra OLSE uygulamaktır. Dayanıklı regresyonda ise sapan değerlerden etkilenmeyen dayanıklı tahmin ediciler üretilmeye çalışılır.

2.2.1. Dayanıklı Tahmin Edicilerin Özellikleri

Dayanıklı bir tahmin edicide bulunması gereken temel iki özellik dirençlilik ve sağlam etkinliktir.

Bir tahmin edici az sayıdaki büyük hatadan veya herhangi bir miktardaki yuvarlama-gruplama hatalarından sınırlı miktarda etkileniyorsa dirençlidir denir. Örneklemin küçük bir alt kümesi tahminler üzerinde gereğinden çok bir etkiye sahip değilse bu tahmin edici büyük hatalara karşı dirençlidir yorumu yapılabilir. Eğer tahmin edici küçük hatalara karşı sürekli olarak tepki gösteriyorsa hatta tahminler bu gözlemlerin küçük bir parçasının yuvarlanma ya da gruplanması ile belirlenmiyorsa bu tahmin edici için gruplama-yuvarlama hatalarına karşı dirençlidir denir .

Bir tahmin edicinin varyansı her bir dağılım için minimuma yakın bir değer alıyorsa bu tahmin edici sağlam etkinliğe sahiptir denir.

2.2.2. Dayanıklı M-Tahmin Ediciler

M-Tahmin Edicileri (ME): Dayanıklı tahmin ediciler içerisinde yer alan en geniş tahmin edici sınıflarından birisi de ilk defa Huber (1964) tarafından ortaya atılmış olan M-tahmin edicidir (ME). ME, y- yönündeki sapan değerlere karşı dayanıklıdır. ME'de temel yaklaşım; hataların karelerini minimum yapmak yerine, hataların bir başka fonksiyonu olan $\rho(\varepsilon_i)$ minimize etmeyi amaçlamaktadır ve

$$\min \sum_{i=1}^{n} \rho(\varepsilon_i) = \min \sum_{i=1}^{n} \rho(y_i - x_i^{'}\beta)$$
$$= \min \sum_{i=1}^{n} \rho(y_i - \sum_{i=1}^{n} x_{ij}\beta_j), \qquad j = 1,2,...,p$$

olarak tanımlanır. ρ için aranan özellikler aşağıdaki gibidir :

i) $\rho(0) = 0,$

ii) $\rho(\varepsilon) \geq 0,$

iii) $\rho(\varepsilon) = \rho(-\varepsilon)$ (simetrik),

iv) ρ sürekli ve türevlenebilir,

v) $0 < \varepsilon_1 < \varepsilon_2$ için $\rho(\varepsilon_1) < \rho(\varepsilon_2).$

ME Hesaplanması: ME, ölçek değişmezliği gerektirmediğinden, ME genelde

$$\min \sum_{i=1}^{n} \rho\left(\frac{\varepsilon_i}{s_0}\right) = \min \sum_{i=1}^{n} \rho\left(\frac{y_i - x_i^{'}\beta}{s_0}\right) \qquad (2.6.)$$

denklemi ile çözülür. Burada, $s_0 = \dfrac{median\left|\varepsilon_i - median(\varepsilon_i)\right|}{0.6745}$ ile verilir. (2.6.) ifadesinin minimumunu bulmak için, β'ya göre türevi alınıp sıfıra eşitlendiğinde;

$$\frac{\partial}{\partial\beta} \sum_{i=1}^{n} \rho\left(\frac{y_i - x_i^{'}\beta}{s_0}\right) = 0,$$
$$\sum_{i=1}^{n} \psi\left(\frac{y_i - x_i^{'}\beta}{s_0}\right) x_{ij} = 0, \qquad j = 1,2,...,p \qquad (2.7.)$$

elde edilir. Burada ψ, ρ fonksiyonun türevidir. Eğer ρ fonksiyonu konveks ise (2.6.) ve (2.7.) ifadeleri birbirine denk olur. Aksi takdirde (2.6.) ifadesinin birden

çok çözümü olacağından en iyi tahmini belirlemekte sorun yaşanabilir. (2.7.) de verilen ψ fonksiyonu genelde lineer değildir ve iteratif yöntemlerle çözülmelidir. Literatürde bir çok doğrusal olmayan optimizasyon teknikleri var olmakla birlikte iteratif olarak yeniden ağırlıklandırılmış en küçük kareler (Iteratively Reweighted Least Squares, IRWLS) yöntemi çok yaygın kullanılmaktadır. Bu durumda (2.7.) ile verilen denklem IRWLS yöntemi ile çözülecektir.

(2.7.) ile verilen denklemin IRWLS yöntemi ile çözümü aşağıdaki gibidir:

IRWLS yöntemini kullanmak için, β ve s_0'ın sırasıyla $\hat{\beta}_0$ ve s gibi başlangıç tahminlerinin var olduğu kabul edilir ve (2.7.) denklemi;

$$\sum_{i=1}^{n} \psi\left(\frac{y_i - x_i'\beta}{s_0}\right) x_{ij} = \sum_{i=1}^{n} \frac{x_{ij}\{\psi((y_i - x_i'\beta)/s_0)/(y_i - x_i'\beta/s_0)\}(y_i - x_i'\beta)}{s_0} = 0, \quad (2.8.)$$

şeklinde verilir ve (2.8.) denklemi,

$$\sum_{i=1}^{n} w_{i0}\left(\frac{y_i - x_i'\beta}{s_0}\right) x_{ij} = 0, \quad j = 1, 2, ..., p \quad (2.9.)$$

şeklinde de ifade edilebilir. Burada,

$$w_{i0} = \begin{cases} \dfrac{\psi((y_i - x_i'\hat{\beta}_0)/s_0)}{(y_i - x_i'\hat{\beta}_0)/s_0}, & y_i \neq x_i'\hat{\beta}_0 \\ 1 & y_i = x_i'\hat{\beta}_0 \end{cases}$$

şeklinde verilir (Montgomery ve ark., 2001). Ayrıca, (2.9.) un matrisi gösterimi;

$$X'W_0X\beta = X'W_0y$$

şeklindedir. Burada, W_0 köşegenleri w_{i0} ağırlıklarından oluşan köşegen (diagonal) matristir. Yukarıda cebirsel ifadeler ile verilen IRWLS yönteminin adım adım algoritması aşağıda verilmiştir.

IRWLS Algoritması:

1) Başlangıç tahmin edicisi seçilir (genellikle OLSE),
2) Başlangıç tahmin edicisi kullanılarak rezidüler (residuals) hesaplanır,
3) Bu rezidüler yardımıyla ölçek tahmini hesaplanır,
4) Rezidüler ölçek tahmini yardımıyla standartlaştırılır,
5) Standartlaştırılmış rezidüler ve ME' nin ağırlık fonksiyonu yardımıyla ağırlıklar bulunur. Ağırlıklar yardımıyla ağırlıklı en küçük kareler tekniği kullanılarak regresyon katsayıları tahmin edilir.

$$\hat{\beta}_M = (X'WX)^{-1}X'Wy \qquad (2.10.)$$

W köşegen elemanları ağırlık (w) olmak üzere köşegen matristir.

6) Ağırlıklı en küçük kareler ile elde edilen regresyon katsayıları ile bir önceki aşamada elde edilen katsayılar arasındaki fark çok küçük ise işlem sonlandırılır. Değilse bulunan katsayılarla 2. Adıma dönülerek yakınsama sağlanana kadar işlemlere devam edilir.

ME hesaplanması için Huber'in (1981) önerdiği ρ fonksiyonu aşağıdaki gibi tanımlanır. Diğer çalışmalar tablo olarak verilmiştir.

Huber'in ρ-fonksiyonu: Huber (1981) tarafından standartlaştırılmış hata terimleri için önerilen fonksiyon,

$$\rho(r)=\begin{cases}\dfrac{r^2}{2}, & -c\le r\le c\\ c|r|-\dfrac{c^2}{2}, & r<-c \text{ veya } c<r\end{cases} \qquad (2.11.)$$

olarak tanımlanır. Burada $r=\dfrac{e}{s}$, $s=\dfrac{median\left\{\left|e_i-median\left(e_i\right)\right|\right\}}{0.6745}$ ve $c=1.345$ dir. (2.10.) eşitliğinin *r*' ye göre türevi alındığında,

$$\psi(r)=\begin{cases}-c, & r<-c\\ r, & |r|\le c\\ c, & r>c\end{cases} \qquad (2.12.)$$

fonksiyonu elde edilir. $\psi(r)$ fonksiyonunun *r*' ye bölünmesiyle

$$w=\begin{cases}1, & |r|\le c\\ \dfrac{c}{r}, & |r|>c\end{cases} \qquad (2.13.)$$

ağırlık fonksiyonu elde edilir . ME için kullanılan bazı ψ-fonksiyonları Çizelge 1 de verilmiştir .

ME'nin asimptotik kovaryansı: Kabul edelim ki ME, $\hat{\beta}_M \sim N(\beta, A^2(X'X)^{-1})$ olsun. Pratikte bu varsayım; $n^{1/2}(\hat{\beta}_M-\beta)\to N(0,A^2(X'X)^{-1})$ olduğundan dolayı yaklaşık olarak sağlanır. Burada;

$$A^2 = s_0^2 E\left[\psi^2(\varepsilon / s_0)\right] / \left[E\psi'(\varepsilon / s_0)\right]^2, \quad (2.14.)$$

ve s_0^2 hataların ölçeğidir . Burada bilinmeyen değerler yerine tahminleri yazıldığında $\hat{\beta}_M$ için asimptotik kovaryansın tahmini;

$$\hat{COV}(\hat{\beta}_M) = \hat{\Omega} = \hat{A}^2 (X'X)^{-1}, \quad (2.15.)$$

$$\hat{A}^2 = s^2 (n-p)^{-1} \sum_{i=1}^{n} \left[\psi\left(e_i / s\right)\right]^2 / \sum_{i=1}^{n} \left[\frac{1}{n}\psi'(\varepsilon / s)\right]^2 \quad (2.16.)$$

şeklindedir (Huber, 1981).

Çizelge 2. 1. Uygulamada ME için en çok kullanılan ψ -fonksiyonları

Fonksiyonun Adı	$\psi(r)$**-Fonksiyonu**	**Düzeltme Sabiti**
Andrews	$\psi(r) = \begin{cases} \sin(r/c) & r \le c\pi \\ 0 & r > c\pi \end{cases}$	$c = 1.399$
Cauchy	$\psi(r) = \dfrac{r}{1 + (r/c)^2} \quad \lvert r\rvert < \infty$	$c = 2.385$
Fair	$\psi(r) = \dfrac{r}{1 + \lvert r\rvert/c} \quad \lvert r\rvert < \infty$	$c = 1.4$
Hampel	$\psi(r) = \begin{cases} r & \lvert r\rvert < a \\ a\,sign(r) & a < \lvert r\rvert \le b \\ \dfrac{a\,sign(r)(c - \lvert r\rvert)}{c - b} & b < \lvert r\rvert \le c \\ 0 & \lvert r\rvert > c \end{cases}$	$a = 1.7$ $b = 3.4$ $c = 8.5$
Huber	$\psi(r) = \begin{cases} r & \lvert r\rvert \le c \\ c\,sign(r) & \lvert r\rvert > c \end{cases}$	$c = 1.385$

Logistics	$\psi(r) = ctanh(r/c) \quad \lvert r\rvert < \infty$	$c = 1.205$
Welsch	$\psi(r) = rexp(-(r/c)^2) \quad \lvert r\rvert < \infty$	$c = 2.985$
Talwar	$\psi(r) = \begin{cases} r & \lvert r\rvert \le c \\ 0 & \lvert r\rvert > c \end{cases}$	$c = 2.795$
Tukey (Bisquare)	$\psi(r) = \begin{cases} r\left[1-\left(\frac{r}{c}\right)^2\right]^2 & \lvert r\rvert \le c \\ 0 & \lvert r\rvert > c \end{cases}$	$c = 4.685$

2.3. Model Parametrelerinin Tahmini

2.3.1. En Küçük Kareler Yöntemi

EKK tekniğinin gerek değişkenler gerekse hata terimine ait bir takım varsayımları bulunmaktadır. Bunlar temel varsayımlar ve ek varsayımlar olarak iki grupta incelenebilir.

Temel varsayımlar,

(i) X , $n \times p$ tipinde $p < n$ ile stokastik olmayan bir matristir.

(ii) X matrisinin rankı p 'dir yani, tam kolon ranklıdır.

(iii) $n \times 1$ tipindeki y vektörü, elemanları gözlemlenebilen rasgele vektördür.

(iv) $n \times 1$ tipindeki ε vektörünün elemanları gözlenemeyen rasgele değişkenlerdir öyle ki, $E(\varepsilon) = 0$ ve $Cov(\varepsilon) = \sigma^2 I_n$ dir.

Ek varsayım,

(v) ε vektörü n değişkenli normal dağılır.

Temel varsayımlara ilaveten ek varsayım da sağlanıyorsa bu model "Klasik Lineer Regresyon Modeli" olarak adlandırılır.Bu varsayımlara ek olarak tüm

gözlemlerin tahmin edilmiş regresyon doğrusu üzerinde aynı etkiye sahip olması gerekmektedir. Yani veri kümesinde sapan değer olmamalıdır (Huber, 1981).

Bilinmeyen parametre vektörü β'nın EKK tahmin edicisi hata kareler toplamı minimum yapılarak bulunur. $\varepsilon = y - X\beta$ olmak üzere hata kareler toplamı;

$$\begin{aligned}\sum_{i=1}^{n} \varepsilon_i^2 &= \varepsilon'\varepsilon \\ &= (y - X\beta)'(y - X\beta) \\ &= y'y - y'X\beta - \beta'X'y + \beta'X'X\beta\end{aligned}$$

olarak elde edilir. Buna göre hata kareler toplamı;

$$\varepsilon'\varepsilon = y'y - 2\beta'X'y + \beta'X'X\beta$$

olur. O halde

$$\frac{\partial(\varepsilon'\varepsilon)}{\partial\beta} = -2X'y + 2X'X\beta = 0$$

$$X'X\beta = X'y$$

normal denklemi elde edilir. Buradan β'nın EKK tahmini edicisi

$$\hat{\beta} = (X'X)^{-1}X'y \qquad (2.17.)$$

olarak elde edilir. $\hat{\beta}$ 'nın ortalaması ve kovaryansı aşağıda verilmiştir:

$$E(\hat{\beta}) = \beta$$

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

EKK tahmin edicisi Gaus Markov teoremi gereğince en iyi lineer yansız tahmin edicidir (BLUE).

2.3.2. Ridge Tahmin Edici

Hoerl ve Kennard'ın (1970) önerdiği ridge tahmin edici (ORRE),

$$\hat{\beta}(k) = (X'X + kI)^{-1} X'y, \quad k \geq 0 \tag{2.18.}$$

şeklinde tanımlanır. Ridge tahmin edici k yanlılık parametresine bağlıdır ve k nın seçimi tahmin edicinin performansını etkiler. Ridge tahmin edici $k = 0$ için EKK tahmin edicisini vermektedir. Ayrıca ridge tahmin edici EKK tahmin edicinin bir lineer dönüşümü olarak da yazılabilmektedir:

$$\begin{aligned} \hat{\beta}(k) &= (X'X + kI)^{-1} X'X\hat{\beta}, \\ &= W(k)\hat{\beta}, \end{aligned} \tag{2.19.}$$

Burada, $W(k) = (X'X + kI)^{-1} X'X$.

Hoerl ve Kennard (1970) ridge tahmin edicinin toplam varyansının k nın sürekli, monoton azalan bir fonksiyonu ve yanlılığın karesinin k nın sürekli,

monoton artan bir fonksiyonu olduğunu göstermiştir. Bu nedenle varyanstaki azalma yanlılığın karesindeki artıştan fazla olduğu sürece ridge tahmin edicinin iyi bir teknik olduğu söylenebilir . $S_k = X'X + kI$ olsun. Bu durumda,

$$E(\hat{\beta}(k)) = S_k^{-1}X'X\beta,\ Bias(\hat{\beta}(k),\beta) = -kS_k^{-1}\beta,\ Var(\hat{\beta}(k)) = \sigma^2 S_k^{-1}X'XS_k^{-1}$$

olduğundan ridge tahmin edicinin matris hata kareler ortalaması

$$MMSE(\hat{\beta}(k),\beta) = S_k^{-1}(\sigma^2 X'X + k^2\beta\beta')S_k^{-1}$$

olur. $\lambda_1,\ldots,\lambda_p$, $X'X$ in özdeğerleri olmak üzere skaler hata kareler ortalaması

$$SMSE(\hat{\beta}(k),\beta) = tr(MMSE(\hat{\beta}(k),\beta)) = \sum_{i=1}^{p}\frac{\sigma^2\lambda_i + k^2\beta_i^2}{(\lambda_i + k)^2}$$

olarak elde edilir. Ayrıca, $\hat{\beta}(k)$ nın MSE'si $\hat{\beta}$ dan daha iyi olacak şekilde her zaman bir $k > 0$ vardır.

k Yanlılık Parametresinin Seçimi

Ridge regresyonda k nın seçimi için pek çok yöntem geliştirilmiştir. Bunlardan bazıları şu şekildedir:

1. Hoerl ve Kennard (1970) ridge tahmin edicinin EKK'dan daha küçük MSE'ye sahip yanlılık parametresinin tahminini $\hat{k}_{HK} = \dfrac{\hat{\sigma}^2}{\hat{\alpha}^2_{maks}}$ olarak vermiştir. Burada $\hat{\alpha}_{maks}$, kanonik formda regresyon katsayısının tahmininin maksimum değeridir.

2. Theobald (1974) her W için $E\left[(\hat{\beta}(k)-\beta)'W(\hat{\beta}(k)-\beta)\right]$, $E\left[(\hat{\beta}-\beta)'W(\hat{\beta}-\beta)\right]$ dan daha küçük olacak şekildeki yanlılık parametresinin tahminini $\hat{k}_T = \frac{2\hat{\sigma}^2}{\hat{\beta}'\hat{\beta}}$ olarak vermiştir.

3. Hoerl, Kennard ve Baldwin (1975) ridge tahmin edicinin MSE değerini minimum yapan $k_i = \frac{\sigma^2}{\alpha_i^2}$ değerlerinin harmonik ortalaması $\hat{k}_{HKB} = \frac{p\hat{\sigma}^2}{\hat{\beta}'\hat{\beta}}$ yı önermiştir.

4. Lawless ve Wang (1976) bayesian yaklaşımla $\hat{k}_{LW} = \frac{p\hat{\sigma}^2}{\hat{\beta}'X'X\hat{\beta}}$ tahminini vermişlerdir.

2.3.3. Liu Tahmin Edici

Çoklu iç ilişki olması durumunda OLSE nin dezavantajları bilinmektedir. Bu dezavantajları gidermek için pek çok tahmin edici önerilmiştir. Ridge ve Stein ($\hat{\beta}_s = c\hat{\beta}$, $0<c<1$) tahmin ediciler bunlardan ikisidir. Fakat her iki tahmin edicinin de olumsuz yanları vardır.

$\hat{\beta}(k)$, k nın karmaşık bir fonksiyonudur. Bu nedenle k nın seçimi için önerilen metotlarda karmaşık denklemlerle karşılaşabiliriz. Ridge tahmin metodu $X'X$ matrisinin aksine $X'X+kI$, $k\geq 0$ matrisine bağlı olduğundan EKK deki zorluklar önlenmiş olunur. $X'X+kI$ nın koşul sayısı k nın azalan bir fonksiyonu olduğundan yeteri kadar büyük k için $X'X+kI$ nın koşul sayısı küçük seviyeye düşürülebilir. Fakat uygulamada k nın küçük olması önerilir. Bu

nedenle $X'X + kI$ nın koşul sayısını küçültebilecek bir k bulunamayabilir. Seçilen k değerine göre hala $X'X + kI$ çoklu iç ilişki problemine sahip olabilir.

Stein tahmin edicinin avantajı c nin bir lineer fonksiyonu olmasıdır. Fakat $\hat{\beta}_s$ nin her elemanının büzülmesi aynıdır. Bu da uygulamada iyi sonuç vermez.

İki tahmin edicinin kombinasyonunun bu iki tahmin edicinin avantajlarını birleştireceği düşüncesi Liu'yu (1993) yeni bir tahmin edici önermeye sevk etmiştir. Çoklu iç ilişki problemini ortadan kaldırmak için Liu (1993), Stein (1956) tahmin edici ile ridge tahmin ediciyi birleştirmiş ve

$$\hat{\beta}(d) = \left(X'X + I\right)^{-1}\left(X'y + d\hat{\beta}\right), \quad 0 < d < 1$$
$$= \left(X'X + I\right)^{-1}\left(X'X + dI\right)\hat{\beta} = F(d)\hat{\beta} \qquad (2.20.)$$

tahmin edicisini önermiştir. Burada $F(d) = \left(X'X + I\right)^{-1}\left(X'X + dI\right)$. Bu tahmin edici Akdeniz ve Kaçıranlar (1995) ve Gruber (1998) tarafından Liu tahmin edici olarak adlandırılmış. $\hat{\beta}(d)$ nın $\hat{\beta}(k)$ üzerindeki avantajı d nin bir lineer fonksiyonu olmasıdır. Dolayısıyla d nin seçimi k nın seçiminden daha kolaydır.

