BÖLÜM 7

7. MATEMATİKSEL MODELİN OLUŞTURULMASI ve EN İYİ REGRESYON DENKLEMİNİN SEÇİLMESİ

Regresyon teknikleri faydalı teknikler olmasına karşın yanlış kullanıldıkları ve yorumlandıklarında oldukça tehlikeli sonuçlar oluşabilir. Bu nedenle regresyon teknikleri bir probleme uygulanmadan önce proje için çalışma amaçlarının ve iş akışındaki kontrol noktalarının belirlenmesini içeren bir ön planın hazırlanması uygun olacaktır. Planlama işlemi açıklanmadan önce bilim adamları tarafından kullanılan üç matematiksel model tipi ele alınacaktır:

a) *Fonksiyonel model*: Açıklayıcı değişkenler ve yanıt arasındaki gerçek fonksiyonel ilişki biliniyor ise, araştırmacı yanıtın davranışlarını anlayabilecek, kontrol edebilecek ve kestirimleyebilecek bir avantaja sahiptir. Bununla birlikte karşılaşabilecek durumların pek azı bu avantaja sahiptir. Fonksiyonel denklemler genellikle karmaşık bir yapıya sahip olup kullanılmaları ve yorumlanmaları oldukça zordur ve doğrusal olmayan bir yapıya sahiptirler. Örneğin pek çok fiziksel ve kimyasal süreç, doğrusal olamayan modelleri ifade eden fark denklemleri sistemi ile belirtilirler. Bu gibi durumlarda doğrusal regresyon teknikleri uygulanmaz, sadece iteratif tahminleme prosedürlerinde gerçek modele yaklaşımda bulunabilmek amacıyla kullanılabilirler.

b) *Kontrol modeli*: Fonksiyonel model tamamen bilinse bile bazı durumlarda yanıtı kontrol etmek için kullanılmaya uygun değildir. Çünkü yanıtı etkileyen değişkenlerden bazıları araştırmacı tarafından süreç içinde kontrol altında tutulamamaktadır. Örneğin polyester levha üreten bir süreçte, polyester miktarı, katkı maddelerinin miktarı süreç hızı kontrol edilebilmesine rağmen çıktıyı etkileyen fırın sıcaklığı kontrol edilememektedir. Bu nedenle araştırmacının kontrolü altındaki değişkenleri içeren bir model kullanılarak yanıtın kontrol edilmesi daha uygun bir yaklaşımdır.

Regresyon tekniklerinin uygun bir şekilde kullanılmasıyla bazı durumlarda faydalı bir kontrol modeli oluşturulabilir. Kontrol edilebilir değişkenleri kullanan tasarlanmış bir deneyin mümkün olması durumunda bu değişkenlerin yanıt üzerindeki etkisi regresyon teknikleri ile elde edilebilir. Bununla birlikte tasarlanmış deneylerin kullanılmayacağı pek çok durum mevcuttur. Bu gibi durumlarda kontrol edilemeyen faktörler, oluşturulan modeldeki kontrol edilebilen faktörlerin etkilerinin karışmasına ve etkin bir şekilde hesaplanmasına engel olabilirler. Araştırmacıların bu durumda kestirimsel modeli kullanması uygun bir yaklaşım olacaktır.

c) *Kestirim modeli*: Fonksiyonel model karmaşık olduğunda ve kontrol değişkenlerinin etkilerin birbirinden bağımsız tahminleme yeteneğinin sınırlı olduğu durumlarda araştırmacı doğrusal kestirimsel model oluşturabilecektir. Bu yaklaşım bazı durumlarda gerçekçi olmasa da en azından ilgilenilen yanıtın davranışlarının ana özelliklerini ortaya çıkarabilir. Kestirimsel modeller belirli şartlar altında süreç ya da problemin anlaşılmasına yardımcı olabilir. Bu problemler genelikle dağınık verili problemler olarak adlandırılmaktadır. Dağınık veriler yüksek derecede içsel ilişkiye (intercorelation) sahiptir. Kestirimsel modelin fonksiyonel olması ve kontrol amacıyla kullanılması

gerekli değildir. Şüphesiz bu durum kestirimsel modelin yararsız olduğu anlamına gelmemektedir. En azından daha ileri seviyedeki araştırmacılar için nelere dikkat edilmesi gerektiğini, önemli değişkenlerinin hangileri olduğunu ortaya çıkararak, bazı değişkenlerin elenmesini sağlar.

Aşağıda tanıtılan plan kestirim amacıyla oluşturulan matematiksel modelin geliştirilmesini kapsasa da, fonksiyonel ve/veya kontrol modellerinin oluşturulmasını da kapsayacak yeterliliktedir. Burada vurgulanacak olan dağınık veri tipindeki problemlerdir. Plan üç aşamalı olup bu aşamalar; planlama, geliştirme ve idame ettirmedir.

7.1 PLANLAMA VE MODEL KURMA SÜRECİ

Problemi Tanımlama, Yanıtı Seçme, Açıklayıcı değişkenleri belirleme: Herhangi bir problem çözme yönteminin en önemli aşaması problemin açık bir tanımının yapılmasıdır. Problemin tanımlanması, hem yanıtın hem de kestirici değişkenlerin açık olarak ortaya koymalıdır. Bu aşamada, yanıt ve kestirici değişkenler için araştırmacı her hangi bir kısıt getirmemelidir. Oluşturulan liste büyük olabilir. Fakat daha sonraki tartışmalarla uygun bir sayıda karar verilebilir. Değişkenlerin elenmesinde unutulmaması gereken önemli nokta her hangi bir istatistiksel yöntemin asla tek karar verici olarak kullanılmamasıdır. Sonuç olarak, belirlenmiş bir potansiyel kestirici değişkenler seti ile ilişkisi araştırılacak belirlenmiş yanıt veya yanıtları içeren bir problem tanımı oluşturulmalıdır.

Bir sonraki aşamada problemin tanımlanması sırasında oluşturulan değişkenler listesi dikkatli bir şekilde incelenmelidir. Kestirici değişkenlerin pek çoğu süreç de ölçülemediği için çalışma dışında kalacaklardır. Bu durumda ölçülemeyen değişkenin yerini tutacak ölçülebilen bir değişken kullanılacak ya da orijinal değişkenin süreç içinde ölçülmesini sağlayacak bir ekipman alınacaktır. Araştırmacı bu iki alternatiften uygun olanını maliyetleri de dikkate alarak belirlemelidir. Bu çalışma planlamanın bu aşamamasında veriler toplanmadan önce tüm değişkenler için uygulanmalıdır. Cevaplanması gereken bir diğer soruda, belirlenmiş tüm x ve y'lere ait tam bir gözlemler setinin aynı zaman diliminde elde edilip edilmeyeceğidir. Bunun mümkün olmadığı pek çok durum mevcuttur. Bu gibi durumlarda bazı konularda fedakarlık yapılması gerekebilir. Tüm değişkenlerin tam bir kontrolü yapıldıktan sonra, problemin çözümünün mümkün olup olmadığı yeniden değerlendirilmelidir.