Liu tahmin edicinin bazı özellikleri ise şu şekildedir:

1. $d = 1$ iken $\hat{\beta}(d) = \hat{\beta}$ dır.
2. $\left\|\hat{\beta}(d)\right\| < \left\|\hat{\beta}\right\|$ dır.
3. $E(\hat{\beta}(d)) = (X'X + I)^{-1}(X'X + dI)\beta$ olup yanlı bir tahmin edicidir.
4. $\hat{\beta}(d)$ nın MSE'si $\hat{\beta}$ dan daha iyi olacak şekilde bir $0 < d < 1$ vardır.

5. Kanonik formda

$$SMSE(\hat{\beta}(d),\beta)=\sigma^2\sum_{i=1}^{p}\frac{(\lambda_i+d)^2}{\lambda_i(\lambda_i+1)}+(d-1)^2\sum_{i=1}^{p}\frac{\alpha_i^2}{(\lambda_i+1)^2}$$

olur. $SMSE(\hat{\beta}(d),\beta)$ yi minimum yapan d yanlılık parametresinin tahminini

$$\hat{d}_{mm}=1-\hat{\sigma}^2\left(\sum_{i=1}^{p}\frac{1}{\lambda_i(\lambda_i+1)}\Big/\sum_{i=1}^{p}\frac{\hat{\alpha}_i^2}{(\lambda_i+1)^2}\right) \qquad (2.21.)$$

olarak Liu (1993)'de verilmiştir. Bu makalede ayrıca ridge tahmin edicide k seçme yöntemlerine benzer şekilde diğer d yanlılık parametresi seçimleri de verilmiştir.

2.3.4. Dayanıklı Ridge Tahmin Edici

Eşitlik (2.1.) ile verilen çoklu lineer regresyon modeli kanonik formda

$$y=Za+\varepsilon, \qquad (2.22.)$$

şeklinde verilir. Burada $Z=XT$, $a=T'\beta$ ve T de kolonları $X'X$ matrisinin özvektörlerinden oluşan ortogonal bir matris olup;

$$Z'Z=T'X'XT=\Lambda=diag(\lambda_1,\lambda_2,...,\lambda_p) \qquad (2.23.)$$

ile elde edilir. Burada $\lambda_1\geq\lambda_2\geq...\geq\lambda_p>0$, $X'X$ matrisinin özdeğerleridir. (2.22.) denklemi için OLSE;

$$\hat{a} = \Lambda^{-1} Z' y \tag{2.24.}$$

ile verilir. Benzer şekilde ridge tahmin edicisi

$$\hat{a}(k) = (\Lambda + kI)^{-1} \Lambda \hat{a} = T_k \hat{a} \tag{2.25.}$$

ile verilir. Burada, $T_k = (\Lambda + kI)^{-1}$ dır.

RE, y-yönündeki sapanlara karşı hassas olduğundan Silvapulle (1991) ME'ye dayalı Ridge M- tahmin edici (RME);

$$\hat{a}_M(k) = T_k \hat{a}_M, \qquad k > 0 \tag{2.26.}$$

ile önermiştir. Burada; $\hat{a}_M$, (2.10) daki ME'nin kanonik formudur. Ayrıca, Silvapulle (1991) RME'nin RE ve ME'den üstünlüklerini MSE kriterine göre incelemiş ve RME için k yanlılık parametresini şu şekilde önermiştir:

$$\hat{k}_1 = \frac{pA^2}{\hat{a}M' \hat{a}M} \tag{2.27.}$$

$$\hat{k}_2 = \frac{pA^2}{\hat{a}M' \Lambda \hat{a}M} \tag{2.28.}$$

2.3.5. Dayanıklı Jacknifed Ridge Tahmin Edici

Jadhav ve Kashid'in (2014) önerdiği dayanıklı jacknifed ridge tahmin edici,

$$\hat{\beta}_{JRM}(k)=\left[I-k^2\varphi'(X'X+kI)^{-2}\varphi\right]^{-2}\hat{\beta}_M=R\hat{\beta}_M,\quad k\geq 0 \qquad (2.29.)$$

olarak tanımlanmıştır. Burada, $R=\left[I-k^2\varphi'(X'X+kI)^{-2}\varphi\right]^{-2}$, φ, $X'X$ matrisinin özvektörlerine karşılık gelen matris ve $\hat{\beta}_M$ bilinmeyen β parametresi için Huber'in M-tahmin edicisidir. Bu tahmin edici veride hem çoklu iç ilişki hemde sapan değerler olması durumunda ortaya çıkan problemi çözmek için tanımlanmıştır. Bu tahmin edicinin performansı, MSE anlamında, OLSE, M-tahmin edici ve ORRE ile karşılaştırılmış ve daha iyi MSE'ye sahip olduğu gösterilmiştir. Jadhav ve Kashid (2014) yanlılık parametresi k için

$$\tilde{k}=\frac{ps^2}{\hat{\beta}_M'\hat{\beta}_M} \qquad (2.30.)$$

önermişlerdir. Burada p açıklayıcı değişken sayısı, $s=1.4826median\left|e_i-median(e_i)\right|$ ve e_i, M-Tahmin edicinin kullanılmasıyla bulunan i-inci rezidüdür.

2.3.6. Dayanıklı Liu Tahmin Edici

Arslan ve Billor'un (2000) önerdiği dayanıklı Liu tahmin edici,

$$\hat{\beta}_{LM}(d)=\left[(X'X+I)^{-1}(X'X+dI)\right]\hat{\beta}_M,\ 0<d<1, \qquad (2.31.)$$
$$=F(d)\hat{\beta}_M,$$

olarak tanımlanmıştır. Burada $\hat{\beta}_M$ bilinmeyen β parametresi için Huber'in M-tahmin edicisidir. Bu tahmin edici de dayanıklı ridge tahmin ediciye benzer şekilde

veride hem çoklu iç ilişki hemde sapan değerler olması durumunda ortaya çıkan problemi çözmek için tanımlanmıştır. (2.22.) kanonik formun kullanılmasıyla bu tahmin edici

$$\hat{\alpha}_{LM}(d) = \left[(\Lambda + I)^{-1}(\Lambda + dI)\right]\hat{\alpha}_M,\ 0 < d < 1, \tag{2.32.}$$

şeklinde verilir. Burada Λ, $X'X$ matrisinin özdeğerlerine karşılık gelen köşegen bir matristir. Arslan ve Billor (2000) yanlılık parametresi d için

$$\hat{d}_M = 1 - \hat{A}^2\left[\sum_{i=1}^{p}\frac{1}{\lambda_i(\lambda_i+1)} \Big/ \sum_{i=1}^{p}\frac{\hat{\alpha}_{Mi}^2}{(\lambda_i+1)^2}\right] \tag{2.33.}$$

önermişlerdir. Detaylar makaleden görülebilir. Bu tahmin edicinin performansı, MSE anlamında, M-tahmin edici ve Liu tahmin edici ile karşılaştırılmış ve yanlılık parametresinin seçimine bağlı olarak daha iyi MSE'ye sahip olduğu gösterilmiştir.

2.3.7. Dayanıklı Genelleştirilmiş Liu ve Hemen Hemen Yansız Genelleştirilmiş Liu Tahmin Edici

Liu (1993) LE'nin genelleştirilmiş formu olan, genelleştirilmiş Liu tahmin edicisini kanonik formda (GLE):

$$\begin{aligned}\hat{a}_L &= (\Lambda + I)^{-1}(X'y + D\hat{a}) = (\Lambda + I)^{-1}(\Lambda + DI)\hat{a} \\ &= \left[I - (\Lambda + I)^{-1}(I - D)\right]\hat{a},\end{aligned} \tag{2.34.}$$

tanımlamıştır. Burada; $D = diag(d_1, d_2, ..., d_p)$, $0 < d_i < 1$, $i = 1, 2, ..., p$. Aynı zamanda Liu (1993) GLE'nin MSE'sini minimum yapan d_i'yi şu şekilde önermiştir:

$$d_i = \frac{a_i^2 - \sigma^2}{a_i^2 + (\sigma^2 / \lambda_i)}, \; i = 1, 2, ..., p \tag{2.35.}$$

elde edilir . Burada a_i^2 ve σ^2 bilinmeyen parametrelerinin yerlerine sırasıyla $\hat{a}_i^2 - \frac{\hat{\sigma}^2}{\lambda_i}$ ve $\hat{\sigma}^2$ yansız tahminleri yazıldığında

$$\hat{d}_i = \frac{\lambda_i (\hat{a}_i^2 - \hat{\sigma}^2)}{\lambda_i \hat{a}_i^2 + \hat{\sigma}^2}, \; i = 1, 2, ..., p \tag{2.36.}$$

elde edilir (Liu, 1993 ve Akdeniz ve Kaçıranlar, 1995).

Diğer taraftan, Akdeniz ve Kaçıranlar (1995) GLE'nin yanlılığını azaltmak için Hemen Hemen Yansız Genelleştirilmiş Liu Tahmin Ediciyi (AUGLE) şu şekilde tanımlamışlardır:

$$\begin{aligned} a_L^* &= \left[I + (\Lambda + I)^{-1} (I - D) \right] \hat{a}_L \\ &= \left[I - (\Lambda + I)^{-2} (I - D)^2 \right] \hat{a} \end{aligned} \tag{2.37.}$$

Burada, $D = diag(d_1, d_2, ..., d_p)$, $0 < d_i < 1$, $i = 1, 2, ..., p$ 'dır. Aynı zamanda Akdeniz ve Kaçıranlar (1995) AUGLE'nin, GLE ve OLSE'den üstünlüklerini MSE kriterine göre incelemiş olup, AUGLE'nin MSE'sini minimum yapan d_i'yi şu şekilde önermişlerdir:

$$MSE\left(a_L^*\right)=\sigma^2\sum_{i=1}^{p}\frac{1}{\lambda_i}\left(1-\left(\frac{1-d_i}{\lambda_i+1}\right)^2\right)^2+\sum_{i=1}^{p}\left(\frac{1-d_i}{\lambda_i+1}\right)^4 a_i^2 \qquad (2.38.)$$

Eşitlik (2.38.)'in d_i 'ye göre türevinin alınıp sıfıra eşitlenmesiyle;

$$d_i^*=1-\frac{\sigma(\lambda_i+1)}{\sqrt{\lambda_i a_i^2+\sigma^2}}, \qquad (2.39.)$$

ifadesi elde edilir. Burada bilinmeyen a_i^2 ve σ^2 parametrelerinin yerlerine sırasıyla $\hat{a}_i^2$ ve $\hat{\sigma}^2$ tahminleri yazıldığında;

$$d_i^*=1-\frac{\hat{\sigma}(\lambda_i+1)}{\sqrt{\lambda_i\,\hat{a}_i^2+\hat{\sigma}^2}}, \quad i=1,2,...,p, \qquad (2.40.)$$

elde edilir (Akdeniz ve Kaçıranlar, 1995).

Eşitlik (2.34.) ve (2.37.) ile verilen tahmin ediciler *y*-yönündeki sapan değerlere karşı hassas olduğundan sırasıyla ME'ye dayalı Genelleştirilmiş Liu-M (GLME) ve Hemen Hemen Yansız Genelleştirilmiş Liu-M tahmin edicileri (AUGLME) aşağıdaki gibi verilebilir.

$$\hat{a}_{ML}=\left[I-(\Lambda+I)^{-1}(I-D)\right]\hat{a}_M, \qquad (2.41.)$$

$$\tilde{a}_{ML}^*=\left[I-(\Lambda+I)^{-2}(I-D)^2\right]\hat{a}_M \qquad (2.42.)$$

GLME ve AUGLME tahmin edicileri için yanlılık parametre tahminleri sırasıyla aşağıdaki gibi verilmiştir.

$$\hat{d}_{mi} = \frac{\lambda_i(\hat{a}_{Mi}^2 - \hat{A}^2)}{\lambda_i \hat{a}_{Mi}^2 + \hat{A}^2}, \tag{2.43.}$$

$$\hat{d}_{mi}^{*} = 1 - \frac{A(\lambda_i + 1)}{\sqrt{\lambda_i \hat{a}_{Mi}^2 + \hat{A}^2}}, \tag{2.44.}$$

Ayrıca, burada Akdeniz ve Kaçıranlar'a (1995) benzer şekilde AUGLME ve GLME'nin MSE'sini minimum yapan d_i'ler kullanılarak AUGLME nin parametre tahmininin GLME'nin parametre tahminine eşit olduğu görülmektedir (Ertaş, 2015).

3. DEĞİŞKEN SEÇİMİ VE MODEL OLUŞTURMA

Regresyon için anlamlı olan açıklayıcı değişkenlerin uygun bir alt kümesini belirleme işlemine "değişken seçimi" denir. Burada iki amaç sözkonusudur.

1. Model mümkün olan en fazla değişkeni içermelidir ki bilgi kaybı olmasın ve $\hat{y}$ değerleri olumsuz etkilenmesin.
2. Model mümkün olan en az değişkeni içermelidir ki $\hat{y}$ nın varyansı fazla büyümesin. Ayrıca gereksiz değişkenler veri toplama maliyet ve süresinin artmasına neden olur.

En iyinin tek bir açıklaması olmayıp, çoğu kez aday değişkenlerin farklı alt kümeleri en iyi olarak belirtilebilir.

En iyi regresyon modelinin belirlenmesinde yaygın olarak çoklu belirleyicilik katsayısı R^2 veya düzeltilmiş çoklu belirleyicilik katsayısı $\overline{R}^2$, hata kareleri ortalaması (MSE), öngörü hata kareler ortalaması (MSEP) gibi kriterler kullanılır. Eşit sayıda açıklayıcı değişken içeren modellerin karşılaştırılmasında çoklu belirleyicilik katsayısı R^2 ve farklı sayıda açıklayıcı değişken içeren modellerin karşılaştırılmasında $\overline{R}^2$ kullanılır. En iyi regresyon modelinin belirlenmesinde R^2 veya $\overline{R}^2$ nın yüksek , MSE'sinin düşük ve az sayıda açıklayıcı değişken içeren modeller tercih edilir (Montgomery ve ark., 2001).

Ayrıca, Mallows (1973)'un C_p kriteri, Golub ve arkadaşları (1979) tarafından tanımlanan genelleştirilmiş çapraz geçerlilik kriteri (GCV), Allen (1971) tarafından tanımlanmış öngörü hata kareler toplamı (PRESS) kriteri, Akaike (1974) bilgi kriteri (AIC), Hurvich ve Tsai (1989)'un tanımladığı düzeltilmiş Akaike bilgi

kriteri (AIC_c), Schwarz (1978) tarafından tanımlanmış bayesyen bilgi kriteri (BIC) gibi bir çok kriter kullanılmaktadır (Draper ve Smith, 2003, Miller, 1990, Montgomery ve ark., 2001, Lindsey ve sheather, 2010, Delaney ve Chatterjee, 1986, Linhart ve Zucchini,1986).

Bazen mevcut açıklayıcı değişkenlerden bazılarının bağımlı değişkendeki toplam değişimi açıklamada etkileri olmayabilir. Bu durumda bu değişkenler modelden çıkarılabilir. Bazen de mevcut açıklayıcı değişkenlerden yanıt değişkendeki toplam değişimi açıklamada yetersiz kalabilir bu durumda da değişkenler eklenebilir. Açıklayıcı değişken sayısının artması durumunda ileriye doğru seçim ya da geriye doğru ayıklama gibi adımsal seçim yöntemleri uygulanabilir. Değişken sayısının çok fazla olması durumunda adımsal seçim yöntemleri kullanılamaz. Bu durumda genetik algoritma ve/veya bilgi kriterleri kullanılarak model seçimi yapılabilir (Chatterjee ve ark., 2000, Miller, 1990, Bozdoğan, 2003).

Çoklu lineer regresyon modeli oluşturulurken genellikle çoklu iç ilişki, aykırı ya da sapan değerlerin, etkili gözlemlerin bulunmadığı varsayımı ile çalışılır. Ancak özellikle uygulamalı alanlarda bu varsayımlar bozulur. Model seçiminden önce bunların kontrol edilmesi ve mevcut duruma göre model seçim yönteminin belirlenmesi gerekmektedir. Açıklayıcı değişkenlerin yapısı, hata ya da rezidü analiziyle bu problemlerin var olup olmadığı tespit edilmelidir.

Modelin yanlış belirlenmesi durumunda elde edilen EKK tahmin edici yanlı olabilir. Bu durumda alt küme modellerinden ve tam modelden bulunan parametre tahminlerinin doğruluğunu MSE'ye göre karşılaştırmak daha uygun olacaktır. Benzer şekilde alt küme modelinden bulunan varyansın tahmin edicisi ve bağımlı değişkenin tahmini de yanlı olacaktır.

Değişken seçimi problemi iki aşamada ele alınır. İlk aşamada alt küme modelleri oluşturulur. İkinci aşamada modeller karşılaştırılarak hangilerinin daha

iyi olduğuna karar verilir. Şimdi karşılaştırma kriterlerini ve hesaplama yöntemlerini sırasıyla ele alacağız.

3.1. Değişken Seçme Kriterleri

3.1.1. Çoklu Belirleyicilik Katsayısı

Regresyon modelinin uygunluğun bir ölçütü olarak yaygın bir şekilde kullanılan ölçü çoklu belirleyicilik katsayısı olup R^2 ile gösterilir. R_p^2 , p terimli , $p-1$ açıklayıcı değişkenli ve sabit terimli bir regresyon denklemi için çoklu belirleyicilik katsayısını göstersin. Bu durumda,

$$\text{SST}=\sum_{i=1}^{n}(\hat{y}_i-\overline{y})^2 = y'y-n\overline{y}^2 = (\hat{\beta}'X'y-n\overline{y}^2)+(y'y-\hat{\beta}'X'y) = \text{SSR+RSS}$$

olmak üzere,

$$R^2=\frac{\sum_{i=1}^{n}(\hat{y}_i-\overline{y})^2}{\sum_{i=1}^{n}(y_i-\overline{y})^2}=\frac{SSR}{SST}=1-\frac{RSS}{SST}=\frac{\hat{\beta}'X'y-n\overline{y}^2}{y'y-n\overline{y}^2}, \qquad (3.1.)$$

$$R_p^2=\frac{SSR_p}{SST}=1-\frac{RSS_p}{SST}=1-\frac{SSE_p}{S_{yy}}, \qquad (3.2.)$$

şeklinde tanımlanır. Burada, SSR, $RSS(SSE)$ ve $SST(S_{yy})$ sırasıyla tam modelde regresyon kareler toplamını, rezidü kareler toplamını ve genel kareler toplamını göstermektedir. SSR_p ve $RSS_p(SSE_p)$ sırasıyla p terimli regresyon kareler toplamını ve rezidü kareler toplamını göstermektedir. p artarken R_p^2

değeri de artar. Bir alt küme regresyon modeli için R_p^2 nin optimum değerini değil beklentileri karşılayan bir değeri aranmalıdır. Genel olarak modele dahil edilecek değişken sayısına karar vermek için R^2 nin kullanılması doğru olmaz. R_p^2 alt küme modellerini karşılaştırmada kullanılır ve R_p^2 değeri büyük olan modeller tercih edilir (Montgomery ve ark., 2001). En iyi modelin seçilmesinde yardımcı olan R^2 ye bağlı kural ve algoritmalar vardır. R^2 istatistiği tek başına aday modeller içinden en iyi öngörü modelinin seçim kriteri olarak tavsiye edilmez (Myers,1990).

Düzeltilmiş Çoklu Belirleyicilik Katsayısı:

Model seçiminde R^2 nin kullanımı tehlikeli olabilir. Çünkü yeni bir değişkenin eklenmesinin R^2 nin değerinin artmasına ya da en azından azalmamasına neden olduğunu biliyoruz. Bu nedenle R^2 istatistiğinde yer alan $RSS_p(SSE_p)$ ve SST değerleri yerine, ortalamalarının yerleştirilmesiyle düzeltilmiş R^2 aşağıdaki gibi tanımlanmıştır.

$$\bar{R}^2 = 1 - \frac{RSS/n-p}{SST/n-1} = 1 - \frac{MSE}{SST/n-1} = 1 - \frac{s^2(n-1)}{SST} \qquad (3.3.)$$

Burada,

$$s^2 = \frac{(y-\hat{y})'(y-\hat{y})}{n-p} = \sum_{i=1}^{n} \frac{(y_i-\hat{y}_i)^2}{n-p}$$

model rezidü kareler ortalamasıdır (MSE). Düzeltilmiş R^2 , p terimli bir regresyon denklemi için

$$\bar{R}_p^2 = 1-(\frac{n-1}{n-p})(1-R_p^2)$$

şeklinde de yazılabilir. $\bar{R}_p^2$ istatistiği, R_p^2 istatistiğine göre daha tutarlıdır. $\bar{R}_p^2$ modele yeni eklenen değişkenlerden fazla etkilenmez. Model seçiminde maksimum $\bar{R}_p^2$ değerine sahip olan model seçilir (Montgomery ve ark., 2001) Bununla birlikte bu kritere denk olan bir başka kriter aşağıda verilmiştir.

3.1.2. Artık (Hata) Kareler Ortalaması

Alt küme regresyon modeli için artık kareler ortalaması,

$$MSE_p = \frac{RSS_p}{n-p} \text{ yada diğer gösterimle } MSE_p = \frac{SSE_p}{n-p} \qquad (3.4.)$$

şeklinde tanımlanır. p değeri artarken MSE_p değeri azalarak gider ve daha sonra çok az artar. MSE_p deki artış, modele yeni değişken eklendiğinde SSE_p deki azalmadan kaynaklanmaktadır. Aralarındaki ilişki,

$$\bar{R}_p^2 = 1-(\frac{n-1}{n-p})(1-R_p^2) = 1-\frac{n-1}{n-p}\frac{SSE_p}{S_{yy}}$$

$$= 1-\frac{n-1}{S_{yy}}\frac{SSE_p}{n-p} = 1-\frac{n-1}{S_{yy}}MSE_p$$

şeklinde gösterilebilir. Böylece MSE_p'yi minimum yapmak ile $\overline{R}_p^2$'yi maksimum yapmak birbirine denktir.

3.1.3. Akaike Bilgi Kriteri (AIC)

AIC, uydurulan aday model ve dataya uygun model arasındaki Kullback-Leibler bilginin beklenen değerini tahmin etmek için Akaike (1974) tarafından tanımlanmıştır. Bu kriter ters yönde çalışır. AIC ne kadar küçük ise model o kadar arzu edilebilir bir model olmaktadır. Modelin açıklama gücü, normallik varsayımı altında, hata varyansı ve açıklayıcı değişkenlerin tahmin edilmiş katsayılarının log-likelihood'unun maksimum edilmesiyle ölçülmektedir. Yani;

$$AIC = 2\left\{-\log L(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, ..., \hat{\beta}_p, \hat{\sigma}^2) + k + 2\right\} \tag{3.5.}$$

burada k açıklayıcı değişken sayısıdır. Normal dağılım olabilirlik fonksiyonundan yararlanarak ve gerekli işlemlerin yapılmasıyla,

$$AIC = n\log\frac{RSS}{n} + 2k + n + n\log(2\pi) \tag{3.6.}$$

bulunur. AIC, $AIC = n(\log\hat{\sigma}^2 + 1) + 2p$ şeklinde de verilebilir. Burada p ve $\hat{\sigma}^2$ sırasıyla parametre sayısı ve altküme modellerin varyansıdır.

3.1.4. Düzeltilmiş Akaike Bilgi Kriteri

Hurvich ve Tsai (1989) AIC'in yanlılık-düzeltilmiş versiyonunu geliştirdiler ve AIC_C ile gösterdiler ve aşağıdaki gibi tanımladılar.

$$AIC_C = AIC + \frac{2(p+1)(p+2)}{n-p-2} \tag{3.7.}$$

Burada p parametre sayısıdır. AIC_C kriteri örneklem hacmi küçük iken ya da bağımsız değişkenlerin sayısı nispeten örneklem hacmine göre büyük ise tercih edilmektedir.

3.1.5. Bayesyen Bilgi Kriteri

Son bir bilgi kriteri olarak , Schwarz(1978) tarafından tanımlanan, bayesyen bilgi kriteri (BIC) ele alınacaktır. BIC, AIC kriterine benzer fakat BIC örneklem hacmine dayanan karmaşıklığı gidermek için bir ceza terimi ile düzeltme yapar ve

$$\begin{aligned} BIC &= -2\log L(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, ..., \hat{\beta}_p, \hat{\sigma}^2) + (k+2)\log n \\ &= n\log\frac{RSS}{n} + k\log n + n + n\log(2\pi) \end{aligned} \tag{3.8.}$$

şeklinde tanımlanır.

3.1.6. Mallows'un C_p Kriteri

Mallows'un C_p kriteri, Mallows (1973) tarafından geliştirilmiştir. Modelin tahmin edilmiş değerinin hata kareler ortalamasına bağlı olarak tanımlanmış bir kriterdir. Yani;

$$E(\hat{y}_i - E(y_i))^2 = \left[E(y_i) - E(\hat{y}_i)\right]^2 + Var(\hat{y}_i) \tag{3.9.}$$

dir. $E(y_i)$, gerçek regresyon denklemindeki bağımlı değişkenin ortalaması ve $E(\hat{y}_i)$ ise p terimli alt küme modelinden tahmin edilmiş değerin ortalamasıdır. Böylece $E(y_i)-E(\hat{y}_i)$, i-inci veri noktasındaki yanlılıktır. Böylece, (3.5.)'in sağ tarafındaki iki terim sırasıyla hata kareler ortalamasının yanlılık karesi ve varyans elemanlarıdır. p terimli model için (3.9.) matematiksel işlemler sonucunda aşağıdaki gibi elde edilir ve bu değeri C_p ile gösterirsek,

$$C_p = \frac{SSE_p}{\hat{\sigma}^2} - n + 2p \qquad (3.10.)$$

elde edilir. Burada $\hat{\sigma}^2$, σ^2'nin bir tahminidir. p terimli model göz ardı edilebilir yanlılığa sahip ise (yani sıfıra ise) $E(SSE_p) = \frac{(n-p)\sigma^2}{\hat{\sigma}^2} - n + 2p = p$ olur. 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 değeri, $C_p = p$ doğrusunun yakınına düşer. C_p'yi hesaplamak için σ^2'nin yansız tahmin edicisi kullanılır. Bu tam model için $C_p = p$ olmasını gerektirir.Sonuç olarak iyi modeller için C_p'nin küçük değerleri tercih edilir. Yani; $C_p \approx p$ olması tercih edilir (Montgomery ve ark., 2001, Hocking, 1976).

Mallows (1973) k yı belirlemek için C_p istatistiğini ridge tahmin edici için C_k istatistiğine

$$C_k = \frac{SS_{\text{Res},k}}{\hat{\sigma}^2} - n + 2 + 2tr[H_k] \qquad (3.11.)$$

olarak uyarlamıştır. Burada $H_k = X\left(X'X + kI\right)^{-1}X'$ dır ve $SS_{\text{Res},k}$ ridge regresyon kullanıldığında elde edilen rezidü kareler toplamıdır. C_k yı minimum yapacak şekilde k nın seçimi önerilir. Prosedür olarak k ya karşı C_k nın grafiği çizilerek C_k yı minimum yapan k değeri belirlenir.

Liu (1993) k nın tahminlerine paralel olarak, d nin bazı tahminlerini vermiştir. Skaler hata kareler ortalamasını minimum yapan d nin tahminini (2.21.) ile vermiştik. Liu (1993) aynı zamanda Liu tahmin edici için ön tahmin amaçlı kriterler olan C_p istatistiğini

$$C_d = \frac{SS_{\text{Res},d}}{\hat{\sigma}^2} + 2tr\left[\tilde{H}_d\right] - (n-2) \qquad (3.12.)$$

şeklinde tanımlamıştır.

3.1.7. Öngörü Hata Kareler Toplamı (PRESS)

Bir modelin ön tahmin performansını ölçmek için bir diğer kriter de PRESS istatistiğini minimum yapmaktır. Montgomery ve Friedman'ın (1993) belirttiği gibi bu prosedür Allen (1974) tarafından önerilmiş ve Golub, Heath ve Wahba (1979) tarafından geliştirilmiştir. PRESS istatistiğinin küçük değerleri, küçük değerli toplam hata kareler ortalamasına sahip modeli temsil eder yani regresyon modeli yeni gözlemlerde iyi ön tahminler sağlar.

Regresyon modelleri bilindiği gibi veri tanımlama, kestirim ve tahmin, parametre tahmini ve kontrol olmak üzere çeşitli kullanım alanlarına sahiptir. Model seçimi için kullanılacak kriterlerde modelin kullanım amacına uygun olmalıdır. Amaç iyi bir tanımlama elde etmek veya karmaşık bir sistemin modelini elde etmekse , SSE değeri küçük olan regresyon modelleri tercih edilmektedir. Ancak regresyon denklemleri çoğu kez gözlemlerin ön tahmini veya bağımlı

değişkenin ortalamasının tahmini için kullanılır. Genel olarak kestirim (öngörü) hata kareler ortalamasının minimum olduğu modeller seçilir. Bu da az etkili açıklayıcı değişkenlerin modelden atılacağı anlamına gelir. Bu nedenle,

$$PRESS_p = \sum_{i=1}^{n}\left(y_i - \hat{y}_{(i)}\right)^2 = \sum_{i=1}^{n}\left(\frac{e_i}{1-h_{ii}}\right)^2 \qquad (3.13.)$$

şeklinde tanımlanan $PRESS_p$ istatistiği kullanılır (Chatterjee ve ark., 2000). $PRESS_p$'nin küçük değerine sahip alt regresyon modeli tercih edilir. Bu kritere bağlı seçim işlemi hesaplama zorluklarına sahiptir. $PRESS_p$ istatistiği alternatif modelleri ayırt etmede kullanışlıdır.

$PRESS$, öngörü kapasitesini yansıtacak R^2 benzeri başka bir istatistik tanımlamak içinde kullanılabilir. Bu istatistik aşağıdaki gibi tanımlanmıştır (Myers,1990).

$$R^2_{\text{Pred}} = 1 - \frac{PRESS}{\sum_{I=1}^{n}\left(y_i - \bar{y}\right)^2} \qquad (3.14.)$$

PRESS istatistiği ridge tahmin ediciye

$$PRESS_R(k) = \sum_{i=1}^{n}\left[\frac{\hat{e}_{Ri}(k)}{1-h_{k,ii}}\right]^2 = \sum_{i=1}^{n}\left[\frac{y_i - x'_i\,\hat{\beta}_R(k)}{1-x'_i\,(X'X+kI)^{-1}x_i}\right]^2 \qquad (3.15.)$$

şeklinde uyarlanmıştır.

Şimdi PRESS istatistiğini Liu tahmin edici için ele alalım. Bu durumda $\hat{\beta}_d$, d nin lineer bir fonksiyonu olduğundan hesaplamalar daha kolay olacaktır. Liu tahmin edici için PRESS istatistik

$$PRESS_d = \sum_{i=1}^{n} (\hat{e}_{d(i)})^2 = \sum_{i=1}^{n} \left[(1-d) \frac{\hat{e}_{Ri(1)}}{1-h_{1-ii}} + d \frac{\hat{e}_i}{1-h_{ii}} \right]^2 \qquad (3.16.)$$

dir. Bu istatistiği minimum yapan değerin tahmini de aşağıdaki gibi olur (Özkale ve Kaçıranlar, 2007).

$$\hat{d}_{PR} = \frac{\sum_{i=1}^{n} \frac{\hat{e}_{Ri}(1)}{1-h_{1-ii}} \left(\frac{\hat{e}_{Ri}(1)}{1-h_{1-ii}} - \frac{\hat{e}_i}{1-h_{ii}} \right)}{\sum_{i=1}^{n} \left(\frac{\hat{e}_{Ri}(1)}{1-h_{1-ii}} - \frac{\hat{e}_i}{1-h_{ii}} \right)^2} \qquad (3.17.)$$

3.1.8. Genelleştirilmiş Çapraz Geçerlilik Kriteri (GCV)

Allen'in $PRESS$ istatistiğine göre bazı avantajlara sahip olan Wahba ve ark. (1979) tarafından tanımlanan GCV kriteri aşağıdaki gibi hesaplanır.

$$GCV = n^{-1} \sum_{i=1}^{n} e_i^2 \,/\, n^{-1} \sum_{i=1}^{n} (1-h_{ii})^2 \qquad (3.18.)$$

Golub, Heath ve Wahba (1979) ridge tahmin edicide k nın seçimi için genelleştirilmiş çapraz geçerlilik (GCV) istatistiğini

$$GCV_k = \frac{\sum_{i=1}^{n} \hat{e}_{Ri}^2(k)}{\left\{ n - \left[1 + tr(H_k) \right] \right\}^2} \qquad (3.19.)$$

olarak önermiştir. Burada $\hat{e}_{Ri}(k)$ belli bir k değeri için i -inci rezidüdür. GCV yi minimum yapacak şekildeki k nın seçimi önerilir. Prosedür olarak k ya karşı GCV nın grafiği k yı belirlemede yararlı olacaktır. Benzer şekilde Liu tahmin edici için

$$GCV_d = \frac{SS_{\text{Res},d}}{\left\{n - \left[1 + tr\left(\tilde{H}_d\right)\right]\right\}^2} \tag{3.20.}$$

olarak verilmiştir. Burada $SS_{\text{Res},d}$ Liu tahmin edici için rezidü kareler toplamı ve $\tilde{H}_d = X(X'X + I)^{-1}(X'X + dI)(X'X)^{-1}X'$ dır.

3.2. Değişken Seçme İçin Hesaplama Teknikleri

Bu bölümde R^2 , s^2 ve C_p kriterlerini kullanarak olası bütün regresyon denklemlerinden model seçimini; R^2 , $\bar{R}^2$ ve C_p kriterlerini kullanarak en iyi alt küme regresyon denklemi seçimini ve adımsal seçim yöntemlerini inceleyeceğiz.

3.2.1. Olası Bütün Regresyonlar

Sabit terimli modelden başlayarak, tek bağımsız değişken , iki bağımsız değişken,…., k- bağımsız değişken içeren bütün modeller oluşturulur. Daha sonra bu modeller seçilen kritere göre değerlendirilir ve en iyi model seçilir. k- bağımsız değişken için 2^k tane aday regresyon denklemi olacaktır. k'nın değeri büyüdükçe seçim işlemide hesaplama zamanı açısından zorlaşacaktır . Montgomery ve ark.2001 de Hald (1952) sayfa 647 de yer alan Portlant çimento verisi ele alınarak dört bağımsız değişken , 16 model için R_p^2 , $\bar{R}_p^2$, MSE_p ve C_p kriterlerini kullanarak elde edilen sonuçlar verilmiştir.

3.2.2. t Testine Bağlı Seçim

$p = k + 1$ açıklayıcı değişkenli tam model için $H_0 : \beta_j = 0$ testi için test istatistiği

$$t_{k,j} = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)} \qquad (3.21.)$$

şeklinde tanımlanır. Tam modele önemli katkıda bulunabilecek regresörler büyük $|t_{k,j}|$ değerlerine sahip olacaktır ve p-regresörlü en iyi alt kümeye eklenme meyilinde olacaklardır. En iyilik için minnumum RSS veya Mallows'un C_p kriterini kullanarak karar verebiliriz. $|t_{k,j}|$, $j = 1, 2, ..., k$ için azalan şekilde sıralanır ve en iyi model elde edilecek şekilde değişken eklemeye devam edilir. Bu yöntem Daniel ve Wood (1980) tarafından t yönünde arama olarak adlandırılmıştır. Regresör sayısı nispeten büyük iken $(k > 20 \; veya \; 30)$ çok etkili bir yöntemdir.

Montgomery ve ark. (2001) de Hald (1952) Portlant çimento verisi için sonuçlar verilmiştir.

3.2.3. F Testine Bağlı Seçim

Doğrusal bir regresyonda gereksiz bağımsız değişkenler, rezidü kareler toplamında (RSS) bir azalma ve tahminlerde artan bir varyans pahasına daha az yanlı tahminler sağlar.

Az sayıda bağımsız değişkenin bulunduğu ortamlarda, belirli bağımsız değişken gruplarının modele dahil edilip edilmeyeceğini belirlemek için kısmi F testi kullanılabilir. Değişkenleri iki gruba ayırırız. Bir grup, temel grup, modelimize dahil edilecektir. Diğer grup, şüpheli grup modele dahil edilebilir veya edilmeyebilir - henüz emin değiliz. Temel ve şüpheli her iki gruptaki tüm değişkenleri içeren regresyon modeline tam (FULL) model diyoruz. Yalnızca temel

değişkenleri içeren regresyon modeline indirgenmiş (RED) model denir.

Kısmi F testinin bir test istatistiği

$$F = \frac{\frac{RSS_{RED} - RSS_{FULL}}{df_{RED} - df_{FULL}}}{\frac{RSS_{FULL}}{df_{FULL}}}$$

şeklinde tanımlanır. RED modelinin doğru olduğu sıfır hipotezi altında (şüpheli grup için tüm tahmin katsayıları sıfırdır), F, F (df_{RED} - df_{FULL}, df_{FULL}) dağılımına sahiptir. Sıfır hipotezinin kabulü, regresyon modelimiz olarak RED modeli kullanmamıza yol açar. Sıfır hipotezin reddedilmesi, şüphelenilen gruptaki değişkenleri göz ardı etmememiz gerektiğini gösterir (tahmin katsayılarından en az biri sıfır değildir). Ardından, modele hangi değişkenlerin dahil edileceğini belirlemek için şüpheli grubun alt kümelerini kullanarak testi yeniden gerçekleştirebiliriz. Kısmi F testi, Stata'da nestreg aracılığıyla kolayca gerçekleştirilebilir (bkz. [R] nestreg, Lindsey ve Sheather, 2010).

Çok sayıda değişkenin bulunduğu modellerle çalıştığımızda şüpheli tahmin edici listesi de büyüyecektir. Bu nedenle nihai regresyon modelini belirlemek için kısmi F testi yöntemini kullanmak uygun olmayacaktır. Bu durumla başa çıkmak için çeşitli algoritmalar oluşturulmuştur.

3.2.4. Adımsal Seçim Yöntemleri

Tüm aday regresyon modellerini göz önüne alarak değerlendirme yapmak zor olacağı için adımsal seçim yöntemleri geliştirilmiştir. Bu yöntemler, Mallows'un C_p kriteri hariç, sadece bir bilgi kriteri ile çalışırlar. Teknik olarak C_p kriteri adımsal seçimde kullanılabilir fakat modele hangi değişkenin ekleneceği ya da muhafaza edileceğine karar vermek oldukça zor olacaktır.Diğer tüm ölçüm

kriterleri kendi değerleri arasında bir içsel sıralamaya sahiptirler. AIC değerinin en küçük olması, $\overline{R}^2$ değerinin büyük olması tercih edilir. Mallows'un C_p kriterine göre iyilik bu değerin değişken sayısına yakın olmasıydı. Hocking (1976) da belirtildiği gibi, C_p değerinin küçüklüğü olabildiği kadar iyi modeli verir. Adımsal seçim algoritmaları modelde bulunan değişkenlere yada modele eklenen değişkenlere göre otomatik karar vermemizi sağlar.Bu bir bilgi kriterine bağlı olarak mümkün modellerin basit sıralanmasıyla yapılabilecektir. Mallows'un C_p değerinin yorumlanması için her iki öneriyide kullanırsak, algoritma modellerin basit sıralanmasına bağlı olarak karar vermeyi gerçekleştiremez. Bu nedenle, C_p değeri adımsal seçimde kullanılamaz. Bu değer leaps (sıçramalar)- ve –bounds (sınırlar) değişken seçiminde kullanılabilecektir. Şimdi bunları sırasıyla kısaca ele alacağız.

3.2.4.1. İleriye Doğru Seçim Yöntemi

Bu yöntem iteratif bir yöntemdir. Başlangıç model sadece sabit terimi içeren model olarak ele alınır. Her iterasyonda , modele değişken eklenir ve modele bu değişken eklendiğinde bilgi kriterine göre en iyi optimalı veren modeli seçeriz. Bilgi kriterine göre, değişken eklendiğinde kriterin değerinde önemli bir değişiklik yaratmıyorsa önceki iterasyonda algoritma sonuçlandırılır ve o iterasyondaki model seçilmiş olur. Algoritmada aşağıdaki adımlar uygulanır:

y ile en büyük basit korelasyona sahip değişken modele eklenecek ilk bağımsız değişken olarak seçilir. Buna x_1 diyelim. Bu değişken aynı zamanda modelin önemlilik testindeki F istatistiği için en büyük değeri üretecek olan değişkendir. F istatistiğinin değeri, önceden seçilmiş F değerini (F_{IN}) aşarsa bu değişken modele dahil edilir. Daha sonra y ile en yüksek korelasyona sahip olan ikinci değişken modele eklenir. Bu korelasyonlar kısmi korelasyonlar gibidir.

Bunlar, $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1$ modelinin rezidüleri ile x_1 üzerinden diğer aday değişkenlerin, $\hat{x}_j = \hat{\alpha}_{0j} + \hat{\alpha}_{1j} x_1$, $j = 1, 2, ..., k$, modelleri tarafından oluşturulan rezidüler arasındaki korelasyonlardır. Bu adımda y ile en büyük korelasyona sahip değişkenin x_2 olduğunu kabul edelim. Bunun anlamı en büyük kısmi F istatistiği

$$F = \frac{SSR(x_2 / x_1)}{MSE(x_1, x_2)} \tag{3.22.}$$

dir. Eğer bu F değeri, F_{IN} değerini aşarsa x_2 modele dahil edilir. Algoritma bu şekilde F değeri, F_{IN} değerini aşmadığı zaman ya da son aday değişken modele eklendiğinde sonlandırılır.

Montgomery ve ark. (2001) de Hald (1952) Portlant çimento verisi için sonuçlar verilmiştir.

3.2.4.2. Geriye Doğru Ayıklama Yöntemi

Bu yöntem de iteratif bir yöntemdir. Bu durumda başlangıç model tüm değişkenleri içeren modeldir. Her iterasyonda , modelden değişken çıkarılır ve modelden bu değişken çıkarıldığında bilgi kriterine göre en büyük değişikliği veren model aranır. Bilgi kriterine göre, değişken çıkarıldığında kriterin değerinde önemli bir değişiklik yaratmıyorsa önceki iterasyonda algoritma sonuçlandırılır ve o iterasyondaki model seçilmiş olur.