Problemin Çözümü Potansiyel Olarak Mevcut mudur? Yukarıda açıklanan değişken eleme sürecinde dikkate alına kestirici değişkenlerin bir çoğu muhtemelen elenebilecektir. Böyle bir durum problemin çözüm şansını azaltır. Bu durumda planlama aşamasında aşağıdaki üç mümkün karardan birinin gerçekleştirilmesi için karara varılmalıdır.

a) Projeye devam etmemek,

b) Edinilen bilgiler ışığında projeyi yeniden tanımlamak,

c)Projenin çözümü mümkün görünmekte, planlama devam etmeli.

Korelasyon Matrisi ve İlk Regresyon İşlemleri: Projenin devam etmesine karar verilmesi durumunda, projenin tamamlanmasına yönelik bir zaman tablosu oluşturmak gerekir. Bu nedenle problemin analitik çözümünde karşılaşılabilecek, korelasyon matrisi ve korelasyon matrisinin tersi hesaplanmalıdır. Kestirici x değişkenlerinin korelasyon matrisinin tersinin köşegen elemanları varyans

artış faktörleri (*VIF*) olarak adlandırılır. Bu faktörlerin istenilen değerlerinin 1'den büyük fakat kesinlikle 10'dan küçük olması gerektiğini Marguardt (1970) 'de belirtilmiştir. Eğer herhangi bir *VIF* değeri ondan büyük ise ait olduğu en küçük kareler katsayısı kötü tahminlenmiştir ve modelin yeniden düzenlenmesi gerekebilir. Bu gibi durumlarda orijinal *x* değişkenleri arasında yüksek korelasyon mevcuttur. Bu nedenle model oluşturma sürecinin mevcut verilerle geliştirilmesi kolay olmayabilir.

Bir sonraki adımda, sırayla her bir yanıt değişkeni için *x* değişkenleri ile arasındaki korelasyon incelenir. Her bir yanıt için bir veya iki büyük korelasyon olmalıdır. Eğer böyle bir durumla karşılaşılmaz ise aşağıdaki soruların cevapları aranmalıdır.

a) Önemli bir değişken ihmal edildi mi?

b) *x* değişkenlerinin sınırları yeterince büyük mü?

Planlamanın bu aşamasında, iyi bir kestirim modeli oluşturma şansın ne olduğunun değerlendirilmesi, gerekli olan zaman ve bütçenin tahminlenmesi gerçekleştirilmelidir.

Amaçların Oluşturulması ve Bütçenin Hazırlanması: Araştırmanın bu aşamasında, analizci proje amaçlarını oluşturmalı, problem çözümünün zaman tablosunu tahminlemeli ve bilgisayar iş gücü gereksinimlerini belirlemelidir. Zaman, iş gücü, kontrol noktaları tasarlanan zaman tablosunda ilgililerle mutabık kalınarak tanımlanmalıdır. Bu çalışma için bir proje formunun oluşturulması ve bu formda sorumluların proje amacının, analiz tipinin, zaman tablosunun ve bütçenin içerilmesi uygun olacaktır. Tüm ilgilileri tarafından onaylanan proje teklifi geliştirme adımına geçer. Bu proje gerçekleştirilmez bir yapıda ise gözden geçirilmeli ve iyileştirilmelidir. Aksi durumda proje durdurulmalıdır.

7.2 MATEMATİKSEL MODELİN GELİŞTİRİLMESİ

Proje artık kestirim modelinin geliştirilmesi için hazır durumdadır. Regresyon modelinin geliştirilmesinde kullanılan teknikler özellikle Bölüm 6 ve 11 de tanıtılmıştır. Bu kısımda prosesdeki kontrol noktaları için izlenebilecek yöntemin ana hatları tanıtılacaktır. Burada sadece yöntem tanıtılacaktır. Uygulanabilecek pek çok değişik durum olduğu unutulmamalıdır.

Veri Toplama, Verilerin Kontrolü, Grafik ve Modelin Geliştirilmesi: Veri toplama sürecinde, planlama adımında tanımlanan kısıtların karşılandığından emin olunacak şekilde dikkatli olunmalıdır. Sayılar mümkün olduğu ölçüde kontrol edilmelidir. Dağınık verilerin kalitesi genellikle oldukça düşüktür. Kötü bir veri seti ile model oluşturmaya çalışılmamalıdır. İnsanların gerçekleştirdiği ölçümlerde insan hatasının olduğu unutulmamalıdır.

Veriler toplanıp kontrol edildikten sonra modelleme süreci başlar. Verilerin grafiğinin çizilmesi, model uyumu, artıkların incelenmesi ve daha önce belirtilen bölümlerdeki tüm analatik prosüdürler bu amaçla uygulanır. Potansiyel denklemler elde edilerek uzmanların yorum ve değerlendirmelerine sunulur. Uzmanların yeni bir değişkenin kullanılmasını önermeleri durumunda bu değişkene ait veriler elde edilerek ilgili yanıt değerleri ile birlikte, uyumu yapılan denklemden elde edilen artıklar incelenmelidir.

Uyumu Yapılan Bazı Regresyon Denklemlerinin İncelenmesi: Regresyon denklemine seçilen kestirici değişkenler incelenmelidir. Bu amaçla kestirici değişkenlerin transformasyonları da gündeme gelebilir. Örneğin artık plotları bir kestirici değişkenin kara kökünün, değişkenin kendisinden daha kullanışlı olduğunu belirtebilir. Bununla birlikte tüm değişkenlere uygun transformasyon bulunması için uğraşmak nadiren olumlu sonuç vermektedir. Bu nedenle kullanılan teknikler her hangi bir değişkene transformasyon kullanılmasının faydalı olabileceğini belirtmedikçe bu yaklaşımın kullanılması zaman kaybına yol açacaktır.

Eldeki potansiyel denklemler üzerinde yapılan çalışmalar sonucunda araştırmacı en iyi olduğunu belirlediği denklemi planlama aşamasında oluşturulan amaçlar doğrultusunda incelenmelidir. Elde edilen denklemin bu standartları karşılamaması durumunda, projenin durdurulacak mı? yoksa geliştirme döngüsünün tekrarlanacak mı? Sorusuna cevap verilmelidir. Bu süreçteki bir diğer kontrol noktası da bu sorgulamanın gerçekleştirilmesidir.

7.3 MATEMATİKSEL MODELİN DOĞRULANMASI VE İYİLEŞTİRİLMESİ

Proje amaçlarını karşılayan bir denklem kestirici model olarak kabul edildikten sonra onun doğrulanması ve iyileştirilmesi için çalışmalara başlanmalıdır.

Örnek Uzayı İçin Parametreler Durağan mıdır? Parametre durağanlı incelendiğinde, veri setleri iki gruba ayrılabilir: Zaman verileri ve kesit verileri.

Zaman Verileri: Eğer uzun zaman dönemine ait gözlemler kullanılarak model oluşturulmuş ise, araştırmacı bu uzun dönemi daha kısa dönemlere ayırarak her biri için model uyumunu gerçekleştirip tahminlenmiş regresyon katsayılarının durağanlığını test edebilir. Örneğin model son beş yılın aylık verilerine dayanarak oluşturulmuş ise, her yılın verileri ayrı ayrı kullanılarak regresyon katsayı tahminleri durağan değilse, tüm verileri kullanarak elde edilen denklemin kestirim amacıyla kullanılması akıllı davranış olacaktır.

Kesit Verileri: Eğer veriler ayı zaman diliminde ya da ayni vardiyada veya aynı hammadde partisinden elde edilmiş ise kesit verisi olarak adlandırılır ve durağanlığı incelemenin birkaç yöntemi mevcuttur. Bu yöntemlerdeki temel fikir veri setini bazı kriterlere göre alt setlere bölmektir. Daha sonra verilerin bir bölümü kullanılarak kestirim modeli elde edilir ve verilerin diğer bölümü ile bu model doğrulanır. Amaç kestirimlerin ne derece iyi olduğu görebilmektir. Değişken seçiminde olduğu gibi bu problem için de tek bir çözüm ya da en iyi çözümü veren bir yöntem yoktur. Bu amaçla kullanılan yöntemlerden bazıları aşağıda tanıtılmıştır.

Bir gözlemin ayrılması yaklaşımı; PRESS Prosedürü Allen (1971) olarak adlandırılan bu yöntem bir doğrulama yöntemidir. Bu yöntemde gözlemlerden bir tanesi veri setinden ayrılarak kalan gözlemleri ile modelin uyumu yapılır ve gözlem değeri ile tahminlenen değer arasındaki farkın karesi alınır. Daha sonra bu işlem teker teker tüm gözlemler için tekrarlanarak, farkların kareler toplamı elde edilir. Değişik modeller için bu yöntem uygulanarak veri seti için doğrulanmış en iyi model elde edilir. Aynı zamanda veri tutarsızlıkları bireysel farklılıklar yardımı ile incelenebilir. Ayrıca $\boldsymbol{\beta}$ tahminlerinin özelliklerinin incelenmesi gerekebilir.

Birden fazla gözlemin ayrılması yaklaşımı: Allen 'in yaklaşımına benzer bir yöntem Geisser (1975) tarafından önerilmiştir. Aradaki fark tek gözlem yerine *m* adet gözlemin bırakılarak kalan *n-m* adet gözlemi ile modelin uyumu yapılmakta ve doğrulama işlemi kalan *m* adet gözlem kullanılarak gerçekleştirilmektedir.

Yarı yarıya yaklaşımı: Modeli oluşturmak için verilerin yarısı kullanılır ve diğer yarısı doğrulama amacıyla kullanılır. Snee (1977) modeli oluşturmak için kullanılacak alt setin seçilmesi problemi ile ilgilenmiştir. Önermiş olduğu Duplex algoritması, hem modeli oluşturan hem de doğrulama için kullanılan verilerin $\mathbf{X}^T\mathbf{X}$ matrisinin determinant özelliklerinin benzer olmasını sağlamaya yöneliktir. Snee nin kullandığı kural, iki determinantın oranının *k*-ıncı kökünün hesaplanmasına dayanmaktadır:

$$\left\{\frac{\left|\mathbf{X}^T\mathbf{X}\right|_{tah\,\min}}{\left|\mathbf{X}^T\mathbf{X}\right|_{kestirim}}\right\}^{1/k}$$

Burada tüm orijinal *x* değişkenleri standadize ve ortonormalize edilmiş olduğundan, bkz. Bölüm 8, $\left|\mathbf{X}^T\mathbf{X}\right|$ matrisi korelasyon formundadır, *k* ise **X** matrisindeki değişken sayısıdır Eğer veri seti uygun bir şekilde bölünmüş ise bu istatistik yaklaşık olarak bire eşit olmalıdır. Bununla birlikte $n>2p+25$ olmadıkça, *n*=toplam örnek hacmi, *p*=modeldeki parametre sayısı, veri setinin ayrıştırılmasının sakıncalı olduğu Snee tarafından belirtilmiştir.

*Sistematik Uyum Yetersizliği Var mıdır***?** Durağan parametreleri olan uyumu yapılmış bir model kabul edildiğinde bile değişkenlerin modelden çıkarılması söz konusu olabilecektir. Böyle bir duruma gerek olup olmadığını ortaya koymak amacıyla artıklar bilinen tüm yöntemler (bkz. Bölüm 6 ve 11) kullanılarak incelenmelidir.

Parametre tahminleri uygun mudur? Bu soru özellikle model uzman olmayan kişiler tarafından kullanıldığı durumlarda önem kazanmaktadır. En küçük kareler katsayıları regresyondaki diğer değişkenleri de etkilemektedirler. Modeli kullananlar sadece bir değişkenin katsayısında değişiklik yaparak yanıtın kestirimlenmesini gerçekleştirmek isteyebilir Bu durunda eğer tüm tahminlenmiş katsayılar birbirinden bağımsız ise oluşan zarar muhtemelen küçük olacaktır. Bununla birlikte kestirici değişkenler arasındaki yüksek korelasyon olması durumunda, tahminlenmiş katsayılarda korelasyonlu olacaktır. Bu gibi durumlarda bireysel katsayılara güvenmek tehlikeli sonuçlar doğurabilir. Yapılacak en güvenilir hareket, kestirimleme işleminin, orijinal verilerin elde edildiği **X**-uzayı içinde gerçekleştirilmesi bu uzayın dışında modelin kestirimleme amacıyla kullanılmamasıdır. Ayrıca bireysel katsayıların yönlerinin (işaretlerinin) doğru olup olamadığı da kontrol edilebilir. Örneğin x_1 gübre miktarı *y* alınan ürün ise b_1 katsayısının işareti pozitif olmalıdır.

Denklem güven veriyor mu? Denklem uzmanlarının incelemesinden geçmiş midir? Denklemdeki değişkenler uygun mudur? ve denklemde bulunması gereken her hangi bir değişken unutulmuş mudur? sorularının cevabı evet ise denkleme güvenilecektir.

Denklem kullanılabilir mi? Final modeli, muhtemelen kontrol amacıyla değil kestirim amacıyla kullanılabilecek değişkenleri içermektedir. Bu durum bazı problemler oluşturabilir. Örneğin; bir sürecin standart çalışma şartlarının tanımlanabilmesi için *r* adet değişkenin bu süreç üzerindeki etkisinin belirlenmesi gereklidir. Buna karşın süreç çıktısının kestirimlenebilmesi için sadece *p* adet ($p<r$) değişkeni içeren bir kestirim modeli oluşturulmuştur. Eğer araştırmacı süreç çıktısını bu modeli kullanarak kestirimlemek ister ise denklemde bulunmayan *r-p* adet değişkeni ihmal etmiş olacaktır. Standart çalışma koşulları altında böyle bir yaklaşım sorunlar çıkaracaktır. Çünkü kestirim için gerekli olmasa da *r-p* adet değişkenin süreç çalışma değerelerinin belirlenmesi *p* adet değişken de olduğu kadar önemlidir.