Her iki adımsal seçim algoritması mümkün model sayısı olan 2^p nin en az tanesinde inceleme yapar. Değişkenler yüksek korelasyonlu ise adımsal seçim sonuçları ve olası tüm alt küme seçim metodları birbirinden çok farklıdırlar. Algoritmalar sezgisel ve anlaması basittir. Bir çok durumda, algoritmalar mümkün olabildiğince iyi modelin seçimi ile sonuçlanır (Lindsey ve Sheather, 2010).

3.2.4.3. Efroymson Algoritması

Efroymson (1960) değişken seçimi için adımsal regresyon yöntemini önermiştir. Bu yöntemde modele daha önce eklenen değişkenler kısmi F-istatistikleri yardımıyla yeniden değerlendirilir.Modele daha önce eklenen bir değişken daha sonraki adımlarda modelden çıkarılabilir. Bir değişken için kısmi F-istatistiğinin değeri, $F_{Ç}$ yada $F-$çıkarılanın değerinden daha az ise o değişken modelden atılır. Bu yöntem iki değere ihtiyaç duyar. Bunlar F_G yada $F-$girilen ve $F_{Ç}$ değerleridir. Bu yöntem için bir örnek Hald verisi için Montgomery ve ark. (2001) de verilmiştir.

4. YANLI TAHMİN EDİCİLERDE MODEL SEÇİMİ

4.1. Çoklu İç İlişki Olması Durumu

Çoklu iç ilişki olması durumunda yaygın olarak kullanılan ridge tahmin edici ile model seçimi kaynaklarda yer almıştır. Bunların bazıları daha önceki kısımlarda incelenmiş olup burada bu amaç için tanımlanan yeni bir istatistik ele alınacak ve daha sonra bu istatistiğin, ridge tahmin ediciye alternatif olarak tanımlanan Liu tahmin edici için tanımlanması ve uygulamalarına yer verilecektir.

4.1.1. Ridge Tahmin Ediciye Bağlı Model Seçimi

Çoklu bağlantı varlığında, regresyon parametrelerinin en küçük kareler tahmin edicisinin performansının tatmin edici olmadığı iyi bilinmektedir. Sonuç olarak, OLSE tahminlerine dayanan Mallow's C_p gibi alt küme seçim yöntemleri yetersiz alt kümelerin seçimine yol açar. Alt küme seçiminde çoklu bağlantı sorununun üstesinden gelmek için, ridge tahmin edicisine dayalı yeni bir alt küme seçim algoritması önerilmiştir. Veriler çoklu bağlantı gösterdiğinde yeni algoritmanın Mallow's C_p'ye daha iyi bir alternatif olduğu gösterilmiştir.

$$y = X\beta + \varepsilon,\ E(\varepsilon) = 0, V(\varepsilon) = \sigma^2 I_n \qquad (4.1.)$$

modelini ele alalım.

Burada, y_{nx1}, $X_{nxk} = (1, X_1, X_2, ..., X_{k-1})$, $\beta = (\beta_0, \beta_1, ..., \beta_{k-1})$ ve $\hat{\beta} = (X'X)^{-1}X'y$ EKK tahmin edici ve $\hat{\beta}(\lambda) = (X'X + \lambda I)^{-1}X'y$ ridge tahmin edicidir. Ridge tahmin edicinin parametre seçimi için Hoerl ve ark.(1975) tarafından

$$\hat{\lambda}_{HKB} = \frac{(k-1)\hat{\sigma}^2}{\hat{\beta}'\hat{\beta}} \tag{4.2.}$$

yönteminin verildiğini biliyoruz. Model verilere uydurulduğunda, ilgilenilen problem y değerlerini tahmin etmektir. Tam modele dayanan uygun regresyon denklemini kullanarak, y'nin tahmin edilen değerleri

$$\hat{y}_{ik} = X_i^{'}\hat{\beta}(\lambda)\ , i = 1,2,3,...,n \tag{4.3.}$$

şeklinde verilir. Burada, $X_i^{'} = (1, X_{i1}, X_{i2}, ..., X_{ik-1})$, $\hat{y}_k = (\hat{y}_{1k}, \hat{y}_{2k}, ..., \hat{y}_{nk})'$ ve $\hat{y}_k$, y nin tahmin edilmiş değerlerini gösterir. Şimdi, açıklayıcı değişkenlerin bir alt kümesine dayalı bir alt model 'A' nın verilere uydurulduğunu varsayalım. Bu model

$$y = X_A\beta_A + \varepsilon \tag{4.4.}$$

Burada X_A, $p-1$ açıklayıcı değişkenli ve sabit terimli bir modelin nxp matrisidir. β_A, uydurulan alt modele bağlı regresyon katsayılarının $px1$ vektörüdür. $p-1$ hacimli alt kümeye dayanan y'nin tahmin edilen değerleri $\hat{y}_p = (\hat{y}_{1p}, \hat{y}_{2p}, ..., \hat{y}_{np})'$ ile verilir.

Alt küme seçim problemi, mümkün olan tüm alt kümelerden doğru öngörü veren p hacimli en iyi alt kümenin araştırılması demektir.

k hacimli tam kümeye dayalı olarak y nin tahmin edilmiş değerleri ile p hacimli bir alt kümesine dayalı olan y nin tahmin edilmiş değerleri arasındaki

Öklid uzaklığına bağlı olan yeni bir alt küme seçim kriteri aşağıdaki gibi tanımlanır.

Tanım 4.1. R_p istatistiği

$$R_p(\lambda) = \sum_{i=1}^{n} \frac{(\hat{y}_{ik} - \hat{y}_{ip})^2}{\sigma^2} - tr(H_R^{'} H_R) + tr(H_{RA}^{'} H_{RA}) + p \qquad (4.5.)$$

şeklinde tanımlanır. Burada, p alt modeldeki parametrelerin sayısıdır. $H_R = X(X'X + \lambda I)^{-1} X'$ *ve* $H_{RA} = X_A (X_A^{'} X_A + \lambda I)^{-1} X_A^{'}$. H_R ve H_{RA} en küçük kareler kullanıldığı zamanki H şapka matrisine denktir. $R_p(\lambda)$ istatistiğine dayalı alt küme seçme algoritması için aşağıdaki adımlar uygulanır.

Adım 1: $p-1$ hacimli tüm mümkün alt kümeler için $R_p(\lambda)$ istatistiğini hesapla

Adım 2: $R_p(\lambda) \leq p$ olacak şekilde minimum hacimli bir alt küme seç

Adım 3: Denk olarak p ye karşı $R_p(\lambda)$ nin değerlerinin grafiğini çiz ve $R_p(\lambda) = p$ doğrusuna daha yakın olan alt kümeyi seç.

Bazı Sonuçlar:

Sonuç 4.1. Altküme modeli 'A' yeterli ise

$$E\left[(\hat{y}_k - \hat{y}_p)'(\hat{y}_k - \hat{y}_p)\right] \cong \sigma^2 (tr(H_R^{'} H_R) - tr(H_{RA}^{'} H_{RA})). \qquad (4.6.)$$

Sonuç 4.2. $\lambda = 0$ ise $R_p(\lambda) = C_p$ dir. Yani, EKK tahmin edici kullanıldığı zaman $R_p(\lambda)$, Mallow'sun C_p istatisiğine indirgenir.

Sonuç 4.3. Altküme modeli yeterli olduğunda, $E(R_p(\lambda)) \cong p$ dir (Dorugade ve Kashid, 2010).

Örnek 4.1. Hald verisi birçok yazar tarafından analiz edilmiştir. Burada da C_p, C_p^* ve önerilen R_p kriterini Hald verisini kullanarak karşılaştıracağız. Sonuçlar Tablo 1'de verilmiştir (burada C_p^*, Mallow's C_p 'de EKK tahmincisi yerine ridge tahmin edicisi kullanılarak hesaplanır).

Çizelge 4.1. Cp, Cp* ve Rp Değerleri

Model	*Cp*	*Cp**	*Rp*
(1)	202.549	200.3	199.383
(2)	142.486	140.5	138.468
(3)	315.154	314.1	309.287
(4)	138.731	136.8	135.481
(12)	2.678	2.525	3.57
(13)	198.095	202.1	200.545
(14)	5.496	5.301	6.874
(23)	62.438	61.74	59.723
(24)	138.226	138	135.191
(34)	22.373	21.96	21.066
(123)	3.041	2.942	4.292
(124)	3.018	2.915	3.712
(134)	3.497	3.374	4.202
(234)	7.337	7.428	7.15
(1234)	5	5	5

Not: (i, j) modeldeki Xi ve Xj değişkenini gösterir.

Tablo 1'den, bu üç kriterin aynı alt kümeyi, yani $\{X_1, X_2\}$'yi seçtiği açıktır. Bununla birlikte, bazı durumlarda çoklu bağlantı varlığında C_p -istatistiği uygun bir alt küme seçememektedir. Bu gerçek, aşağıdaki örnek kullanılarak gösterilmiştir.

Örnek 4.2. Burada C_p, C_p^* ve önerilen R_p kriterini karşılaştırıyoruz. Sonuçlar Tablo 3'te verilmiştir (burada C_p^*, Mallow's C_p 'de LS tahmin edicisi yerine ridge tahmin edicisi kullanılarak hesaplanır.)

Burada, X_1, X_2, X_3, X_4 ve X_5 üzerinde $N_5(0, \sum)$'den 25 boyutunda rastgele örnekler oluşturduk, burada

$$\Sigma = \begin{bmatrix} 1 & \rho & 1-\rho & 1-\rho^2 & 1-\rho^3 \\ \rho & 1 & \rho & 1-\rho & 1-\rho^2 \\ 1-\rho & \rho & 1 & \rho & 1-\rho \\ 1-\rho^2 & 1-\rho & \rho & 1 & \rho \\ 1-\rho^3 & 1-\rho^2 & 1-\rho & \rho & 1 \end{bmatrix} \quad \rho = 0.6'da.$$

(X_1, X_3, X_4) alt modelini şu şekilde ele alıyoruz:

$$Y = 3 + X_1 + X_3 + X_4 + \varepsilon,$$

burada ε ~ N (0, 7).

VIF değerleri Tablo 2'de hesaplanmış ve raporlanmıştır ve durum endeksleri 1, 8.9, 17.26, 26.76 ve 4418.2'dir.

Tablo 2'ten, X_1 ve X_2 tahmin değişkenleri seti için VIF'lerin 100'ü ve koşul indekslerinin 1000'i aştığı açıktır. Dolayısıyla, simüle edilen verilerde çoklu bağlantı açıkça mevcuttur.

Farklı alt modeller için C_p, C_p^* ve R_p değerleri Tablo 3'te verilmiştir.

Çizelge 4.2. VIF Değerleri.

Variable	X_1	X_2	X_3	X_4	X_5
VIF	672.9	462.7	3.2	7.9	12.4

Çizelge 4.3. CP, CP* ve Rp Değerleri.

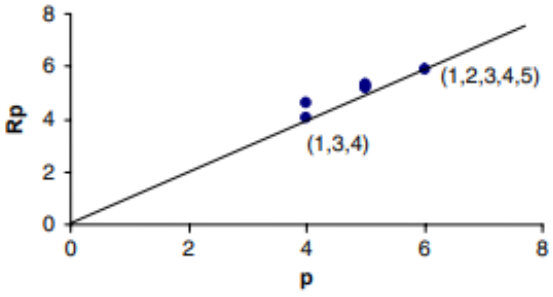
Model	*Cp*	*Cp**	*Rp*
(1)	22.1728	21.58	19.664
(2)	32.5056	31.771	29.171
(3)	34.8537	34.087	36.561
(4)	49.2361	48.271	50.389
(5)	85.2025	83.823	75.809
(12)	16.5853	16.081	14.646
(13)	11.2585	10.843	11.244
(14)	11.2313	10.816	11.133
(15)	24.2431	23.642	21.319
(23)	15.2756	14.805	15.168
(24)	13.6634	13.214	13.553
(25)	31.7519	31.048	27.914
(34)	6.9945	6.637	10.78
(35)	29.7271	29.057	29.767
(45)	20.8885	20.341	20.954
(123)	10.021	248.446	254.91
(124)	12.8596	12.414	12.062
(125)	16.0253	15.566	14.636
(134)	2.2181	1.954	4.048
(135)	13.3736	12.948	12.947
(145)	12.1089	11.705	11.374
(234)	2.6529	2.383	4.622
(235)	17.2773	16.8	16.404
(245)	12.3381	11.933	11.606
(345)	5.4108	5.103	7.845
(1234)	4.3348	4.06	5.333
(1235)	8.7128	8.381	9.455
(1245)	14.0705	13.67	12.803
(1345)	4.1309	3.866	5.203
(2345)	4.1271	3.861	5.27
(12345)	5	5	5

Not: (i, j) modeldeki Xi ve Xj değişkenini gösterir.

Alt küme seçim algoritmasını kullanarak, Tablo 3 ve Şekil 1'den hem C_p hem de C_p^* istatistiğinin $\{X_1, X_2, X_3, X_4\}$ alt kümesini seçtiği açıktır. Bu nedenle, hem C_p hem de C_p^* uygun bir $\{X_1, X_3, X_4\}$ alt kümesi seçemez. Öte yandan, önerilen istatistik R_p, uygun alt kümeyi $\{X_1, X_3, X_4\}$ seçer.

Son olarak, burada R_p'nin tanımında, $\hat{\sigma}^2$'nin tam modele dayalı uygun tahmini ile değiştirilmesi gerektiğini belirtmiştik. İyi bilindiği gibi, çoklu bağlantı

mevcut olduğunda ve ridge tahmin edicisi kullanılıyorsa, $\hat{\sigma}^2$ tahmini bir yanlı tahmin edicidir. Bu nedenle, bu aşamada σ^2'nin uygun seçiminin ne olması gerektiği ve önerilen kriterin performansını nasıl etkilediği doğal bir soru ortaya çıkar. Bir sonraki bölümde, bu soruyu cevaplamaya çalışacağız. (Dorugade ve Kashid, 2010).



Şekil 4.1. Rp - p grafiği.

4.1.2. Liu Tahmin Ediciye Bağlı Model Seçimi

Bu kısımda bir alt küme seçiminde çoklu bağlantı sorununun üstesinden gelmek için, ridge tahmin edicisine dayalı alt küme seçim algoritması , ridge tahmin ediciye alternatif olarak tanımlanan Liu tahmin ediciye uyarlanacaktır. Liu tahmin edicinin parametre seçimi Liu (1993) tarafından

$$\hat{d}_{mm} = 1 - \hat{\sigma}^2 \left(\sum_{i=1}^{p} \frac{1}{\lambda_i (\lambda_i + 1)} \Big/ \sum_{i=1}^{p} \frac{\hat{\alpha}_i^2}{(\lambda_i + 1)^2} \right) \qquad (4.7.)$$

şeklinde verilmiştir.

k hacimli tam kümeye dayalı olarak y'nin tahmin edilmiş değerleri ile p hacimli bir alt kümesine dayalı olan y'nin tahmin edilmiş değerleri arasındaki

Öklid uzaklığına bağlı olan yeni bir alt küme seçim kriteri Liu tahmin edici için aşağıdaki gibi tanımlanır.

Tanım 4.1. R_p istatistiği

$$R_p(d) = \sum_{i=1}^{n} \frac{(\hat{y}_{ik} - \hat{y}_{ip})^2}{\sigma^2} - tr(H_L^{'} H_L) + tr(H_{LA}^{'} H_{LA}) + p \qquad (4.8.)$$

şeklinde tanımlanır. Burada,

$$H_L = X(X'X + I)^{-1}(X'X + dI)(X'X)^{-1}X'$$

ve

$$H_{LA} = X_A(X_A^{'}X_A + I)^{-1}(X_A^{'}X_A + dI)(X_A^{'}X_A)^{-1}X_A^{'}$$

dır. $R_p(d)$ istatistiğine dayalı alt küme seçme algoritması için aşağıdaki adımlar uygulanır.

Adım 1: $p-1$ hacimli tüm mümkün alt kümeler için $R_p(d)$ istatistiğini hesapla

Adım 2: $R_p(d) \le p$ olacak şekilde minimum hacimli bir alt küme seç

Adım 3: Denk olarak p ye karşı $R_p(d)$ nin değerlerinin grafiğini çiz ve $R_p(d) = p$ doğrusuna daha yakın olan alt kümeyi seç.

4.2. Sapan Değerler Olması Durumu

EKK tahmin ediciye dayanan C_p istatistiğinin sapan değerler olması durumunda yada hataların normallik varsayımının sağlanmaması durumunda çok hassas olduğunu biliyoruz. Son zamanlarda bir çok dayanıklı parametre tahmin metodu ve alt küme seçim metodu geliştirilmiştir.

Kullback-Leibler (K-L) bilgi kriterine bağlı olarak model seçimi kriterleri çok yaygındır. Akaike bilgi kriteri (AIC) ve düzeltilmiş AIC bu gruba ait kriterlerdir. K-L bilgi kriterine bağlı olarak tahminin yanlılığında azaltma daha iyi değişken seçimine götürecektir. K-L bilgi kriteri gerçek model ile aday model arasındaki ayrışımın ortalaması olarak kullanılır. X sürekli rasgele değişken ve $f(x/\theta)$ olasılık yoğunluk fonksiyonu olsun. Burada θ, p-boyutlu parametre vektörüdür. θ^*, $f(x/\theta^*)$ yoğunluk fonksiyonu ile θ nın gerçek değeri olsun. K-L bilgi kriteri $f(x/\theta^*)$'ın $f(x/\theta)$'ya yakınlığının ölçüsüdür:

$$\begin{aligned} B(\theta^*;\theta) &= -I(\theta^*;\theta) \\ &= E\left[\log f(x/\theta) - \log f(x/\theta^*)\right] \\ &= \int f(x/\theta^*)\log f(x/\theta)dx - \int f(x/\theta^*)\log f(x/\theta^*)dx \\ &= H(\theta^*;\theta) - H(\theta^*;\theta^*), \end{aligned} \qquad (4.9.)$$

Burada I , K-L bilgiyi , E de beklenen değeri gösterir. K-L bilgi:

$$I(\theta^*;\theta) = -B(\theta^*;\theta) = H(\theta^*;\theta^*) - H(\theta^*;\theta) \qquad (4.10.)$$

şeklinde yazılabilir. (4.9.)'u maksimum yapma yerine (4.10.) minimum yapılır. $H(\theta^*;\theta^*)$ bir sabit olduğundan $B(\theta^*;\theta)$ aşağıdaki gibi yazılabilir:

$$B(\theta^*;\theta) = \int f(x/\theta^*)\log f(x/\theta)dx$$

$H(\theta^*;\theta)$ nın ($\theta = \theta^*$) da θ'ya göre türevi Fisher bilgi kriterine eşittir.

Fisher Bilgi (F-B): Fisher bilgi istatistiksel tahminde ve teorik çalışmalarda önemli katkı sağlar.Etkinlik ve yeterlilik kavramlarıyla yakından ilişkilidir. n bağımsız gözlem için

$$I_F = nE_\theta \left\{ \left(\frac{\partial}{\partial \theta} \sum_{i=1}^{n} \ln f(x_i;\theta \right)^2 \right\}$$

şeklinde tanımlanır. $f(x;\theta)$ rasgele değişkenin olasılık yoğunluk fonksiyonudur. F-B negatif olmayan değerlere sahiptir ve θ ile ilişkili bilginin miktarını ölçer.F-B yansız tahminin düzeltilmesiyle alakalıdır. F-B ve Kullback'in ayrıştırma bilgisi gözlemlerin gruplandırılması, toplanabilirliği, etkinliği ve yeterliliğinde benzerliklere sahiptir(Çetin ve Erar, 2002).

Ronchetti (1985) AIC kriterinin dayanıklı versiyonunu tanımladı ve RAIC olarak adlandırdı.

$$RAIC(p;\alpha,\rho) = 2\sum \rho(r_{i,p}) + \alpha p$$

Burada $r_{i,p} = (y_i - x_i^{'} T_{n,p}) / \hat{\sigma}$ ve $\hat{\sigma}$, σ nın bazı robust tahminidir, $T_{n,p}$, M-tahmin edici ve $\alpha = 2$ dir. ρ da, (2.11.) verdiğimiz gibi Huberin fonksiyonu olarak seçilir. AIC'in RAIC'e genişletilmesi en çok olabilirlik tahminin M-tahmine genişletilmesinden ibarettir.

Fisher Bilgi kriterinin kullanılmasıyla Akaike bilgi kriteri için yanlılık düzeltmesi model seçimi için Çetin (2006) tarafından aşağıdaki gibi verilmiştir.

$$AICF = \frac{n}{\hat{\sigma}^2}\left[\frac{np}{n-p-2} + 2\right]. \qquad (4.11.)$$

Örnek 4.3.