Eğer daha önce belirtilen tüm kriterler ve kontrol noktaları sağlanmış ise modelin iyileştirilmesi için bir prosedür oluşturulmalıdır. Eğer analizci kestirilmiş değerler ile gerçek gözlemler arasındaki sapmayı izlemek amacıyla standart bir kalite kontrol kartı prosedürü oluşturmuş ise modelin kontrolü açısından böyle bir yaklaşım yeterli olacaktır. Bununla birlikte modelin iyileştirilmesi için periyodik olarak kontrol edilerek yeniden düzenlenmesi ihmal edilmemelidir. Çünkü çalışmanın başında şüphelenilen ya da daha sonra ortaya çıkan bazı karmaşık değişken etkileri ancak bu tipteki kontroller ile ortaya çıkarılıp model yeniden düzenlenerek etkinliği artırılabilir.

7.4 PLANLANMAMIŞ VERİLERİN KULLANIMINDA DİKKAT EDİLECEK DURUMLAR

Regresyon çalışmaları planlanmamış verilere dayanılarak gerçekleştirilmek zorunda olduğunda bazı potansiyel tehlikeler mevcuttur. Planlanmamış veriler, bir deney tasarımından elde edilmemiş ya da kesintisiz operasyonlardan elde edilmemiş verilerdir. Dikkat edilmesi gereken durumlar aşağıda tanıtılmıştır.

Modeldeki hata teriminin tam olarak rassal davranış göstermemesi; böyle bir durum regresyon denkleminde birlikte oluşturdukları etki dikkate alınmayan ya da hiç denkleme dahil edilmeyen birkaç değişkenin ortak etkisinin sonucu olarak ortaya çıkabilir. Bu tip değişkenler görünmeyen (latent) değişken olarak adlandırılır. Denklemdeki bir değişkenin tahmininde gözlenen hatalı bir etki gerçekte ölçümlenmemiş görünmeyen bir değişkenin neden olduğu sapmaya bağlı olabilir. Sürekli sistemlerde, verilerin buna paralel olarak sürekli bir kayıt sistemine sahip olması iyi bir yaklaşımdır. Fakat görünmeyen değişkenler ölçümlenmediği için ondaki değişiklikler kayıt edilmeyecektir. Bu değişkenler ölçümlenmediği için ondaki değişiklikler de kayıt edilmeyecektir Bu değişimler kestirim denklemine de yansıyacağı için kullanılan denklemin güvenirliği azalacaktır. Bu problem kestirim modeli oluşturulurken değişken seçiminin ne kadar önemli olduğunu ortaya koymaktadır.

Kestirim değişkeninin sınırları; yanıt değişkenini spesifikasyon limitleri içinde tutmak amacıyla pek çok etkili kestirim değişkeni oldukça küçük sınırlar içinde tutulurlar; sınırların küçük tutulması sık sık regresyon katsayısının anlamsız olarak yorumlanmasına neden olmaktadır. Eğer araştırmacı tecrübeli ise problemin kaynağının değişkenin sınırları olduğunu fark edebilecektir. Gerçekte etkin bir kestirici değişkenin sınırları yeterli ölçüde değiştirilmez ise etkisiz ya da az etkili olarak değerlendirilebilir.

Sürecin çalışma şartları; bazı değişkenlerin süreçteki çalışma koşulları birlikte hareket etmelerini, örneğin x_1 artarken x_2 azalır gibi, zorunlu hale getirmesi bu kestirici değişkenler ararsında yüksek korelasyon oluşmasına neden olabilmektedir. Bu gibi durumlarda x_1 ve x_2 'nin bireysel ya da birlikte y üzerindeki etkilerini işlemek imkansız hale gelebilecektir.

Yukarıda açıklanan tüm problemler dikkatli bir deney tasarımı ile giderilebilir. Bununla birlikte tasarlanmış deneylerin mümkün olmadığı durumlarda da regresyon analizi hala geçerli bir tekniktir. Çünkü veri analizi sırasında yukarıdaki problemleri ortadan kaldıracak yöntemler geliştirilmiştir.

7.5 REGRESYON DENKLEMİNİN KULLANIM AMAÇLARI VE DEĞİŞKEN SEÇİMİNİN ÖNEMİ

Regresyon denkleminin hangi amaçla kullanılacağı kurulacak model üzerinde uygulanacak olan EKK analizleri açısından oldukça önemlidir. Regresyon denklemlerinin temel kullanım amaçları Hocking (1976) tarafından listelenmiştir:

1) Yanıt değişkeninin davranışının iyi bir tanımını oluşturmak.
2) Yanıtın gelecek değerlerinin kestirilmesi ve ortalama yanıtın tahminlenmesi.
3) Veri setinin sınırları dışındaki yanıtın kestirilmesi, (ekstrapolasyon).
4) Parametrelerin tahminlenmesi.
5) Bağımsız değişken seviyelerini değiştirerek süreci kontrol etmek.
6) Sürecin gerçek modeline yaklaşmak.

Yukarıda tanımlanan amaçlar ile bu bölümün girişinde açıklanan model tipleri arasında bir ilişki olduğu açıktır.

Modelden değişkenlerin elenmesi, her bir amaç için farklı öneme ve anlama sahiptir. Burada anlam ve önem olarak kastedilen, modeldeki değişkenlerin yanıt değişkeni ile olan nedensellik ilişkisi ve modelin gerçekçi olması için ortaya konan çabadır. Nedensellik ve gerçekçilik üzerine verilen kararlar belirli bir veri setinin dışındaki, örneğin sistemin nasıl çalıştığına dair temel bilgilere dayanmalıdır.

Amaç, belirli bir veri setindeki yanıt değişkeninin davranışının basit bir tanıtımının yapılması ise; modelin gerçekliği ile ilgili olarak ya da nedensellik ilişkisi açısından modeldeki değişkenlerin elenmesi fazla önem taşımamaktadır. Çünkü minimum hata kareler toplamını dikkate alan en küçük kareler yöntemi ile yanıt değişkeninin en iyi tanımı tüm değişkenleri içeren (tam) model ile elde edilecektir. Bu açıdan bakıldığında değişkenlerin nedensellik ilişkisinin ya da modelin gerçekçi olup olmadığının bir önemi yoktur.

EKK regresyonun diğer amaçları açısından bakıldığında değişken eleme konusu daha fazla önem kazanmaktadır. Regresyon denklemindeki değişken sayısı azaldıkça denklem daha basit bir yapı kazanacak ve bu denklemin oluşturulması ve kullanılması için gerekli olan bilginin elde edilmesi daha ekonomik alacaktır. Bununla birlikte uygun değişkenlerin elenmesiyle kestirim yeteneği kaybı ve sapmanın oluştuğu da unutulmamalıdır.

Regresyonun kullanım amaçlarından olan ortalama yanıtın kestirim ve tahminlenmesi, değişken elenmesi konusunda oldukça toleranslıdır. Aynı zamanda değişkenlerin nedensel ilişkileri veya modelin

gerçekliği diğer kullanım amaçlarına göre daha az önem taşır. Buradaki gizli varsayım, verilerin elde edildiği andaki sistem çalışma şartlarının süreklilik gösterdiği ve kestirimleme ve tahminlemenin de bu **X**–uzayı içinde gerçekleştirildiğidir. Sonuç olarak bağımlı değişken için kestirimsel bilgiyi içeren değişkenler faydalı değişken olarak nitelenirler. Önemli olan bilginin uygun maliyetlerle elde edilmesidir.

Extrapolasyon, X-veri uzayının dışında gerçekleştirildiği için değişken seçimi daha önemli hale gelmektedir. Diğer bir deyişle sistemin davranışı olabildiğince tam olarak tanımlamaya yarayacak değişkenlerin modelde de kalabilmesi için oldukça dikkatli olmak gerekecektir. Extrapolasyon tehlikeli bir işlem olup, eğer elde edilen denklem gerçek modeli yeterli bir şekilde temsil edemiyor ise oldukça kötü sonuçlara neden olabilir. Extraplasyonda, örnekte gözlenen ilişkisel yapının örnek uzayının dışında da devam ettiği şeklinde çok güçlü bir varsayım yapılmaktadır. Extrapolasyon ya da öngörümleme amacıyla kurulan denklemlerin doğrulanması ve sürekli olarak güncelleştirilmesi gereklidir.

Parametre tahminleme amacıyla yapılan çalışmalarda araştırmacı değişken eleme konusunda tutucu davranabilecektir. Böyle bir yaklaşım uygun bir değişkenin elenmesi sonucu oluşacak sapmayı önleyecektir. Bununla birlikte gerçekten bağımlı değişkenle ilişkisi olmayan değişkenin denklemden çıkarılması, tahminlerin varyansını indirgemek gibi bir avantaja sahiptir.

Bir kontrol sistemi için parametre tahminlerinin oldukça iyi olması gereklidir. Fakat bundan daha da önemli olan bağımsız değişken ile yanıt değişkeni arasında nedensellik ilişkisinin olmasıdır. Aksi takdirde araştırmacının düzeltme amacıyla sisteme müdahalesi imkansız hale gelebilecek ve bağımsız değişkenlerin değerlerinin değiştirilmesi sonucu oluşacak değişikliklerin etkisi izlenemeyecektir.

Gerçek modelin oluşturulması için yapılan çalışmalarda ana amaç sürecin anlaşılmasıdır. Bağımlı değişkenin ifade edilmesinde, her ne kadar bazı nedensel ilişkiler mevcut olsa da asıl önemli olan bu değişkenlerin yeterince açık olsun olmasın iyi bir tanımın yapılmasıdır. Bu amaçla değişken seçim prosedürleri, göreceli olarak önemsiz olan, veri setindeki gözlenmiş korelasyon yapısı üzerine kurulur. Böyle bir çalışma ise daha ileri bir nedensel ilişki çalışmasına temel oluşturacak olan değişken sınıflarının tanımlanmasında kullanılan araç olmaktan öteye geçmeyecektir. Araştırmanın amacı daha çok sürecin anlaşılması yönünde ise, fonksiyonel formu gerçekçi olarak sistemin davranışını ifade eden modelin geliştirilmesi için çalışmalar artacaktır.

Bir araştırmacının, değişkenlerin öneminin oluşturulması ya da verilerin doğası ve kaynağı arasındaki nedensellik ilişkisini hangi detayda araştırması gerektiğini yönünde bir kararı olmalıdır. En küçük kareler regresyon sonuçları sadece analiz edilen verilerin korelasyon yapısını ortaya koyar. EKK analizi nedensellik ilişkisini kurmaz. Nedensellik ilişkisi sadece kontrollü deneylerde, nedensel olduğundan şüphelenilen değişkenin değeri değiştirilip bağımlı değişkendeki etkisi ölçülerek elde edilebilir. Her hangi bir değişken seçim prosedürü sonuçlarının, modellenen sürecin belirtilen bilgileri ile önerilen modelin tutarlı olmasını garanti altına almak için oldukça dikkatli bir şekilde incelenmelidir.

Modelin geliştirilmesi amacıyla değişken seçimi konusunda üç temel problemle karşılaşılır:

1) EKK regresyon sonuçları üzerinde değişken seçiminin teorik etkileri.

2) Farklı değişken sayılarına sahip alt kümeler için en iyi alt kümenin bulunmasında kullanılan hesaplama yöntemleri.

3) Sonuç modeli için alt küme büyüklüğünün seçilmesi, (durdurma kuralı).

Bu konuların problem olarak ortaya çıkmasının temel nedeni sonuç model elde edilirken dikkate alınan ve birbiri ile ters ilişkili iki kriterdir.

a) Kestirim amacıyla kullanılacak denklemler için, güvenilir uyumu yapılmış değerleri belirlemek amacıyla mümkün olduğunca fazla kestirim değişkeninin modelede içerilmesi istenir.

b) Bilginin elde edilmesinde artan maliyetleri dikkate almak ve model yapısının karmaşıklaşmasını önlemek amacıyla olabildiğince az kestirim değişkenin modelde yer alması istenir.

Gereğinden az sayıda terimin bulunduğu modeller *eksik belirlenmiş* (underspecification), gereğinden fazla sayıda terim içeren modeller ise *aşırı belirlenmiş* (overspecification) model olarak adlandırılır.

Regresyon analizinde bu iki uç durum arasında uzlaşı sağlamak amacıyla yapılan çalışmalara *en iyi regresyon denkleminin seçimi* adı verilir. Yukarıda verilen iki kriterin özel bir durumu olarak n örnek hacmin sabit olduğu durumlar için, parametre sayısının artmasının artık serbestlik derecesini azaltmasıdır.

7.6 DEĞİŞKEN SEÇİMİNİN EN KÜÇÜK KARELER YÖNTEMİ ÜZERİNDEKİ ETKİSİ

Değişken seçiminin EKK üzerindeki etkisi seçim işleminin sadece mevcut verilerdeki bilgi üzerine oluşturulmadığı durumlarda ortaya konabilir. Fakat bu durumlarla sık karşılaşılmadığı için, değişken seçiminde teorik sonuçlar dikkate alınır.

Gerçek modelin r değişken içerdiği fakat regresyon denkleminin bu değişkenlerden p adedini içeren bir alt küme ile oluşturduğu varsayılsın. $\mathbf{X}$ ve $\boldsymbol{\beta}$ bu p değişkeni içeren $\mathbf{X}_r$ ve $\boldsymbol{\beta}_r$ 'nin alt matrisleri ve vektörleri olsun. $\hat{y}_i$, $\hat{y}_0$ ve s^2 ise sırasıyla p değişkenli modelden elde edilmiş tahminlenmiş ortalama, kestirimlenmiş değer ve artık kareler ortalamasıdır. Aşağıdaki özellikler Hocking (1976) tarafından özetlenmiştir.