Şimdi sağlam ve klasik Akaike değişken seçim kriterlerini karşılaştırmak için bir simülasyon çalışması sunuyoruz. Kriterlere göre seçilen alt kümelerin yüzdelerini elde etmek için S-Plus fonksiyonları [Çetin, M. 2000.] kullanılarak bir program kodlanmıştır. Sırasıyla 0 beklentisi ve $\sigma^2 = 0.01, 1, 100$ varyansı ile U[−1, +1] ve 15 bağımsız normal dağılımlı hata e_i üzerinde beş bağımsız tek biçimli rastgele değişkenin 15 bağımsız kopyasını oluşturduk. Daha sonra 2 modele göre y_i gözlemleri oluşturduk:

$$Beta1 : (2\sqrt{5}, 4, \sqrt{3}, 1, 0) \text{ ve } Beta2 : (2\sqrt{5}, 4, -\sqrt{3}, 1, 0) .$$

Model, kesintiye uğramadan kullanılmak üzere tasarlanmıştır. Uç değerlerin tahmin ediciler ve seçim kriterleri üzerindeki etkilerini görmek için e için aykırı değer olmaması, bir aykırı değer ve iki aykırı değer olması durumunda değişken seçim kriterleri incelenir. $2^5 = 32$ alt küme elde ettik. Bununla birlikte, aşağıdaki tablolarda, 32'den yalnızca en çok seçilen alt kümeler verilmiştir. Ayrıca 'ye göre sıra seçim kriterlerinin frekansları da aşağıdaki tablolarda verilmiştir.

Bu tablolarda 'P' modeli gösterir. Örneğin, bir '1', x_1 değişkenini içeren alt kümeleri etiketler.

Çizelge 4.4. Beta1 kriterlerine göre seçilen alt kümelerin yüzdesi

		$\sigma^2 = 0.01$			$\sigma^2 = 1$			$\sigma^2 = 100$		
	P	RAIC	AICF	AICC	RAIC	AICF	AICC	RAIC	AICF	AICC
aykırı değer yok	1	87.2	-	-	41.6	-	-	23	-	78.0
	12	-	-	0.22	0.2	-	-	16	1	21
	13	-	-	-	12.0	-	-	16	-	-
	123	-	-	-	17.8	-	-	7.6	-	0.6
	124	-	-	-	7.6	-	-	9	-	-
	125	-	-	-	0.4	-	-	5.2	-	0.4
	1234	12.8	100	99.7	5.0	100	100	6.6	38.4	-
	1235	-	-	-	14.2	-	-	3.4	26	-
	1245	-	-	-	1.2	-	-	5.4	27.4	-
	1345	-	-	-	-	-	-	7.8	7.2	-
bir aykırı değer	1	96.8	-	-	56.0	-	-	10.2	-	100
	12	-	-	0.4	-	-	0.6	0.8	-	-
	13	-	-	-	12.8	-	-	14.8	-	-
	123	-	-	-	-	-	-	2.8	-	-
	124	-	-	-	-	-	-	2	-	-
	125	-	-	-	-	-	-	1.2	-	-
	1234	3.2	100	99.6	0.6	100	99.4	14.2	47.2	-
	1235	-	-	-	0.6	-	-	7	2	-
	1245	-	-	-	-	-	-	0.2	-	-
	1345	-	-	-	-	-	-	46.8	50.8	-
iki aykırı değer	1	99.6	-	-	49.6	-	-	21.8	-	98.8
	12	-	-	-	-	-	32.2	7.8	-	-
	13	-	-	-	49.0	-	-	19.8	0.2	8.2
	123	-	-	-	0.4	-	30.6	11	-	-
	124	-	-	-	0.6	-	-	6.2	-	-
	125	-	-	-	-	-	-	4.8	-	-
	1234	0.4	100	100	-	100	37.2	5.8	24.4	-
	1235	-	-	-	-	-	-	6.6	42.8	-
	1245	-	-	-	0.4	-	-	9	2	-
	1345	-	-	-	-	-	-	7.2	30.6	-

Çizelge 4.5. Beta2 kriterlerine göre seçilen alt kümelerin yüzdesi

		$\sigma^2 = 0.01$			$\sigma^2 = 1$			$\sigma^2 = 100$		
	P	RAIC	AICF	AICC	RAIC	AICF	AICC	RAIC	AICF	AICC
aykırı değer yok	1	96.8	-	-	96.8	-	-	28	-	50.4
	12	-	-	-	-	-	-	13.6	-	40
	13	-	-	-	-	-	-	19	-	6.6
	123	-	-	-	-	-	-	8.4	-	0.8
	124	-	-	-	-	-	-	8.6	-	1.4
	125	-	-	-	-	-	-	5.8	-	0.8
	1234	3.2	100	100	3.2	100	100	5.4	39.4	-
	1235	-	-	-	-	-	-	2.4	27.8	-
	1245	-	-	-	-	-	-	3.4	24.8	-
	1345	-	-	-	-	-	-	5.4	8	-
bir aykırı değer	1	99.4	-	-	39.4	-	-	26.2	-	99.8
	12	-	-	-	3.4	-	1.6	3.6	-	0.2
	13	-	-	-	51.8	-	-	8.6	-	-
	123	-	-	-	-	-	0.2	0.4	-	-
	124	-	-	-	0.2	-	63.2	3	-	-
	125	-	-	-	3.8	-	-	4.2	-	-
	1234	0.6	100	100	0.4	100	35.0	10.4	45.6	-
	1235	-	-	-	1.0	-	-	4.2	2	-
	1245	-	-	-	-	-	-	2	4.4	-
	1345	-	-	-	-	-	-	37.4	48	-
iki aykırı değer	1	100	-	-	31.8	-	-	20.4	1.4	100
	12	-	-	-	3.0	-	32.0	8.2	-	-
	13	-	-	-	56.4	-	-	18.4	0.2	-
	123	-	-	-	-	-	4.2	12.2	-	-
	124	-	-	-	2.8	-	52.8	6.2	-	-
	125	-	-	-	2.8	-	-	6.8	-	-
	1234	-	100	100	-	100	11.0	5	19	-
	1235	-	-	-	-	-	-	11.8	29	-
	1245	-	-	-	3.2	-	-	7.4	31.2	-
	1345	-	-	-	-	-	-	8.6	19.2	-

Tablo 4'e göre, varyans küçük olduğunda, AICC ve AICF kriterleri, aykırı değer içermeyen %100 gerçek model '1234'ü seçer. Aksine, RAIC kriterleri her durumda düşük performans gösterir. Aykırı değer olmaması ve bir aykırı değer olması durumunda, AICF ve AICC kriterlerinin sonuçları $\sigma^2 = 0.01$ ve $\sigma^2 = 1$ için benzerdir. Bir aykırı değerle AICF ayrıca %100 ile '1234' modelini seçer.

$\sigma^2 = 1$ olduğunda, iki aykırı değer olması durumunda AICC kriteri başarısız olur. Genel olarak, $\sigma^2 = 0.01$ ve $\sigma^2 = 1$ sonuçları Beta1 için benzerdir. AICC kriteri $\sigma^2 = 0.01$ ve $\sigma^2 = 1$ için iyi bir performans sağlasa da $\sigma^2 = 100$ için başarısız olur. AICC'nin varyanstaki bir değişiklikten ve varyans büyük olduğunda etkilendiği bilinmektedir [Çetin, M., Erar A. 2006.]. Aykırı değerlere göre büyük bir varyans olması durumunda, AICF kriteri aynı zamanda gerçek "1234" modelini ve "1345" modelini seçer, oysa AICC başarısız olur. Varyans büyüdüğünde, AICC kriteri aykırı değerlerden çok etkilenir. AICF kriteri, iki aykırı değer olsa bile, %100 ile '1234' modelini seçer.

Tablo 5'ten Beta2 sonuçlarının Beta1 sonuçları ile benzer olduğu görülmektedir. Beta2 durumunda, AICC ve AICF Beta1 kadar iyi bir performans sağlarken, RAIC yine iyi bir performans sağlamamaktadır. Ancak, Beta1'in aksine, bir aykırı değer durumunda AICC kriteri $\sigma^2 = 1$ için başarısız olur.

Uygulamanın Hald verisinde kullanımı

Bu veri seti, bu değişkenlerin birbiriyle ilişkili olduğu 4 bağımsız değişkene ilişkin 13 gözlemden oluşmaktadır. Bu veriler için açıklayıcı değişkenler arasında çoklu bağlantı olduğuna dikkat edilmelidir. y yönünde herhangi bir aykırı değer olmadığı gözlemlenmiştir. Aykırı değer olmadığında Hald veri kullanılarak elde edilen 15 alt küme için bazı sonuçlar Tablo 6'da verilmiştir.

Çizelge 4.6. Hald veri için bazı alt küme sonuçları

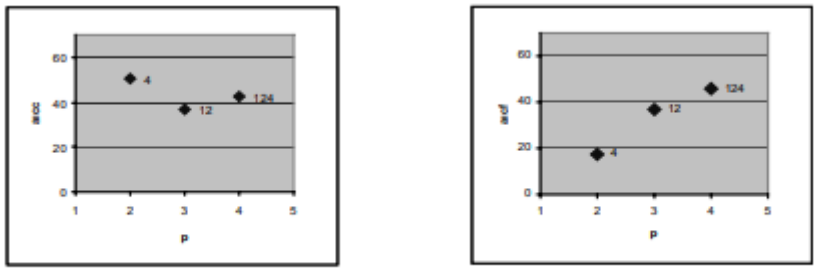
P	RAIC	AICF	AICC
12	11.49	36.88	36.62
13	150.0	1.74	53.8
14	6.10	28.56	38.07
123	11.22	45.80	42.46
124	9.24	45.93	42.44
134	3.91	43.34	42.77
234	17.28	29.85	44.88

Hald veri için model '12', yani x_1 ve x_2'yi içeren model, tahmin edilen model olarak özellikle tercih edilmektedir. Bu alt kümede çoklu bağlantı yoktur. Model '124' 3 değişkenli en iyi modeller olarak önerilebilir.

Tablo 6 ve Şekil 2'den, AICC kriterine göre '12' ve '124' modellerinin seçildiği görülmektedir. AICC kriterine göre '12' modeli seçildiğinde, '124' modeli de uygun bir 3 değişkenli modeldir. Ancak AICF değeri parametre numarasına karşı çizilirse AICF tarafından '12'nin seçildiği görülür.

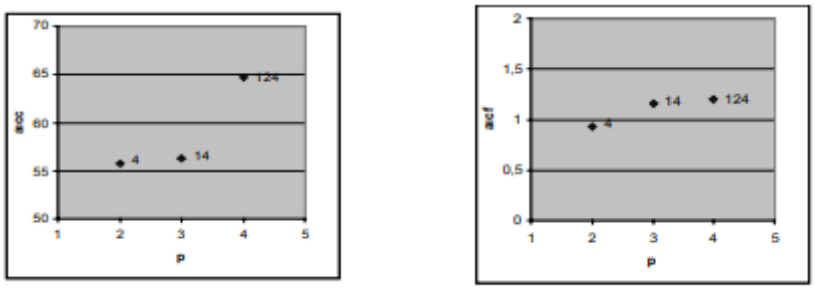
AICF, üç değişkenli '124' modelini seçer. AICF'nin çok değişkenli alt kümeleri seçme eğiliminde olduğu görülmektedir (Çetin, 2000.). RAIC, simülasyon çalışmasında da iyi bir performans sağlamamaktadır.

Aykırı değerlerin etkisini görmek için bağımlı değişken üzerinde uç değerler üretilir: Bağımlı değişken önce $y_6 = 150$, daha sonra $y_6 = 150$ ve $y_8 = 150$ ile değiştirilir. Sonuçlar Şekil 3 ve 4'te verilmiştir.



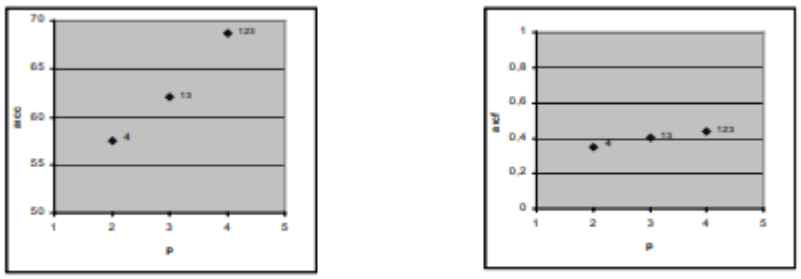
(a) (b)

Şekil 4.2. Aykırı değerler içermeyen AICC-p ve AICF-p



(a) (b)

Şekil 4.3. Bir aykırı değerle AICC-p ve AICF-p



(a) (b)

Şekil 4.4. İki aykırı değer içeren AICC-p ve AICF-p

Bir aykırı değer olması durumunda, AICC ve AICF kriterleri '14' modelini seçer. Hald verisi üzerine yapılan önceki çalışmadan, '14' alt kümesinin en iyi modellerden biri olduğu bilinmektedir. Ayrıca, bir aykırı değer için AICC ve AICF tarafından '14' alt kümesi seçilir(Çetin ve Erar, 2002).

Şekil 3'e göre, AICC ve AICF aynı '13' alt kümesini seçer. Ancak bu durum AICC için pek arzu edilen bir durum değildir. AICF iki aykırı değerden etkilenir.

Ronchetti ve Statudte (1994), C_p kriterinin dayanıklı versiyonunu aşağıdaki gibi tanımladı ve RC_p olarak gösterdi:

$$RC_p = \frac{W_p}{\hat{\sigma}^2} - (U_p - V_p) \tag{4.12.}$$

Burada $W_p = \sum \hat{w}_i^2 r_i^2 = \sum \hat{w}_i^2 / (y_i - \hat{y}_i)^2$, w_i i-inci gözlem için ağırlık ve $\hat{\sigma}^2$, $\hat{\sigma}^2 = \frac{W_{full}}{U_{full}}$ ile verilen tam modeldeki $\hat{\sigma}$ nın robust ve tutarlı tahmin edicisidir.

$U_p = \sum \text{var}(\hat{w}_i r_i)$ ve $V_p = \sum \text{var}(\hat{w}_i x_i^{'}(\hat{\beta}_p - \beta))$ sabitleri altkümeler doğru ve $\sigma = 1$ varsayımı altında hesaplanır. Çetin (2009) da Ronchetti ve Statudte (1994) dan yararlanarak her alt kümeye karşılık değişen $U_p - V_p$ değerini göz önüne almıştır. Çetin (2009) da , Ronchetti ve Statudte tahminleri hesaplarken ağırlıklı en küçük kareleri (WLS) kullanır iken Huber tipi tahmin ve Huber ağırlıklarını kullanmıştır (Çetin , 2000). Bu durumda $U_p - V_p$ aşağıdaki gibi verilmiştir.

$$U_p - V_p \approx nE\|\eta\|^2 - 2tr(NM^{-1}) + tr(LM^{-1}QM^{-1}), \tag{4.13.}$$

burada,

$$E\|\eta\|^2 = \sum_{1\le i\le n} \eta^2(x_i,\varepsilon_i), N = E\left[\eta^2\eta' xx'\right] \text{ ve } L = E\left[w'\varepsilon(w'\varepsilon + 4w)xx'\right]$$

dır.

Mallows'un C_p ve Ronchetti ve Statudte' in RC_p model seçimi faydalı araçlar olmakla birlikte hesaplama zorlukları ve yoğun hesaplama gerektirmesi nedenleriyle birçok dezavantaja sahiptirler. Sommer ve Huggins(1996) ,Wald test istatistiğine dayanan daha esnek ve daha kolay genelleştirilebilen ve sadece tam modelden tahminlerin hesaplanmasını gerektiren RT_p kriterini tanımladılar.

T_p 'nin robust versiyonu regresyon parametrelerinin genelleştirilmiş M-tahmin edicilerine dayanmaktadır ve

$$RT_p = \hat{\beta}_2' \sum_{22}^{-1} \hat{\beta}_2 - k + 2p \qquad (4.14.)$$

şeklinde tanımlanmıştır. Burada, $\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2) = (X'VX)^{-1}X'Vy$, $\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$ kovaryans matrisi, V köşegenleri v_{ii} olan matris, k ve p sırasıyla tam model ve alt modelin boyutlarıdır. $\hat{\beta}_1 = (\hat{\beta}_0,...,\hat{\beta}_{p-1})$ ve $\hat{\beta}_2 = (\hat{\beta}_p,...,\hat{\beta}_{k-1})$. Alt model doğru ise RT_p nin değeri p ye yakın olacaktır (Sommer ve Huggins, 1996).

Çetin (2009)'da daha önce tanımladığımız Liu, dayanıklı Liu tahmin ediciyi ve Liu tahmin edicinin dayanıklı parametre seçimini göz önüne alarak ((2.10), (2.16), (2.32) ve (2.33) denklemleri ile gerekli bilgiler verilmişti) RC_p ve RT_p model seçim kriterlerini kullanarak model seçimini incelemiştir.

Örnek 4.4.

Regresyon modelleri için sağlam Liu tahmin edicisi ile değişken seçim kriterlerinin ne kadar iyi olduğunu görmek için S-Plus'ta kodlayan bir bilgisayar programı geliştirdik (Çetin, 2000).

Aykırı değerler ve çoklu bağlantı varlığında sağlam değişken seçim kriterlerinin performanslarını karşılaştırmak için Myers'da (1990) verilen tütün veri setini kullandık. Myers, bu veri seti için açıklayıcı değişkenler arasında ciddi bir çoklu doğrusal bağlantı olduğuna ve ayrıca y yönünde iki aykırı değer olduğuna işaret etmektedir.

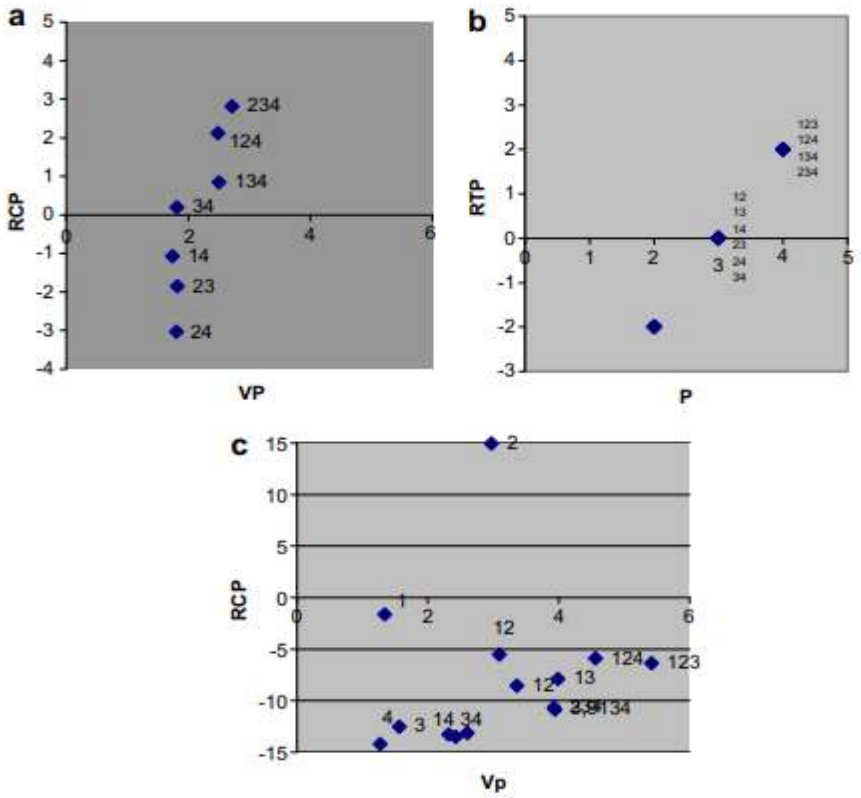
Bu veriler için sağlam model seçim kriterleri Huber-M ve sağlam Liu-M tahmin edicileri kullanılarak hesaplanmıştır. Tütün verileri için model seçim sonuçları Tablo 7'de verilmiştir. Ayrıca Şekil 5'de RCP'ye karşı V_P ve RTP'ye karşı p grafiğini de gösteriyoruz.

Tablo 8 ve Şekil 5, Liu tahmin edicili RCP'nin üç değişkenin önemi üzerinde hemfikir olduğunu göstermektedir: x_1, x_2, x_4 ve x_1, x_3, x_4, çünkü bu modeller RCP≤V_P'yi karşılamaktadır. Bu şekilde RTP, x_1, x_2, x_3'ü önerir; x_2, x_3, x_4; x_1, x_2, x_4 ve x_1, x_3, x_4 modelleri. Ancak Huber tahmin edicisi ile RCP bunların hiçbirini seçmez. Dolayısıyla Huber tahmin edicisi ile RCP uygulamasının bu veri seti için uygun olmadığı açıktır.

Örnek 4.5.