1. Elenen tüm değişkenlerin gerçek regresyon katsayıları sıfır olmadıkça s^2 değeri σ^2'nin pozitif sapmalı bir tahminini verir, bkz. Kısım 4.9.1 ve 7.6.1
2. Elenmiş her bir değişken için gerçek regresyon katsayıları sıfır olmadıkça veya elenmiş değişkenlerin her biri modelde kalan p adet değişkene ortogonal olmadıkça $\mathbf{b}$ değeri $\boldsymbol{\beta}$'nın, $\hat{y}_i$ değeride $E(y_i)$ nin sapmalı tahminlerini verir, sırasıyla bkz. Kısım 4.4 ve Kısım 7.6.2
3. $\mathbf{b}$, $\hat{\mathbf{y}}_i$ ve $\hat{\mathbf{y}}_0$ değerleri r değişkenli modelden elde edilen istatistiklere göre daha az değişkendirler, bkz. Kısım 7.6.3
4. $\mathbf{b}$, $\hat{\mathbf{y}}_i$ ve $\hat{\mathbf{y}}_0$ değerlerinin ortalama karesel hatalarının (varyans artı sapma kare) r değişkenli modelden elde edilen tahminlerin varyansından daha küçük oldukları şartlarda mevcuttur, bkz. Kısım 7.6.4

$\beta_j \neq 0$ durumunda uygun değişkenlerin modelden çıkarılması (özellik 1 ve 2) sonucunda bir sapma cezası ortaya çıkar. Buna karşılık eğer modelden değişken çıkarılırsa hem tahminleme hem de kestirimlerin varyansı azalmaktadır, (özellik 3). Ayrıca, gerçek regresyon katsayıları sıfırdan farklı olduğu halde değişkenin denklemden çıkarılması durumunda hem tahminlerin hem de kestirimlerin ortalama karesel hatasının azaldığı durumlarla da karşılaşılabilir, (özellik 4).

Veri setindeki bağımsız değişkenlerin ortogonal olması durumunda (deney tasarımlarında karşılaşılabilecek bir durumdur) her bir değişkenin EKK sonucu, modelde hangi değişkenlerin olduğu önemli olmaksızın değişmeyecektir. Böyle durumlarda tek bir EKK analizinden elde edilen sonuçlar modelde kalacak değişkenleri seçmek için kullanılabilecektir. Bununla birlikte bağımsız değişkenler genellikle ortogonal değildir. Ortogonalite bozukluğu gözlemsel verilerde beklenen umulmadık aksiliklere bağlı olarak tasarlanmış deneylerde de ortaya çıkabilmektedir. Bağımsız değişkenler arasındaki ortogonalite yetersizliği her bir bağımsız değişken için elde edilen EKK sonuçların modeldeki diğer değişkenlere bağımlı olmasın neden olur.

Seçilecek modelin kullanım amacının kestirim olduğu varsayılsın. Burada kestirim, ilerideki bir $\mathbf{x}^T = \mathbf{x}_0^T$ noktasındaki $E(y_0)$ beklenen değerinin $\hat{y}_0$ ile tahminlenmesi işlemidir. Sıradan e_i artıkları genellikle regresyon modelinin kestirim yeteneğini ortaya koymazlar. Bunun nedeni EKK prosedürünün, gerçek kestirim hatalarından daha küçük artıklar üreten bir regresyon fonksiyonunu oluşturacak şekilde tasarlanmış olmasıdır. Hatırlanacağı gibi $\hat{y}_i$ değeri y_i değerinden bağımsız değildir ve artık kareler toplamını minimize etmek için regresyon doğrusunu kendine çekmektedir. Bu nedenle artıklar uyumun kalitesini ölçmekle birlikte kestirimin kalitesini değerlendirmezler.

7.6.1 Eksik Ve Aşırı Belirlenmiş Modeller İçin Artık Kareler Ortalamasının Özellileri

Eksik belirlenmiş bir model için artık kareler ortalamasının beklenen değeri Kısım 4.9.1 de eşitlik (4.55) ile,

$$E\left(s_p^2\right) = \sigma^2 + \frac{\boldsymbol{\beta}_2^T \left[\mathbf{X}_2^T\mathbf{X}_2 - \mathbf{X}_2^T\mathbf{X}_1 \left(\mathbf{X}_1^T\mathbf{X}_1\right)^{-1} \mathbf{X}_1^T\mathbf{X}_2 \right] \boldsymbol{\beta}_2}{n-p}$$

tanımlanmıştır. Bu modelin parametre sayısı p olup artık kareler toplamının serbestlik derecesi n-p dir. Eşitlik (4.55) eksik belirlenmiş bir model için s^2 değerinin sapmalı ve sapmanın pozitif olduğunu göstermektedir. Eşitliğin ikinci bileşeni sapmayı ifade etmektedir. Görüldüğü gibi sapma miktarının değeri modele alınmayan değişkenlere ait parametrelerin değerleri ile belirlenmektedir. Modelin oldulça eksik belirlendiği durumlarda s^2 de oluşan sapmanın büyük olması beklenir. Farklı bir bakış açısı ile $r>p$ olmak üzere $s_r^2 > s_p^2$ olması durumunda ise p terimli modeldeki eksik belirlemenin bir sapmaya neden olmadığı ve muhtemelen $\beta_2=0$ olduğundan p terimli modelin uygun bir model olduğu kabul edilebilir. Sonuç olarak artık kareler ortalaması değerlerinin aday modellerin kıyaslanmasında kullanılabilecek bir kriter olduğu görülmektedir. Eşitlik (4.55) in ikinci bileşenindeki karesel formun tanım matrisi $\left[\mathbf{X}_2^T\mathbf{X}_2 - \mathbf{X}_2^T\mathbf{X}_1 \left(\mathbf{X}_1^T\mathbf{X}_1\right)^{-1} \mathbf{X}_1^T\mathbf{X}_2 \right]$, β_2 vektörü için EKK tahminlerinin varyans-

kovaryans matrisinin tersini tanımlamaktadır. Model dışında kalan parametreler nedeniyle s^2 de oluşan sapma ihmal edilen parametrelerin standardize değerine eşittir.

Şimdi de aşırı belirlenmiş bir model ele alınacaktır. Eşitlik (4.55) den modelin eksik belirlendiği durumlarda modelin sapmalı olduğu ve bu sapmanın artık kareler ortalamasına da yansıdığı görülmektedir. Aşağıda ise aşırı belirlenmiş bir model

$$\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon} \tag{7.1}$$

verilmiştir. β_2=0 olması durumunda r adet parametreli model aşırı belirlenmiştir. Eğer bağımsız değişkenler matrisi,

$$\mathbf{X} = \left[\mathbf{X}_1\mathbf{X}_2\right]$$

olarak bölümlenmiş ise, aşırı belirlenmiş model için artık kareler ortalaması, eşitlik (4.46*c*) kullanılarak,

$$s_r^2 = \frac{\mathbf{y}^T\left[\mathbf{I} - \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\right]\mathbf{y}}{n-r}$$

ve beklenen değeri,

$$E\left(s_r^2\right) = \frac{1}{n-r}E\left\{\mathbf{y}^T\left[\mathbf{I} - \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\right]\mathbf{y}\right\}$$

bu ifade **y** vektörü için karesel form oluşturduğundan,

$$E\left(s_r^2\right) = \frac{1}{n-r}\left\{\sigma^2 tr\left[\mathbf{I} - \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\right] + \left[E\left(\mathbf{y}\right)\right]^T\left[\mathbf{I} - \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\right]\left[E\left(\mathbf{y}\right)\right]\right\}$$

Eşitlik (7.1) deki modelin aşırı belirlenmiş olması nedeniyle β_2=0 olup $E(\mathbf{y})=\mathbf{X}_1\boldsymbol{\beta}_1$ dır ve

$$E\left(s_r^2\right) = \frac{1}{n-r}\left\{\sigma^2\left(n-r\right) + \boldsymbol{\beta}_1^T\mathbf{X}_1^T\left[\mathbf{I} - \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\right]\mathbf{X}_1\boldsymbol{\beta}_1\right\} \tag{7.2a}$$

aşırı belirlenmiş model için artık kareler ortalamasının beklenen değeri,

$$E\left(s_r^2\right) = \sigma^2 + \frac{1}{n-r}\boldsymbol{\beta}_1^T\mathbf{X}_1^T\left[\mathbf{I} - \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\right]\mathbf{X}_1\boldsymbol{\beta}_1 \tag{7.2b}$$

ve $\mathbf{X}^T\left[\mathbf{I} - \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\right] = \mathbf{0}$ ile $\mathbf{X}_1^T\left[\mathbf{I} - \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\right] = \mathbf{0}$ olduğundan,

$$E\left(s_r^2\right) = \sigma^2 \tag{7.3}$$

elde edilir. Bu sonuçtan aşırı belirlenmiş modeller için artık kareler ortalamasının σ^2 parametresinin sapmasız bir tahmin olduğu görülmektedir. Bununla birlikte tahminleyici doğru modelin artık kareler ortalamasına göre daha az serbestlik derecesi içermektedir. Normal dağılım varsayımı ile,

$$\frac{s_r^2\left(n-r\right)}{\sigma^2} \sim \chi_{n-r}^2$$

ve

$$V\left(s_r^2\right) = \frac{\sigma^2 V\left(\chi_{n-r}^2\right)}{\left(n-r\right)^2}$$

$$= \frac{2\sigma^4}{n-r} \tag{7.4}$$

elde edilir. Sonuç olarak aşırı belirlilik durumunda serbestlik derecesinin azalması s^2 nin daha büyük bir varyansa sahip olmasına neden olmaktadır. Öyleyse aşırı belirli bir model için σ^2 nin tahmininin varyansı, doğru modelden hesaplanmış σ^2 nin tahmininin varyansından daha büyüktür. Elde edilen bu teorik sonuç s^2 nin rakip modellerin kıyaslanmasında kullanılabilecek bir kriter olduğunu göstermektedir.

7.6.2 Eksik Belirlenmiş Model ile Kestirim

Kısım 4.4 de eksik belirlenmiş modelin parametre tahminleri üzerindeki etkisi incelenmişti. Eşitlik (4.23*a*)

$$E(\mathbf{b}_1) = \boldsymbol{\beta}_1 + \mathbf{A}\boldsymbol{\beta}_2$$

olup burada $\mathbf{A} = \left(\mathbf{X}_1^T\mathbf{X}_1\right)^{-1}\mathbf{X}_1^T\mathbf{X}_2$ sapma ya da eş yapı matrisi olarak adlandırılır.

Şimdi eksik belirlenmiş bir modelini kestirilmiş yanıt $\hat{\mathbf{y}}_0$ üzerindeki etkisi araştırılacaktır. Eksik belirlenen model için kestirilmiş yanıt,

$$\hat{y}_{01} = \mathbf{x}_{01}^T\mathbf{b}_1$$

burada $\mathbf{x}_{01}^T$, veri uzayının dışında ilgilenilen bir noktadır, veri setindeki bir nokta için $\mathbf{x}_{i1}^T$ ile gösterilir. Beklenen değeri ise,

$$\hat{y}_{01} = \mathbf{x}_{01}^T\left[\boldsymbol{\beta}_1 + \mathbf{A}\boldsymbol{\beta}_2\right]$$

şeklindedir. İlgilenilen nokta tüm uygun model terimlerini dikkate alınarak $\mathbf{x}_0^T = \left[\mathbf{x}_{01}^T\mathbf{x}_{02}^T\right]$ ifade edildiğinde gerçek ortalama yanıt,

$$\hat{y}_0 = \mathbf{x}_{01}^T\boldsymbol{\beta}_1 + \mathbf{x}_{02}^T\boldsymbol{\beta}_2$$

olup $\mathbf{x}_0$ noktasındaki kestirim sapması,

$$E(\hat{y}_{01}) - E(y_0) = \left(\mathbf{x}_{01}^T\mathbf{A} - \mathbf{x}_{02}^T\right)\boldsymbol{\beta}_2 . \tag{7.5a}$$

Sapmasının karesi bir karesel form olarak,

$$\left[E(\hat{y}_{01}) - E(y_0)\right]^2 = \left[Sapma(\hat{y}_{01})\right]^2 = \boldsymbol{\beta}_2\left(\mathbf{x}_{01}^T\mathbf{A} - \mathbf{x}_{02}^T\right)^T\left(\mathbf{x}_{01}^T\mathbf{A} - \mathbf{x}_{02}^T\right)\boldsymbol{\beta}_2 \tag{7.5b}$$

eşitlik (4.23*a*) ve (7.5) kullanılarak eksik belirlenmiş bir modelin s^2 değerindeki sapma, ilgili veri noktalarındaki kestirim değerinde oluşan karesel sapmaların toplamına göre,

$$E\left(s_p^2\right) = \sigma^2 + \frac{1}{n-p}\sum_{i=1}^{n}\left[Sapma\ \hat{y}_i\right]^2 \tag{7.6}$$

yazılabilir. Burada s_p^2 değeri p adet terim içeren modelin, ortalama karesel hatasını belirtir.

Eşitlikten görüldüğü gibi eksik belirlenmiş bir modelin, kestirilmiş değerler, parametre tahminleri ve hata varyansı tahmini üzerindeki etkisi sabit bir sapma miktarı şeklinde ortaya çıkmaktadır. Eşitlik (7.6) bir model seçim kriteri geliştirmede kullanılabilecek önemli bir yapıdır. Bu konu Kısım 7.7 de

ele alınacaktır. Temel olarak, eksik belirleme ya da diğer bir ifade ile önemli değişkenlerin modele alınmaması durumunda önemli değişkenlerin şans değişkeni üzerindeki etkisi artık kareler toplamına ekleneceği için artık kareler toplamı büyüyecektir. Bu artış ilgilenilen veri noktasındaki kestirimde ortaya çıkan sapma şeklinde görülebilir.