Bu bölümde, Liu ve sağlam Liu-M tahmin edicilerini kullanarak sağlam değişken seçim kriterlerini incelemek için bir simülasyon çalışması gerçekleştiriyoruz. 15 durum ve x_1, x_2, x_3, x_4, x_5 gibi beş bağımsız değişkene sahip veriler normal dağılımdan üretilmiştir. Bu değişkenler bir regresyon modeline eklenir ve bu modelde katsayı değerleri sırasıyla beş bağımsız değişkene göre (5, 3, $\sqrt{6}$, 0, 0) olarak alınır (Çetin, 2000). Ek olarak, çoklu bağlantı ve aykırı

değerin sağlam seçim kriterleri üzerindeki birleşik etkisini analiz etmek için x_1 - x_4 ve x_2 - x_3 ikili değişkenleri arasında aykırı ve doğrusal bağımlılıklar sağlıyoruz. Simülasyonun ilk aşaması için bu veri setleri için Liu tahminleri elde edilmiş ve ardından Tablo 9'da verilen tüm alt kümelere 100 tekrarlama ile RCP ve RTP kriterleri uygulanmıştır. Bu simülasyonun sonuçları Tablo 9'da verilmiştir. Tablo 9'dan, RCP'nin replikasyonlarda 100 kez doğru modeli (x_1, x_2, x_3) seçtiğini, RTP'nin ise koşul olarak herhangi bir modeli seçemediğini gözlemliyoruz, RTP < parametre sayısı , memnun değil. RCP kriteri de modeli (x_1, x_2, x_3, x_4) 100 tekrarda 99 kez alternatif uygun model olarak seçer.



Şekil 4.5. Tütün verileri için model seçimi. (a) RCP vs VP (Liu-M); (b) RTP'ye karşı p c) RCP'ye karşı VP (Huber).

Çizelge 4.7. Tütün verileri için sağlam Liu-M ve Huber-M parametre tahminleri

Robust $-\hat{\alpha}_{LM}$	$\hat{\beta}_{LM}$	Huber $-$ M$\hat{\beta}$
1.0108	0.5064	1.0274
−0.4233	0.3641	−0.4824
−0.6008	−0.3666	−0.6750
0.9676	−0.2799	1.1036
d = 0.8785		

Ancak, Liu-M tahmin edicili RCP kriteri, Liu tahmin edicili RCP kriterinden daha iyi performans gösteremez. RTP kriterinin üç ve dört değişkenli modelleri seçme eğiliminde olduğunu belirtmek isteriz. Sağlam Liu-M tahmin edicisi ile RTP kriteri ve RTP kriteri Huber-M tahmin edicisi (Çetin, 2000) benzer sonuçlar elde etti.

Çizelge 4.8. Alt kümelerin RCP, VP ve RTP değerleri.

Subsets	With Robust Liu-M estimators			With Huber-M estimators		
	RCP	V_p	RTP	RCP	$V_{H\,p}$	RTP
x_1	100.077	2.093758	1.9930	1.6177	1.3421	1.9962
x_2	154.6534	1.947888	−1.9971	14.942	2.970	−1.9873
x_3	55.34135	1.309717	−1.9908	−12.559	1.5586	−1.9948
x_4	39.82961	1.422388	−1.9914	−14.226	1.2675	−1.9991
x_1, x_2	47.50134	3.957321	0.00235644	−5.5328	3.0952	0.0021
x_1, x_3	22.99912	2.537972	0.00534423	−7.9121	3.9835	0.0035
x_1, x_4	−1.070512	1.740131	0.00650112	−13.2970	2.3207	0.00057
x_2, x_3	1.852897	1.824981	0.00150020	8.5489	3.3626	0.00506
x_2, x_4	−3.036544	1.80666	0.00154391	−13.1299	2.6057	0.00021
x_3, x_4	0.1903111	1.811402	0.00848519	−13.4978	2.4231	0.0006017
x_1, x_2, x_3	19.6815	3.654669	2.000962	−6.3446	5.4184	**2.00136**
x_1, x_2, x_4	**2.118568***	**2.490464**	2.001316	−5.8920	4.5610	**2.00020**
x_1, x_3, x_4	0.8426743	2.507251	2.004977	−10.8790	3.9413	**2.00056**
x_2, x_3, x_4	**2.815754***	**2.717739**	2.001341	−10.6702	3.9290	**2.00018**

Çizelge 4.9. Liu ve sağlam Liu-M tahmin edicilerinin sonuçları.

Subsets	With Liu estimators		With Robust Liu-M estimators	
	RCP < **VP**	RTP < P	RCP < **VP**	RTP < P
x_1	0	0	21	0
x_2	0	0	8	2
x_3	0	0	1	2
x_4	0	0	22	0
x_5	0	0	13	2
x_1, x_2	0	0	4	0
x_1, x_3	0	0	7	0
x_1, x_4	0	0	0	19
x_1, x_5	0	0	2	19
x_2, x_3	2	0	0	72
x_2, x_4	0	0	11	0
x_2, x_5	0	0	0	72
x_3, x_4	0	0	30	0
x_3, x_5	0	0	1	72
x_4, x_5	0	0	3	19
x_1, x_2, x_3	**100**	0	0	100
x_1, x_2, x_4	0	0	44	100
x_1, x_2, x_5	0	0	0	100
x_1, x_3, x_4	0	0	**65**	100
x_1, x_3, x_5	0	0	0	100
x_1, x_4, x_5	0	0	0	99
x_2, x_3, x_4	2	0	0	100
x_2, x_3, x_5	0	0	2	67
x_2, x_4, x_5	0	0	0	100
x_3, x_4, x_5	0	0	5	100
x_1, x_2, x_3, x_4	**99**	0	0	100
x_1, x_2, x_3, x_5	0	0	19	100
x_1, x_2, x_4, x_5	0	0	**85**	100
x_1, x_3, x_4, x_5	0	0	**89**	100
x_2, x_3, x_4, x_5	5	0	0	100

Sonuçların simülasyondaki parametrelere dayandığını bildiğimiz için, RCP ve RTP kriterlerini, çeşitli aykırı değerler ve çoklu bağlantı yapılarına sahip farklı bir simülasyon çalışması kullanarak incelemeyi umuyoruz(Çetin, 2009).

Veride etkili gözlemler olması durumunda etkili gözlemlerin atılmasıyla uyarlanmış C_p kriteri Kim ve Hwang (2000) tarafından tanımlanmıştır. Bununla birlikte Ronchetti ve Statudte' nin yaptığı gibi tüm gözlemlerin modelde kalması ve dayanıklı tahmin yönteminin kullanılması daha iyi bir yaklaşım olacaktır. Bir çok yazarın belirttiği gibi OLSE yerine robust tahmin edici yerleştirerek C_p kriterine benzer klasik alt küme seçim yöntemlerini uygulamak yeterli değildir.Bu yüzden C_p nin robust versiyonu olan RC_p tanımlanmıştır. Bu nedenlerle, Kashid

ve Kulkarni (2002) veride sapan değerler olması durumunda alt küme seçimi için daha genel kriter olan S_p kriterini tanımladılar. S_p kriteri diğer dayanıklı alt küme seçim metotlarına göre operasyonel olarak daha basittir. Bu kriter,

$$S_p = \sum_{i=1}^{n} \frac{(\hat{y}_{ik} - \hat{y}_{ip})^2}{\sigma^2} - (k - 2p) \qquad (4.15.)$$

şeklinde tanımlanmıştır. Bilinmeyen σ yerine $\hat{\sigma} = 1.4826 median|r_i - median(r_i)|$ şeklinde uygun tahmini yerleştirilir. Diğer seçimler örnekler kısmında incelenmiştir. Burada, r_i i-inci rezidü, k ve p sırasıyla tam model ve alt modelin parametre sayılarıdır. (4.15.) kriteri hesaplanır iken β ve β_A nın bazı uygun tahmin edicileri kullanılabilir. Alt model yeterli olduğunda $E(S_p) = p$ olur. Veride sapan değerler olmaması durumunda EKK tahmin ediciler kullanılır. Bu durumda S_p, C_p ye indirgenir. Veride sapan değerler olması ya da hataların normal dağılmaması durumunda Huber'in M-tahmin edicileri gibi bazı robust tahmin ediciler kullanılır. Alt model yeterli olduğunda $E(S_p) \cong p$ olur. Bu nedenle S_p nin değeri ne zaman p ye yaklaşır ise o zaman p regresörlü karşılık gelen kümeyi seçeriz.Grafiksel olarak, S_p nin değerlerinin p ye karşı grafikleri doğru alt küme seçiminde yardımcı olacaktır.

SAYISAL ÖRNEKLER

Bu bölümde, iki sayısal örneği göz önünde bulundurarak önerilen prosedürün performansını inceleyeceğiz. İlk örnekte Hald çimento verilerini ve ikinci örnekte simülasyon verilerini ele alacağız. Her iki örnekte de, bükülme

noktası c =1.345 olan Huber fonksiyonunu ψ(r) = max(-c, min(r, c)) kullanıyoruz ve r artıktır.

Örnek 4.6. Hald çimento verisi için:

İki durum için C_p, RC_p ve S_p'yi karşılaştırıyoruz:

Durum I: Gerçek veriler ve

Durum II: En küçük kareler uyumu için en büyük artığa sahip olan yanıt değişkeninin değerini $Y_6 = 200$ olarak değiştirilmesi. Sonuçlar Tablo 10'da verilmiştir.

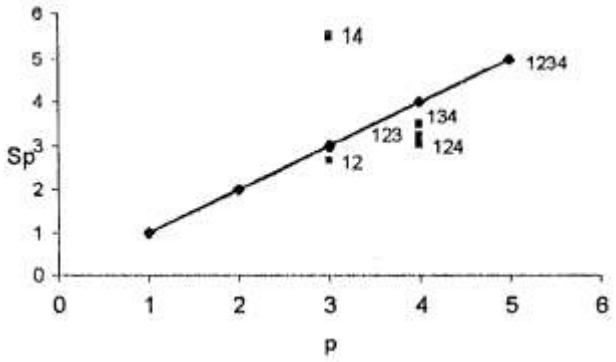
Tablo 10'dan ve Şekil 6'dan, Durum I'de bu üç kriterin iki değişkenin $\{X_1, X_2\}$ önemi üzerinde anlaştıkları açıktır.

Minimum RC_p $\{X_1, X_2, X_3\}$ alt kümesine karşılık gelirken C_p ve S_p, X_1 ve X_2 değişkenlerini içeren aynı alt kümeyi seçer. Öte yandan, Durum II için, tek bir etkili gözlemin C_p değerlerini nasıl büyük ölçüde etkilediği fark edilebilir.

Çizelge 4.10. Case I & II için Cp, RCp, Sp ve Case II için Cp*Değerleri

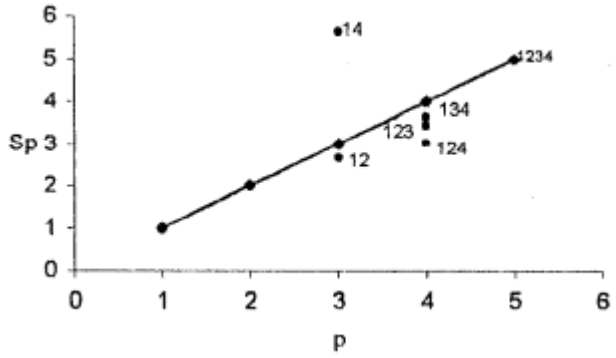
	Case I			Case II			
Model	C_p	RC_p	S_p	C_p	C_p^*	RC_p	S_p
(1)	203.00	142.00	202.64	1.00	1653.53	205.03	204.50
(2)	143.00	55.86	142.55	0.90	1692.80	65.07	142.66
(3)	315.00	200.00	315.31	2.30	1842.18	269.98	318.65
(4)	139.00	79.94	138.80	0.80	1651.14	101.25	137.85
(12)	2.68	0.64	2.67	1.10	1417.59	0.92	2.63
(13)	198.00	141.10	198.19	2.90	1644.54	199.11	199.98
(14)	5.50	2.73	5.49	1.10	1408.46	2.89	5.65
(23)	62.40	23.56	62.46	1.90	1554.64	26.16	63.98
(24)	138.00	58.67	138.29	2.80	1685.16	69.06	141.11
(34)	22.40	1.65	22.38	1.40	1483.28	2.27	23.83
(123)	3.00	−2.27	3.04	3.10	1419.26	−2.30	3.54
(124)	3.00	1.96	3.01	3.10	1412.33	2.13	3.02
(134)	3.50	2.87	3.49	3.10	1407.80	3.02	3.54
(234)	7.30	3.08	7.33	3.00	1399.37	3.11	7.60
(1234)	5.00	3.01	5.00	5.00	1416.71	2.96	5.00

Not: (j, k) sembolü, Xj ve Xk değişkenlerinin modelde olduğu anlamına gelir.



Şekil 4.6. Durum I (Hald Verileri) için Sp ve p.

S_p ve RC_p istatistikleri daha kararlıdır ve her biri daha önce Durum-I için seçilen aynı seti seçer. Her iki durumda da S_p kriteri $\{X_1, X_2\}$'yi, RC_p kriteri ise $\{X_1, X_2, X_3\}$'ü seçer. Ayrıca, RC_p kriterinin gereksiz yere daha büyük alt kümeleri seçtiği de görülebilir.



Şekil 4.7. Durum II (Hald Verileri) için Sp ve p.

Ayrıca, bir alt küme yeterli olduğunda, her iki durumda da S_p'nin değerinin p'ye yakın olduğuna dikkat edelim. Önceki kısımda bahsedildiği gibi, Mallows' C_p, aykırı değer olduğunda EKK yerine sağlam tahminler kullanılarak hesaplanır

ve bunlar C_p^* olarak etiketlenen sütunda verilir. C_p^*' nin saçma sonuçlar verdiği açıktır.

S_p ile p grafiksel olarak karşılaştırıldığında (Şekil 7), her iki durumda da uygun tahmin edici olabilecek dört model olduğu görülmektedir $\{X_1, X_2\}$, $\{X_1, X_2, X_3\}$, $\{X_1, X_3, X_4\}$ ve $\{X_1, X_2, X_4\}$. Açıkçası, en küçük S_p değerine sahip olduğu için $\{X_1, X_2\}$ içeren en basit modeli seçiyoruz, yani durum I, $S_p = 2.67$ ve durum II, $S_p = 2.63$.

Örnek 4.7. Simulasyon verileri için:

Bu örnekte, iki durum için C_p istatistiğini ve S_p istatistiğini karşılaştıracağız. 1) gerçek veriler ve 2) aykırı bir gözlemin tanıtılması $Y_{14} = 100$. Sonuçlar Tablo 11'de sunulmuştur.

Durum I'de, açıkçası her iki teknik de aynı Alt Kümeyi $\{X_1, X_2\}$ alır. II durumunda, S_p ve C_p farklı alt kümeleri alır, ancak S_p her iki durum için de aynı alt kümeyi seçer. Böylece, S_p'nin her iki durumda da bir alt kümeyi doğru seçtiği, C_p'nin bunu yapamadığı görülebilir.

UYARI. Yeni kriterin bir diğer avantajı (yüksek hızlı bilgisayar çağında olmasa da), RC_p'den farklı olarak daha az hesaplama gerektirmesidir. RC_p'de, W_p, U_p ve V_p sabitleri alt modelden alt modele değişir ve bu nedenle hesaplamalar giderek daha fazla zaman alır ve sıkıcı hale gelir.

Çizelge 4.11. Durun I ve Durum II için Cp ve Sp Değerleri

	Case I		Case II	
Set	C_p	S_p	C_p	S_p
(1)	460.00	433.22	4.7	507.87
(2)	362.00	349.64	4.5	410.68
(3)	124.00	117.27	1.1	142.97
(12)	3.80	3.87	2.1	4.23
(13)	103.00	96.19	2.8	122.60
(23)	108.00	104.13	3.0	124.82
(123)	4.00	4.00	4.0	4.00

Not: (j, k) sembolü, Xj ve Xk değişkenlerinin modelde olduğu anlamına gelir.

σ^2 TAHMİN EDİCİ SEÇİMİ

Aykırı gözlem durumunda, β için Huber M-tahmin edicisini kullandık. Daha önce belirtildiği gibi varyans içinde sağlam tahmin edicilerden birini kullanmalıyız. Bu tür çeşitli tahmin ediciler mevcut olduğundan, bu bölümde bu nedenle bunların S_p değerleri üzerindeki etkilerini inceleyeceğiz.

$\tilde{\beta}$, β'nin Huber M kestiricisini ve r_i, tarafından verilen i'nci kalıntıyı göstersin.

$$r_i = Y_i - X_i' \tilde{\beta}, \quad i = 1, 2, ..., n$$

$\tilde{\beta}$ ve W_i ağırlığının hesaplanması için hesaplama prosedürü Ek'te verilmiştir. Spesifik olarak, aşağıdaki σ^2 tahmin edicilerini dikkate alıyoruz:

$$I : \hat{\sigma}_1^2 = [1.4826 \; med \; |r_i - med(r_i)|]$$

(Bu tahmin edici önceki örneklerde kullanılmıştır)

$$II : \hat{\sigma}_2^2 = \left[med\,|r_i| \right]^2$$

$$III : \hat{\sigma}_3^2 = \sum_{i=1}^{n} W_i^2 r_i^2 / (n-k)$$

$$IV : \hat{\sigma}_4^2 = Z_k / U_k$$

Çizelge 4.12. Hald ve Abt Verileri için $\hat{\sigma}_i^2$ değerleri

Data	$\hat{\sigma}_1^2$	$\hat{\sigma}_2^2$	$\hat{\sigma}_3^2$	$\hat{\sigma}_4^2$	$\hat{\sigma}_5^2$
Hald data	6.36	5.11	5.35	7.60	6.41
Abt data	0.49	0.42	0.42	0.47	0.70

Çizelge 4.13. Hald Verileri için Farklı σ2 Tahmincisi Kullanan Sp Değerleri

Set	$\hat{\sigma}_1^2$	$\hat{\sigma}_2^2$	$\hat{\sigma}_3^2$	$\hat{\sigma}_4^2$	$\hat{\sigma}_5^2$
(1)	204.50	239.83	228.77	160.81	191.04
(2)	142.66	167.36	159.62	112.66	133.25
(3)	318.65	373.59	356.39	250.69	297.71
(4)	137.85	161.72	154.25	108.33	128.76
(12)	2.63	2.91	2.82	2.28	2.52
(13)	199.98	234.18	223.47	157.67	186.94
(14)	5.65	6.45	6.20	4.66	5.35
(23)	63.98	74.80	71.41	50.59	59.85
(24)	141.11	165.20	157.65	111.32	131.93
(34)	23.83	27.75	26.52	18.97	22.33
(123)	3.54	3.63	3.60	3.42	3.50
(124)	3.02	3.03	3.03	3.02	3.02
(134)	3.54	3.63	3.60	3.42	3.50
(234)	7.60	8.39	8.14	6.62	7.30
(1234)	5.00	5.00	5.00	5.00	5.00

Not: (j,k) nın anlamı modelde j-inci ve k-ıncı değişkenin yer almasıdır

Burada $Z_k = \sum W_i^2 (\hat{Y}_{ik} - Y_i)^2$ ve $U_k = \sum (W_i r_i)$.

U_k hesaplama ifadesi Ronchetti ve Staudte (1994) tarafından verilmiştir.

$$V:\hat{\sigma}_5^2 = h^2\sum_{i=1}^{n}\psi(r_i/k_1)^2k_1^2 \Big/ \left[(n-k)\left[\sum_{i=1}^{n}\psi'(r_i/k_1)/n\right]^2\right].$$

Burada $h = 1+\frac{k(1-m)}{nm}$, m $-c < r_i < c$ 'yi sağlayan artıkların bağıl frekansıdır ve $k_1 = \left[1.4826\ med\left|r_i - med(r_i)\right|\right]$ ölçek parametresidir . Burada, $c = 1.345$ bükülme noktasına sahip Huber fonksiyonu ile regresyon parametrelerini tahmin etmek için klasik Huber tipi M-tahmin edicisini kullanılmıştır. Bir önceki bölümde ele alınan $Y_6 = 200$ aykırı değere sahip Hald verileri ve $Y_{14} = 100$ aykırı değere sahip Abt verileri için hesaplanmıştır.

Çizelge 4.14. Abt Verileri için Farklı σ2 Tahmincisi Kullanan Sp Değerleri

Set	$\hat{\sigma}_1^2$	$\hat{\sigma}_2^2$	$\hat{\sigma}_3^2$	$\hat{\sigma}_4^2$	$\hat{\sigma}_5^2$
(1)	507.87	380.93	592.59	528.83	357.74
(2)	410.68	308.04	479.19	427.63	289.28
(3)	142.97	107.24	166.82	148.87	100.71
(12)	4.23	3.67	4.60	4.32	3.57
(13)	122.60	92.45	142.71	127.57	86.95
(23)	124.82	94.12	145.31	129.89	88.51
(123)	4.00	4.00	4.00	4.00	4.00

Hald verileri için çeşitli $\hat{\sigma}_i^2$ için S_p, Tablo 13'te ve Abt verileri için Tablo 14'te verilmiştir.

Tablo 13'ten, S_p-nin tüm $\hat{\sigma}_i^2$ için aynı $\{X_1, X_2\}$ alt kümesini seçtiği açıktır. İkinci örnekte, S_p, σ^2'nin tüm tahmin edicileri için aynı $\{X_1, X_2\}$ alt kümesini seçer. Hesaplama açısından bakıldığında, $\hat{\sigma}_4^2$ ve $\hat{\sigma}_5^2$ yi elde etmek için

yapılan hesaplamaların sıkıcı ve zaman alıcı olduğunu , $\hat{\sigma}_1^2$ için nispeten daha basit olduğunu gözlemliyoruz. . Bu nedenle S_p hesaplamasında $\hat{\sigma}_1^2$ kullanılmıştır.

4.3. Çoklu İç İlişki ve Sapan Değerler Olması Durumu

Sapan değerler olması durumunda EKK tahmin ediciye alternatif olarak M-tahmin edicilerin, çoklu iç ilişki olması durumunda ridge tahmin edici gibi bazı yanlı tahmin edicilerin tanımlandığını biliyoruz. Bu durumda Jadhav ve ark.(2014)'te, yukarıda ele aldığımız S_p kriterinin genelleştirilmiş versiyonunu tanımlamışlardır. Daha sonra Jacknifed Ridge - M (JRM) tahmin ediciye bağlı model seçimini örnek ve simülasyon çalışması ile incelemişlerdir.

(2.29.) da verilen JRM tahmin ediciyi, $\hat{\beta}_{JRM}(k) = \left[I - k^2\varphi'(X'X + kI)^{-2}\varphi \right]^{-2} \hat{\beta}_M = R\hat{\beta}_M$, göz önüne alırsak, $\hat{y}_k = X\hat{\beta}_{JRM}(k) = Hy$ şeklinde verilir. Burada $H = XR(X'WX)^{-1}X'W$ tam modele dayanan öngörü matrisidir.Tam model $k-1$ bağımsız değişken içermektedir. (4.1.) de verilen model $y = X_A\beta_A + X_B\beta_B + \varepsilon$ şeklinde yazılabilir. Burada, $X = \left[X_A : X_B \right]$ ve $\beta' = \left[\beta_A' : \beta_B' \right]$ şeklinde parçalanmıştır. Burada X_A, nxp tipinde birinci kolonunda 1'lerin olduğu bir matris ve X_B, $nx(k-p)$ tipinde bir matristir. β_A ve β_B sırasıyla $px1$ ve $(k-p)x1$ tipinde vektörlerdir. $p-1$ açıklayıcı değişken içeren $y = X_A\beta_A + \varepsilon$ alt küme modelini göz önüne alalım. $\hat{\beta}_{AJRM}$ nin β_A nın JRM tahmin edicisi olduğunu varsayalım. Bu durumda $\hat{y}_p = X_A\hat{\beta}_{AJRM} = H_1 y$ ve $H_1 = X_A R_A (X_A' W_A X_A)^{-1} X_A' W_A$ olur. $\hat{y}_k$ ve $\hat{y}_p$ öngörü vektörlerine bağlı olarak S_p kriterinin genelleştirilmiş versiyonunu aşağıdaki gibi tanımlanır:

$$GS_p = \sum_{i=1}^{n} \frac{(\hat{y}_{ik} - \hat{y}_{ip})^2}{\sigma^2} - tr\left[(H - H_1)'(H - H_1)\right] + p \qquad (4.16.)$$

Burada σ bilinmeyen hata varyansı ve uygulamada yerine JRM tahmin ediciye dayanan uygun bir tahmini yerleştirilir. Bu tanımda motivasyon şudur:

$\hat{y}_{ik}$, $\hat{y}_{ip}$ ye yakın olduğu zaman $\sum_{i=1}^{n} \frac{(\hat{y}_{ik} - \hat{y}_{ip})^2}{\sigma^2}$ değeri küçük olacaktır yani $p-1$ açıklayıcı değişken içeren modele bağlı öngörü, $k-1$ açıklayıcı değişken içeren modele bağlı olan öngörü kadar doğru olacaktır. $\sum_{i=1}^{n} \frac{(\hat{y}_{ik} - \hat{y}_{ip})^2}{\sigma^2}$ değeri büyük olduğunda $\hat{y}_{ik}$, $\hat{y}_{ip}$ den uzaklaşacak ve bu da $p-1$ açıklayıcı değişken içeren modele bağlı öngörünün, $k-1$ açıklayıcı değişken içeren modele bağlı olan öngörü kadar doğru olmamasına neden olacaktır.

Bazı Sonuçlar:

Sonuç 4.4. Altküme modeli yeterli ise

$$E(\sum_{i=1}^{n} (\hat{y}_{ik} - \hat{y}_{ip})^2) \cong \sigma^2 tr\left[(H - H_1)'(H - H_1)\right].$$

Sonuç 4.5. Altküme modeli yeterli ise $E(GS_p) \cong p$

Sonuç 4.6. β 'yı tahmin etmek için EKK tahmin edici kullanıldığı zaman GS_p , Mallow'sun C_p istatisiğine indirgenir.

Sonuç 4.7. β 'yı tahmin etmek için $\hat{\beta}_M$ kullanıldığı zaman GS_p , S_p istatisiğine indirgenir.

Sonuç 4.8. β'yı tahmin etmek için ridge tahmin edici kullanıldığı zaman GS_p, $R_p(\lambda)$ istatisiğine indirgenir.

GS_p istatistiğine dayalı alt küme seçme algoritması için aşağıdaki adımlar uygulanır.

Adım 1: Tüm mümkün alt kümeler için GS_p istatistiğini hesapla

Adım 2: GS_p nin değeri p ye yakın olacak şekilde minimum hacimli alt kümeyi seç

Şimdi C_p, S_p, R_p ve GS_p istatistiklerini kullanarak model seçimini simülasyon çalışması ve örneklerle ele alabiliriz.

Örnek 4.8.

Önerilen yöntemin performansını göstermek için bir simülasyon çalışması yapılmıştır. Simülasyon çalışması üç alt kısma ayrılmıştır. Birinci kısımda, C_p, S_p, R_p ve GS_p kriterlerinin performansı aykırı değerin yokluğu ve varlığı ile çoklu bağlantının tüm kombinasyonları için gösterilmiştir. Bu kriterlerin doğru bir model seçme yeteneği ikinci kısımda değerlendirilmektedir. Ayrıca, σ^2 tahmin edicisinin çeşitli seçenekleri üçüncü kısımda ele alınmış ve performansları incelenmiştir.

M-tahmin edicisinin hesaplanması için Huber'in sağlam kriter fonksiyonu kullanılır. $\rho(\cdot)$ biçimi

$$\rho(z)=\begin{cases} \dfrac{1}{2}z^2 & |z|\le t \\ |z|t-\dfrac{1}{2}t^2 & |z|>t \end{cases}$$

burada z, ölçeklenmiş rezidülerdir ve t = 1.345.

Bu alt bölümde dört durumu ele aldık: 1. Temiz veri, 2. Aykırı değer içeren veri, 3. Çoklu bağlantıya sahip veri ve 4. Aykırı ve çoklu bağlantıya sahip veri. İlk iki durum için C_p, S_p, R_p ve GS_p istatistiklerinin performansı Örnek 4.8.1 de gösterilmiştir. Örnek 4.8.2 ise, C_p, S_p, R_p ve GS_p istatistiklerinin doğru alt küme seçim performansını incelemek için son iki durumu ele almaktadır.

Örnek 4.8.1. Burada, Y yanıt değişkeni üzerinde n = 20 gözlem üretmek için aşağıdaki regresyon modeli ele alınmıştır.

$$Y_i = 1 + 2X_{i1} + 3X_{i2} + 0X_{i3} + 0X_{i4} + \varepsilon_i, \quad i = 1, 2, ..., n$$

burada $X_{ij} \sim U(0,1)$, $i = 1, 2, ..., n$ $j = 1, 2, 3, 4$ ve $\varepsilon_i \sim N(0, 0.25)$. Jadhav ve Kashid (2014) tarafından önerilen yanıt değişkeninde aykırı değeri tanıtmak için bir yöntem uygulanmaktadır. Tek bir aykırı gözlem, en büyük mutlak kalıntıya karşılık gelen yanıt değişkenine şu şekilde dahil edilir:

Çizelge 4.15. Tüm alt küme modelleri için Cp, Sp, Rp ve GSp istatistiklerinin değerleri.

modeldeki regresörler	Temiz veri				Bir aykırı veri ile			
	C_p	S_p	R_p	GS_p	C_p	S_p	R_p	GS_p
X_1	41.4566	88.8442	38.9205	89.5066	6.4856	88.9771	3.9139	89.2078
X_2	21.8913	59.5440	20.1670	59.8542	3.7951	63.7876	2.8164	63.7516
X_3	55.3685	127.1283	57.0563	128.0798	4.5724	129.3113	3.0506	129.9719
X_4	59.6552	143.6816	94.6994	144.0413	6.8399	143.4294	4.1313	143.5348
$X_1\ X_2$	2.7568	3.2767	2.9584	3.1723	4.7240	3.3099	3.5026	2.9617
$X_1\ X_3$	42.1485	86.6213	38.7884	87.3354	6.2909	88.7644	3.9701	88.9678
$X_1\ X_4$	43.2062	91.3802	39.9829	92.1708	7.1991	91.0399	4.5070	91.3607
$X_2\ X_3$	19.6384	48.3448	17.5928	48.4336	2.1995	53.3889	2.5820	53.1513
$X_2\ X_4$	19.9072	57.8829	18.1727	57.8715	4.6941	65.2842	3.6406	64.9878
$X_3\ X_4$	54.2756	125.7806	51.6828	126.7200	6.1055	127.9752	3.9853	128.6459
$X_1\ X_2\ X_3$	4.3852	3.6169	4.3647	3.5583	4.0004	3.6296	3.8108	3.5170
$X_1\ X_2\ X_4$	3.6821	5.1481	3.8975	5.0931	5.0321	5.3201	4.3798	5.1078
$X_1\ X_3\ X_4$	43.5963	89.3993	39.5378	90.3715	7.5577	90.4398	4.8287	90.5849
$X_2\ X_3\ X_4$	16.6383	44.2775	14.8272	44.3727	3.4836	51.8037	3.824	51.2683
$X_1\ X_2\ X_3\ X_4$	5.0000	5.0000	5.0000	5.0000	5.0000	5.0000	5.0000	5.0000

Y 'nin gerçek değerinin yirmi ile çarpılması (en büyük mutlak kalıntı Y_{19} 'a karşılık gelir ve 20 ile çarpıldıktan sonraki gerçek değeri ve aykırı değer sırasıyla 4.73744 ve 94.7488'dir). C_p, S_p, R_p ve GS_p istatistiklerinin değerleri tüm olası alt küme modelleri için hesaplanır ve Tablo 15'te sunulur.

$\{X_1, X_2\}$ alt kümesi için C_p, S_p, R_p ve GS_p istatistiklerinin değerlerinin temiz veri için p(=3)'e yakın olduğu açıktır. Bu, dört yöntemin de iki regresör değişkeni X_1 ve X_2 'nin önemi üzerinde anlaştığını ve doğru alt küme modelini seçtiğini gösterir. Aykırı durum için, C_p ve R_p istatistiklerinin yanlış alt kümeyi veya birden fazla alt küme modelini seçtiği görülüyor. Bu her iki kriter de doğru alt küme modelini seçemez. Buna karşılık, S_p ve GS_p istatistiklerinin temiz veriler için seçilen aynı alt kümeyi seçtiğini görebiliriz. Bu nedenle, S_p ve GS_p istatistikleri, verilerde aykırı değerlerin varlığına karşı dayanıklıdır (Jadhav, Kashid ve Kulkarni, 2014).

Örnek 4.8.2. Bu örnekte, çoklu bağlantı derecesi, $\rho = 0.999$ olarak ayarlanır ve aşağıdaki regresyon modeli kullanılarak yanıt değişkeni üzerinde n = 30 gözlem üretilir.

$$Y_i = 15 + 5X_{i1} + 5X_{i2} + 0X_{i3} + 0X_{i4} + \varepsilon_i,$$

burada $\varepsilon_i \sim N(0, 1)$, i = 1, 2, . . . , 30. Örnek 4.8.1'de kullanılan senaryonun aynısı, yanıt değişkeninde bir aykırı değer gözlemi sağlamak için uygulanmaktadır. Her bir regresör değişkenine karşılık gelen VIF'ler 405.0179, 441.0012, 373.2567 ve 509.0688'dir. Tüm olası alt küme modelleri için C_p, S_p, R_p ve GS_p istatistiklerinin değerleri, bir aykırı değer olmadan ve bir aykırı durumla birlikte hesaplanır ve Tablo 16'da rapor edilir.

Tablo 16, verilerde çoklu bağlantı varlığında, R_p ve GS_p istatistiklerinin $\{X_1, X_2\}$ alt kümesi için her ikisinin de kriteri karşıladığını, C_p ve S_p istatistiklerinin ise farklı boyutlardaki birkaç alt küme için kriteri karşıladığını ve bu nedenle kesin bir sonucun olmadığını göstermektedir. Başka bir deyişle, doğru altkümeyi seçemezler.

Aykırı değerin yanı sıra çoklu bağlantı varlığında, R_p istatistiği birkaç alt küme seçer, GS_p istatistiği ise $\{X_1, X_2\}$ olarak adlandırılan alt kümeyi doğru seçer. Dolayısıyla bu çalışma, aykırı değerler ve çoklu bağlantı gibi veri anomalilerinin varlığında GS_p istatistiğinin performansının rakiplerinden daha iyi olduğunu göstermektedir.

Ayrıca, birkaç aykırı değer varlığında GS_p istatistiğinin performansını değerlendirmek için yukarıdaki veri setini ele alınmıştır. Yanıt değişkenine iki ve üç aykırı gözlem dahil edilerek, C_p, S_p, R_p ve GS_p istatistiklerinin doğru model seçim performansı değerlendirilmiştir. Sonuçlar Tablo 17'de sunulmuştur.

Tablo 17, bir aykırı değer durumunda olduğu gibi benzer sonuçları gösterir, yani, GS_p istatistiği doğru alt küme modelini $\{X_1, X_2\}$ seçer, C_p ve S_p istatistikleri ise kriteri karşılayan birkaç alt küme seçer ve böylece sonuçsuz karara yol açar. Birden fazla alt küme modeline karşılık gelen R_p istatistiğinin değerleri p'ye yakındır. Bu nedenle, GS_p istatistiğinin performansı, verilerde çoklu doğrusal bağlantı ve birden fazla aykırı gözlem varlığında diğerlerinden daha iyidir.

Çizelge 4.16. Çoklu bağlantı varlığında tüm alt küme modelleri için Cp, Sp, Rp ve GSp değerleri

modeldeki regresyonlar	Sapan değer olmadan				Bir sapan değerle			
	C_p	S_p	R_p	GS_p	C_p	S_p	R_p	GS_p
X_1	3.0624	4.0741	4.8225	4.0673	0.0868	4.0757	1.9696	4.0402
X_2	4.7698	7.1797	5.8701	7.3400	−0.0061	7.1948	1.9213	7.3243
X_3	13.7610	16.7849	13.5605	16.9714	0.2235	16.7863	2.0500	16.8117
X_4	16.9498	21.3865	12.9645	20.2555	0.0976	21.3939	1.9813	20.1864
$X_1 X_2$	2.7505	3.5041	3.4738	3.2757	1.7151	3.5189	2.8865	3.2695
$X_1 X_3$	3.0103	3.1471	3.7453	3.6414	1.7867	3.1461	2.8929	3.6164
$X_1 X_4$	4.4605	5.2065	4.3467	4.8318	2.0868	5.2072	2.7099	4.7966
$X_2 X_3$	3.9440	4.8946	4.1756	4.7671	1.1404	4.8960	2.8873	4.7696
$X_2 X_4$	6.2422	8.2382	5.5534	7.5530	1.7710	8.2538	2.8762	7.5054
$X_3 X_4$	14.0713	16.7995	11.4287	15.7780	1.5284	16.7907	2.9295	15.7992
$X_1 X_2 X_3$	3.7495	3.8968	3.4489	3.7146	3.1025	3.8993	3.9311	3.7153
$X_1 X_2 X_4$	4.7441	5.4564	4.2877	4.9393	3.6141	5.4687	3.9191	4.9174
$X_1 X_3 X_4$	4.8445	4.9141	4.9912	5.5923	3.4331	4.9156	3.9596	5.5801
$X_2 X_3 X_4$	5.3831	6.2449	5.5784	6.6812	3.0479	6.2465	3.9249	6.6903
$X_1 X_2 X_3 X_4$	5.0000	5.0000	5.0000	5.0000	5.0000	5.0000	5.0000	5.0000

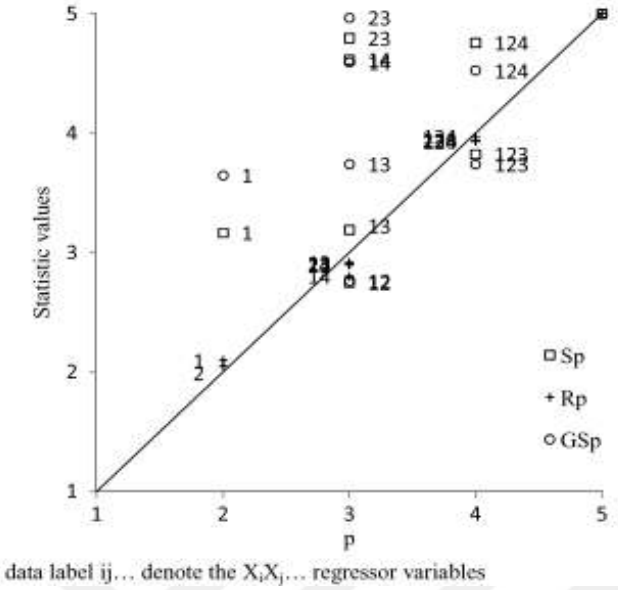
Çizelge 4.17. Çoklu doğrusal bağlantı ve birden fazla aykırı değer varlığında tüm alt küme modelleri için Cp, Sp, Rp ve GSp değerleri.

modeldeki regresyonlar	İki sapan değerli				Üç sapan değerli			
	C_p	S_p	R_p	GS_p	C_p	S_p	R_p	GS_p
X_1	0.2189	3.1665	2.0991	3.6478	−0.8819	3.1855	2.2393	3.6339
X_2	0.1246	5.7824	2.0488	6.8607	−0.8933	5.7982	2.2342	6.8489
X_3	0.4238	17.1435	2.2255	18.4844	−0.7931	17.1703	2.2813	18.8167
X_4	0.2749	20.3236	2.1399	20.8400	−0.7848	20.3956	2.2828	20.4596
$X_1 X_2$	1.9256	2.7502	2.8969	2.7616	1.1066	2.7400	2.9518	2.7213
$X_1 X_3$	1.7165	3.1903	2.9178	3.7398	1.0853	3.1907	2.9094	3.7252
$X_1 X_4$	2.1909	4.6171	2.8028	4.5929	1.0607	4.6175	2.9481	4.5601
$X_2 X_3$	1.0655	4.7947	2.9047	4.9666	1.0642	4.8012	2.9204	4.9436
$X_2 X_4$	1.7747	7.2267	2.9031	7.2986	1.0137	7.2338	2.9766	7.3958
$X_3 X_4$	1.6978	16.9648	2.9623	17.1498	1.2064	16.9686	3.0786	17.0818
$X_1 X_2 X_3$	3.0614	3.8247	3.9376	3.7411	3.0550	3.8260	3.9361	3.7317
$X_1 X_2 X_4$	3.7025	4.7580	3.9313	4.5242	3.0000	4.7454	3.9879	4.4937
$X_1 X_3 X_4$	3.4453	5.0073	3.9738	5.6390	3.0607	5.0082	3.9730	5.7065
$X_2 X_3 X_4$	3.0068	6.2014	3.9417	6.7953	3.0130	6.2008	3.9974	6.7293
$X_1 X_2 X_3 X_4$	5.0000	5.0000	5.0000	5.0000	5.0000	5.0000	5.0000	5.0000

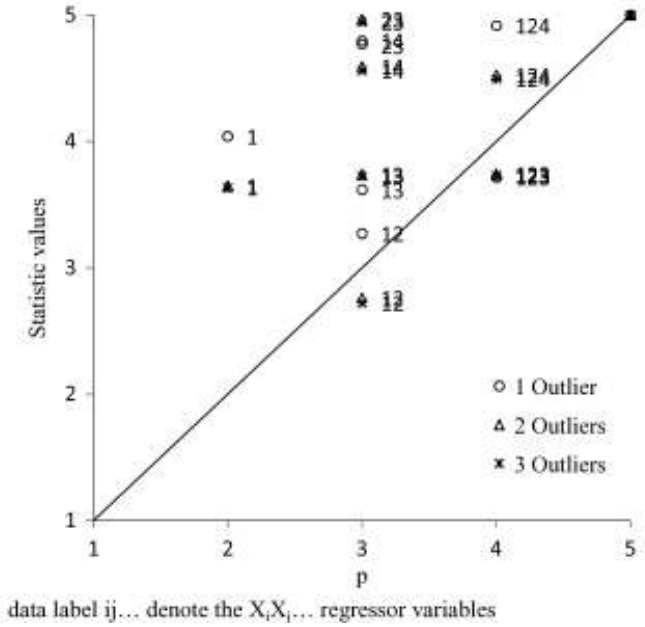
Grafiksel temsil, yukarıdaki gerçekleri açıkça ortaya koymaktadır. Bu amaçla Tablo 17'de verilen çoklu doğrusal bağlantı gösteren ve iki aykırı değere sahip veri seti ele alınmıştır. Alt küme modellerine karşılık gelen S_p, R_p ve GS_p istatistiklerinin değerleri Şekil 8'de gösterilmektedir. Ayrıca Tablo 16 ve 17'den bir aykırı değer, iki aykırı değer ve üç aykırı değer durumuna karşılık gelen alt küme modelleri için GS_p istatistiklerinin değerleri elde edilir ve Şekil 9'da p'ye karşı çizilir. Bu rakamlar, GS_p istatistiğinin, yanıt değişkeninde farklı sayıda aykırı değerle çoklu bağlantı sorunu için tutarlı bir şekilde doğru alt küme modelini seçtiğini açıkça gösterir (Jadhav, Kashid ve Kulkarni, 2014).

Model seçme yeteneği

Bu alt bölümde, C_p, S_p, R_p ve GS_p istatistiklerinin doğru model seçme yeteneği incelenmiştir. McDonald ve Galarneau (1975) tarafından verilen bir simülasyon tasarımı kullanılmıştır.



Şekil 4.8. Sp, Rp ve GSp istatistiklerinin p'ye karşı değerleri.



Şekil 4.9. GSp istatistiğinin p'ye karşı değerleri

Aşağıdaki regresyon modellerinin X_1 ve X_2 açıklayıcı değişkenlerinde çoklu bağlantı olması durumu:

$$Model\ I\ \ Y_i = 5 + 4X_{i1} + 0X_{i2} + 3X_{i3} + 0X_{i4} + \varepsilon_i, \quad i = 1,2,...,30$$
$$Model\ II\ Y_i = 5 + 2X_{i1} + 0X_{i2} + 4X_{i3} + 0X_{i4} + 3X_{i5} + \varepsilon_i, \quad i = 1,2,...,50.$$

Çizelge 4.18. Cp, Sp, Rp ve GSp istatistiklerinin model seçim yeteneği

	ρ	Sapan değer olmadan				Bir sapan değer ile			
		0.6	0.7	0.8	0.9	0.6	0.7	0.8	0.9
Model I	$\sigma^2 = 0.25$								
	C_p	72	81	74	71	23	24	20	18
	S_p	63	66	63	62	63	68	62	64
	R_p	72	79	72	71	25	23	20	21
	GS_p	63	66	64	63	63	68	64	65
	$\sigma^2 = 1$								
	C_p	70	74	74	71	17	19	19	18
	S_p	61	63	64	62	65	60	63	59
	R_p	69	74	73	70	18	21	22	17
	GS_p	62	63	64	63	65	62	64	61
Model II	$\sigma^2 = 0.25$								
	C_p	81	74	76	79	16	24	21	20
	S_p	75	67	69	72	59	65	70	63
	R_p	81	74	75	77	17	29	21	20
	GS_p	76	68	69	73	62	66	72	66
	$\sigma^2 = 1$								
	C_p	63	74	74	79	22	21	19	19
	S_p	62	67	62	72	57	63	63	63
	R_p	61	74	74	79	24	25	18	23
	GS_p	63	69	62	74	62	65	64	67

Model I $(X_3 ve\ X_4)$ ve Model II'de $(X_3 ve\ X_5)$ kalan regresör değişkenleri standart normal dağılımdan üretilir. Model II'yi daha karmaşık hale getirmek için, X_4 regresör değişkeni, X_2 ve X_3 regresör değişkeninin bir ürünü olarak alınır. Hata değişkeni ε_i, i = 1, 2, . . . , n, sıfır ortalama ve $\sigma^2 = 0.25$ ve $\sigma^2 = 1$ varyansı olan normal dağılımdan üretilir. ρ = 0.6, 0.7, 0.8 ve 0.9 ayarlanarak farklı çoklu bağlantı dereceleri elde edilir. Yanıt değişkeninde sapan değer gözlemini tanıtmak için örnek 4.8.1'de kullanılan senaryo izlenir.

Simülasyon deneyi, çoklu doğrusallık derecesi (ρ), hata varyansı (σ^2) ve aykırı değer durumu olan ve olmayan tüm kombinasyonlar için her model için 100 kez tekrarlanır. Tüm olası alt küme modelleri için C_p, S_p, R_p ve GS_p istatistiklerinin değerleri hesaplanır. C_p, S_p, R_p ve GS_p istatistiklerinin doğru alt küme modelini seçme sayısı Tablo 18'de sayılır ve raporlanır.

Tablo 18, aykırı değeri olmayan durum için, C_p ve R_p istatistiklerinin doğru model seçim sıklığının S_p ve GS_p istatistiklerinden daha büyük olduğunu göstermektedir. Ancak, tek aykırı durum için, GS_p istatistiğinin doğru model seçme yeteneği, hem model I hem de model II için C_p ve R_p istatistiklerininkinden daha büyüktür. Bir aykırı durum için, ρ değeri büyük olduğunda, GS_p istatistiği tarafından doğru alt küme modeli seçiminin yüzdesi, S_p istatistiğininkinden daha büyüktür (Jadhav, Kashid ve Kulkarni, 2014).

σ^2 tahmin edicisinin seçimi

GS_p istatistiğinin hesaplanmasında, bir ölçek parametresi σ^2 kullanıyoruz. Bilinmediği için uygun tahmin edicisini kullanmamız gerekiyor. Literatürde σ^2'nin çeşitli tahmin edicileri mevcuttur. GS_p kriterinin performansını incelemek için, JRM tahmin edicisine dayanan dört farklı σ^2 tahmin edicisi türü kullanıyoruz. β'nın JRM tahmin edicisine dayalı i'nci rezidü r_i olsun ve şu şekilde tanımlanır:

$$r_i = Y_i - X_i^{'} \hat{\beta}_{JRM}, \quad i = 1, 2, ..., n.$$

Aşağıdaki σ^2 tahmin edicisini ele alıyoruz

$$1.\hat{\sigma}_1^2 = \left[1.4826 Median\left|r_i - Median(r_i)\right|\right]^2$$

$$2.\hat{\sigma}_2^2 = \left[1.4826 Median\left|r_i\right|\right]^2$$

$$3.\hat{\sigma}_3^2 = \sum_{i=1}^{n} r_i^2 w_i^2 / (n-k)$$

$$4.\hat{\sigma}_4^2 = \sum_{i=1}^{n} r_i^2 w_i^2 / (\sum_{i=1}^{n} \left|r_i w_i\right|).$$

Çizelge 4.19. Farklı σ2 tahminlerine sahip tüm alt küme modelleri için GSp istatistiğinin değerleri

modeldeki regresyonlar	Bir sapan değerli				İki sapan değerli			
	$\hat{\sigma}_1^2$ (0.9881)[a]	$\hat{\sigma}_2^2$ (0.8478)	$\hat{\sigma}_3^2$ (0.8112)	$\hat{\sigma}_4^2$ (0.9949)	$\hat{\sigma}_1^2$ (0.9881)	$\hat{\sigma}_2^2$ (0.8478)	$\hat{\sigma}_3^2$ (0.8050)	$\hat{\sigma}_4^2$ (0.9901)
X_1	337.5378	393.5145	411.3049	335.2256	336.6605	392.5125	413.4319	336.0047
X_2	102.7412	119.8496	125.2869	102.0346	103.7496	121.0292	127.5012	103.5467
X_3	615.8711	717.9162	750.3480	611.6561	615.6479	717.6772	755.8924	614.4499
X_4	124.4055	145.1017	151.6793	123.5507	125.0403	145.8491	153.6430	124.7960
$X_1 X_2$	2.9559	3.2151	3.2974	2.9452	2.9414	3.2006	3.2977	2.9384
$X_1 X_3$	314.6453	366.4917	382.9694	312.5038	313.0681	364.6737	384.0026	312.4621
$X_1 X_4$	12.9300	14.8421	15.4498	12.8510	12.9644	14.8887	15.6094	12.9418
$X_2 X_3$	43.3745	50.3237	52.5323	43.0875	43.6117	50.6038	53.2227	43.5296
$X_2 X_4$	103.6550	120.5901	125.9723	102.9555	104.5139	121.5935	127.9907	104.3133
$X_3 X_4$	11.7165	13.4261	13.9695	11.6459	11.6709	13.3807	14.0210	11.6508
$X_1 X_2 X_3$	3.5522	3.5621	3.5653	3.5518	3.5441	3.5539	3.5576	3.5440
$X_1 X_2 X_4$	3.8178	3.8903	3.9134	3.8148	3.8174	3.8897	3.9168	3.8165
$X_1 X_3 X_4$	6.9981	7.5776	7.7618	6.9741	7.0097	7.5940	7.8129	7.0029
$X_2 X_3 X_4$	11.7398	13.1290	13.5705	11.6825	11.7104	13.0941	13.6124	11.6941
$X_1 X_2 X_3 X_4$	5.0000	5.0000	5.0000	5.0000	5.0000	5.0000	5.0000	5.0000

a Parantez içindeki rakamlar σ_i^2, i = 1, 2, 3, 4'ün tahminini gösterir.

Bu tahmin edicilerin performansı, simülasyon ile üretilmiş bir örnek kullanılarak gösterilmiştir. Burada, X_1, X_2, X_3 ve X_4 üzerinde $N_4(0, \Sigma)$'den n = 30 büyüklüğünde rastgele bir örnek oluşturulur.

$$\sum \begin{bmatrix} 1 & 0.531 & -0.850 & -0.531 \\ 0.531 & 1 & -0.401 & -0.972 \\ -0.850 & -0.401 & 1 & 0.288 \\ -0.531 & -0.972 & 0.288 & 1 \end{bmatrix}$$

Yanıt değişkeni üzerinde n = 30 gözlem ürettiği düşünülen bir regresyon modeli şu şekilde verilmektedir:

$$Y_i = 5 + 2X_{i1} + 4X_{i2} + 0X_{i3} + 0X_{i4} + \varepsilon_i, \quad i = 1, 2, ..., 30,$$

burada $\varepsilon_i \sim N(0, 1)$. Her terimin VIF'leri 31.0458, 182.9686, 33.1639 ve 210.6130'dur. Bir ve iki aykırı gözlem, Örnek 4.8.1'de verilen prosedürün aynısı kullanılarak yanıt değişkenine dahil edilmiştir. Simüle edilen verilere dayanarak, dört farklı σ^2 tahmin edicisi ile bir aykırı ve iki aykırı gözlem durumu için GS_p istatistiğinin değerleri elde edilmiş ve Tablo 19'da sunulmuştur.

Tablo 19'dan, GS_p istatistiğinin, σ^2'nin tüm tahmin edicileri için bir aykırı değer ve iki aykırı değer durumu ile çoklu bağlantı varlığında aynı alt kümeyi $\{X_1, X_2\}$ seçtiği açıktır. $\hat{\sigma}_1^2$ ve $\hat{\sigma}_4^2$ değerleri, $\hat{\sigma}_2^2$ ve $\hat{\sigma}_3^2$ değerlerine kıyasla σ^2'nin gerçek değerine yakındır. Sonuç olarak, $\hat{\sigma}_1^2$ ve $\hat{\sigma}_4^2$ kullanılarak doğru altküme modeline karşılık gelen GS_p istatistiğinin değeri, $\hat{\sigma}_2^2$ ve $\hat{\sigma}_3^2$ ile karşılaştırıldığında p'ye yakındır (Jadhav, Kashid ve Kulkarni, 2014).

5. SONUÇ VE ÖNERİLER

Tez çalışmasında ele aldığımız metodlar, Üçüncü bölümde tanımlanmış olan Dayanıklı Liu tahmin edici, Dayanıklı Genelleştirilmiş Liu ve Hemen Hemen Yansız Genelleştirilmiş Liu tahmin edici gibi tahmin edicilere uygulanacaktır.Ayrıca Dördüncü bölümde ele alınan yöntemler hem Liu-tipi tahmin edicilere hemde diğer alternatif tahmin edicilere uygulanacaktır.

KAYNAKLAR

Akaike, H. 1974. A new look at the statistical model identification. IEEE Transactions on Automatic Control 19: 716–723.

Akdeniz, F., ve Kaçıranlar, S., 1995. On the Almost Unbiased Generalized Liu Estimator and Unbiased Estimation of the Bias and MSE. Comm. Statist. Theory Methods, 24, 1789-1797.

Allen, D. M., 1971. Mean Square Error of Prediction as a Criterion for Selection of Variables. Technometrics, 13, 469-475.

Andrews, D.F., 1974. A Robust Method for Multiple Linear Regression. Technometrics, 16, 4, 523–531.

Arslan, O., ve Billor N., 1996. Robust Ridge Regression Estimation Based on the GM-Estimators. Journal of Math, 9, 1, 1-9.

Arslan, O., ve Billor N., 2000. Robust Ridge Regression Estimation Based on the M-Estimator. Journal of Applied Statistics, 27, 1, 39-47.

Askin, G.R., ve Montgomery, D.C., 1980. Augmented Robust Estimators. Technometrics, 22, 333-341.

Çetin, M. 2000. Sağlam Regresyonda Değişken Seçim Ölçütleri, Hacettepe Üniversitesi Fen Bilimleri Enstitüsü, Doktora tezi.130pp.

Çetin, M. C., Erar, A., 2002.Variable selection with Akaike information criteria: A comparative study. Hacettepe Journal of Mathematics and Statistics, 31:89-97.

Çetin, M. ve Erar, A., 2006. A simulation study on classic and robust variable selection in linear regression, 175,1629-1643.

Çetin, M., 2009.Robust model selection criteria for robust Liu estimator. European Journal of Operational Research, 199:21-24.

Cox, D. R. ve Snell, E. J., 1974. The choice of variables in observational studies, Appl. Statist., 23, 51-59.

Daniel, Cuthbert ve Wood, Fred S., 1980. *Fitting Equations to Data: Computer Analysis of Multifactor Data*. Second Edition, with the assistance of John W. Gorman. Wiley, New York.

Delaney, N.J., Chatterjee, S., 1986. Use of the bootstrap and cross-validation in ridge regression, Journal of Business and Economic Statistics. Vol.4, No.2, 255-262.

Dorugade, A.V., Kashid, D.N. 2010.Variable selection in linear regression based on ridge estimator. Journal of Statistical Computation and Simulation, 80(11):1211-1224.

Draper, N.R., Smith, H., 2003. Applied regression analysis, John Wiley and Sons, NewYork.

Dutter, R., 1977. Numerical solution of robust regression problems: Computational aspects, a comparison. Journal of Statistical Computation and Simulation, 5, 207–238.

Efroymson, M. A., 1960. Multiple regression analysis in A. Ralston and H. S Wilf (Eds.), Mathematical Methods for Digital Computers, Wiley, New York.

Ertaş, H., Kaçıranlar, S., Güler, H. 2017. Robust Liu-type estimator for regression based on M-estimator. Communications in Statistics - Simulation and Computation, 46(5):3907-3932.

Ertaş, H., 2015. DAYANIKLI YANLI TAHMİN EDİCİLER . Çukurova Üniversitesi, Fen Bilimleri Enstitüsü - Adana.

Ertaş, H., Toker, S., Kaçıranlar, S. 2015. Robust two parameter ridge M-estimator for linear regression. Journal Of Applied Statistics, 42(7):1490-1502.

Golub, G.H., Heath, M. ve Wahba, G., 1979. Generalized Cross-Validation as a Method for Choosing a Good Ridge Parameter. Technometrics, Vol.21, No.2, 215-223.

Gruber, M. H. J., 1998. Improving Efficiency by Shrinkage: The James-Stein and Ridge Regression Estimators. Marcell Dekker, Inc. New York.

Hald, A., 1952. Statistical Theory with Engineering Applications. New York: Wiley.

Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J., Stahel, W.A., 1986. Robust Statistics: The Approach Based on Influence Functions. John Wiley, New York.

Hocking, R. R. ve 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., 1972. Criteria for selection of a subset regression: Which one should be used, Technometrics, 14, 967-970.

Hocking, R. R., 1976. The analysis and selection of variables in linear regression, Biometrics, 32, 1-49, 1044.

Hoerl, A.E. Kennard, R.W., 1970. Ridge regression: applications to nonorthogonal problems, Technometrics 12, 69–82.

Hoerl, A.E., Kennard, R.W., 1970. Ridge regression: biased estimation for nonorthogonal problems, Technometrics 12, 55–67.

Hoerl, A.E., Kennard, R.W., Baldwin, K.F., 1975. Ridge regression: some simulations, Commun. Stat. 4, 105–123.

Holland, W. P., 1973. Weighted Ridge and Robust Regression Methods. NBERWorking Paper, Cambridge, MA.

Holland, W. P., ve Welsch, E.R., 1977. Robust Regression Using Iteratively Reweighted Least-Squares. Commun. Statist.-Theor. Meth, A(6), 9, 813–827.

Huber, P.J., 1964. Robust Estimation of a Location Parameter. The Annals of Mathematical Statistics, 35, 73-101.

Huber, P.J., 1981. Robust Statistics, John Wiley and Sons, New York.

Hurvich, C. M., ve Tsai , C.H., 1989. Regression and time series model selection in small samples. Biometrika 76: 297–307.

İyi, P., 2006. Genetik Algoritma uygulanarak ve Bilgi kriterleri kullanılarak çoklu regresyonda medel seçimi. Ç.Ü. Yüksek Lisans Tezi, Adana, 125s.

Jadhav, N. H., Kashid, D. N., Kulkarni, S. R., 2014.Subset selection in multiple linear regression in the presence of outliner and multicollinearity. Statistical Methodology, 19:44-59.

Jadhav, N.H., Kashid, D.N., 2011. A jackknifed ridge M-estimator for regression model with multicollinearity and outliers, J. Stat. Theory Pract. 5 (4), 659–673.

Kashid, D. N., Kulkarni, S. R., 2002.A more general criterion for subset selectionin multiple linear regression. COMMUN. STATIST.-THEORY METH., 31(5):795-811.

Lawless, J. F. ve Wang, P., 1976. A Simulation Study of Ridge and Other Regression Estimators. Communication in Statistics, 7, 139-164.

Lindsey, C., Sheather, S., 2010.Variable selection in linear regression. The Stata Journal, 10(4):650-669.

Linhart, H., Zucchini, W., 1986.Model selection, New York.

Liu, K., 1993. A New Class of Biased Estimate in Linear Regression. Comm. Statist. Theory Methods, 22, 2, 393-402.

Mallows, C. L., 1973. Some comments on Cp. Technometrics 15: 661–675.

Mallows, C.L., 1975. On some topics in Robustness. Unpublished Memorandum, Bell Telephone Laboratories, Murray Hill, NJ.

McDonald, G.C., Galarneau, D.I., 1975. A Monte Carlo evaluation of some ridge-type estimators, J. Amer. Statist. Assoc. 70 (350) 407–416.

Miller, A. J., 1990. Subset selection in regression, London: Chapman and Hall.

Montgomery, D. C., Friedman, D. J., 1993. Prediction using regression models with multicollinear predictor variables. *IIE Trans.* 25:73–85.

Montgomery, D.C., Peck, E.A. ve Vining, G.G., 2001. Introduction to Linear Regression Analysis, 3rd Edition, John Wiley & Sons, New York.

Myers, R. H., 1990. Classical and Modern Regression with Applications, 2nd ed., PWS-Kent Publishers, Boston.

Özkale M. R., Kaçıranlar, S., 2007. The Restricted And Unrestricted Two Parameter Estimators. Communications In Statistics-Theory And Methods, 36(15):2707-2725.

Pfaffenberger, R.C., ve Dielman, T.E., 1985. A Comparison of Robust Ridge Regression. Proceeding of the American Statistical Association Business and Economic Statistics Section, Las Vegas, Nev., 631-635.

Pfaffenberger, R.C., ve Dielman, T.E., 1990. A Comparison of Regression Estimators When Both Multicollinearity and Outliers are Present. In Robust Regression, 243-270

Ronchetti, E., Staudte, R.G.A., 1994. Robust Version of Mallows' Cp. Journal of the American Statistical Association, 89(426), 550–559.

Schwarz, G., 1978. Estimating the dimension of a model. Annals of Statistics. 6, 461-464.

Silvapulle, M. J., 1991. Robust ridge regression based on an M-estimator. Australian Journal of Statistics, 33, 319-333.

Silvey, S. D., 1969. Multicollinearity and Imprecise Estimation. Journal of the Royal Statistical Society, Series B (Methodological), 31, 3, 539-552.

Stein, C., 1956. Inadmissibility of Usual Estimator for the Mean of a Multivariate Normal Distribution. Proceeding of the Third Berkeley Symposium on Mathematical Statistics and Probability. University of California Press, Berkeley, 197-206.

Theobald, C. M., 1974. Generalizations of Mean Square Error Applied to Ridge Regresion. Journal of the Royal Statistical Society. Series B (Methodological), 36, 1, 103-106.

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.

Tukey, J. W., 1970. Exploratory Data Analysis. MA, Addison-Wesley.

ÖZGEÇMİŞ

Azer ÖNCÜ. Mithatpaşa ilköğretim okuluna 2001 yılında başladım. İlkokul ve ortaokul eğitimimi burada 2009 yılında tamamladı. Lise öğrenimimi İskenderun Anadolu Lisesi'nde 2013 yılında tamamladı. 2013 yılında Çukurova Üniversitesi Fen-Edebiyat Fakültesi İstatistik bölümüne girdi. 2017 yılında buradan üçüncülük ile mezun oldu. 2019 yılında Çukurova Üniversitesi Fen Bilimleri Enstitüsü İstatistik Anabilim Dalı'nda yüksek lisans öğrenimime başladı. Halen eğitimime Çukurova Üniversitesi Fen Bilimleri Enstitüsü İstatistik Anabilim Dalı'nda yüksek lisans öğrencisi olarak devam etmektedir.