7.6.3 Aşırı Belirlenmiş Model ile Kestirim

Eksik belirlenmiş model için önemli regresyon sonuçlarının sapmalı olabileceği ve model basitleştikçe artık varyansının artma eğiliminde olduğu bir önceki kısımda açıklandı. Aşırı belirlenmiş model durumu ile ilgili önemli varyans ifadeleri aşağıda özetlenmiştir:

1. $x_1,\ldots,x_k$ değişkenlerini içeren model için EKK parametre tahminleri b_0, $b_1,\ldots,b_k$ olsun. Modele x_{k+1} değişkeninin eklenmesi ile elde edilen modelin parametre tahminleri $b_0^*, b_1^*, \ldots, b_k^*, b_{k+1}^*$ olsun. Parametre tahminlerinin varyansları için,

$$V\left(b_i^*\right) \geq V\left(b_i\right) \qquad i=0,1,\ldots,k \tag{7.7}$$

2. Yukarıda tanımlanan iki regresyon modeli için kestirim değerleri, $\hat{y}_1 = \sum_{i=0}^{k} b_i x_i$ ve $\hat{y}_2 = \sum_{i=0}^{k+1} b_i^* x_i$ olup, $\mathbf{x}_0^T = \begin{bmatrix} 1 & x_{01} & x_{02} & \cdots & x_{0k} & x_{0k+1} \end{bmatrix}$ noktasındaki kestirim varyansı,

$$V\left(\hat{y}_{01}\right) \leq V\left(\hat{y}_{02}\right) \tag{7.8}$$

eşitsizliğini sağlar.

Yukarıda tanımlanan iki özelliğin sonuçları eksik ve aşırı belirlenmiş modeller arasında nasıl denge sağlanacağı konusunda önemli ipuçları vermektedir. Modelin çok basit olması durumunda (eksik belirlenmiş) parametre tahminleri ve kestirim değerleri sapmalı olacaktır. Buna karşın karmaşık yapıdaki (aşırı belirlenmiş) modellerde ise parametre tahminleri ve kestirilmiş değerlerin varyansı büyüyecektir. Sonuç olarak pek çok durum için uygun model, sapmalı bir model ile varyansı büyük model arasındaki dengeyi kuran model olacaktır.

7.6.4 Ortalama Karesel Hata Kriteri

Bağımsız değişkenlerin uygun bir seti ve aday modellerin uygun fonksiyonel formu denge unsuru dikkate alınarak seçilmelidir. Sapma ve varyans arasındaki denge unsuru bir tahminleyicinin, örneğin bir kestirimin, ortalama karesel hatası kullanılarak başarılabilir. Her hangi $\mathbf{x}_0$ noktasındaki $\hat{y}_0$ kestirim değeri bu noktadaki $E(y_0)$ parametresinin bir tahminleyicisidir. Aday model için $\hat{y}_0$ tahminleyicisinin ortalama karesel hatası,

$$OKH\left(\hat{y}_0\right) = E\left[\hat{y}_0 - E\left(y_0\right)\right] \tag{7.9a}$$

$$OKH\left(\hat{y}_0\right) = V\left(\hat{y}_0\right) + \left[E\left(\hat{y}_0\right) - E\left(y_0\right)\right]^2 \tag{7.9b}$$

olup sapma ve varyansı birlikte değerlendiren bir kriterdir. Eşitlik (7.9*b*) nin ikinci bileşeni için eşitlik (7.5) kullanılarak, ihmal edilen parametre $\boldsymbol{\beta}_2$ vektörü olan eksik belirlenmiş model için,

$$OKH\left(\hat{y}_{01}\right) = \sigma^2 \mathbf{x}_{01}^T \left(\mathbf{X}_1^T \mathbf{X}_1\right)^{-1} \mathbf{x}_{01} + \left[\left(\mathbf{x}_{01}^T \mathbf{A} - \mathbf{x}_{02}^T\right)\boldsymbol{\beta}_2\right]^2 \tag{7.10}$$

elde edilebilir. Fakat eşitlik (7.10) doğrudan uygulanabilecek bir kriter değildir. Çünkü $\boldsymbol{\beta}_2$ vektörü bilinmemektedir ve $\mathbf{x}_{01}$ ile $\mathbf{x}_{02}$ vektörlerinin değerlerine bağımlıdır. $OKH(\hat{y}_0)$ değerinin $\mathbf{x}_{01}$ ve $\mathbf{x}_{02}$ vektörlerine olan bağımlılığının oluşturduğu problemi aşmak için, uyumu yapılmış bir $\hat{y}_i$ değerinin ortalama karesel hatası ele alınsın. Ortalama karesel hata bütün veri değerleri üzerinden toplanarak,

$$\sum \frac{OKH(\hat{y}_i)}{\sigma^2} = \sum \frac{V(\hat{y}_i) + [Sapma(\hat{y}_i)]^2}{\sigma^2} \tag{7.11}$$

Elde edilir. Eşitlik (7.11) standardize toplam hata olarak adlandırılır. Bu eşitliğin aday model için herhangi bir ekstrapolasyon ya da interpolasyon yeteneğini ortaya koymadığı açıktır. Çünkü veri setindeki gözlemler kullanılmıştır. Bununla birlikte eşitlik (7.11) in bir tahmini elde edilebilir. Bu tahmin kullanılarak sapma ve varyans arasındaki denge sağlanabilir. Eşitlik (7.11) in bileşenleri ayrı ayrı incelenmelidir. İlk aşamada araştırılan modelin p adet gerçek modelin ise r-p adet ilave parametre içerdiği varsayılsın. İlave değişkenlerin parametre vektörü $\boldsymbol{\beta}_2$ dir. Bu durumda i-inci veri noktasındaki kestirim değeri,

$$\hat{y}_i = \mathbf{x}_{i1}^T \mathbf{b}_1$$

olup burada $\mathbf{b}_1 = \left(\mathbf{X}_1^T \mathbf{X}_1\right)\mathbf{X}_1 \mathbf{y}$. Eşitlik (7.11) in varyans bileşeni,

$$\sum \frac{V(\hat{y}_{i1})}{\sigma^2} = \sum \mathbf{x}_{i1}^T \left(\mathbf{X}_1^T \mathbf{X}_1\right)^{-1} \mathbf{x}_{i1} \tag{7.12a}$$

eşitlik (7.12a) bir skaler olduğundan daha basit bir şekilde,

$$\begin{aligned}\sum \mathbf{x}_{i1}^T \left(\mathbf{X}_1^T \mathbf{X}_1\right)^{-1} \mathbf{x}_{i1} &= \sum tr \mathbf{x}_{i1}^T \left(\mathbf{X}_1^T \mathbf{X}_1\right)^{-1} \mathbf{x}_{i1} \\ &= \sum tr \mathbf{x}_{i1} \mathbf{x}_{i1}^T \left(\mathbf{X}_1^T \mathbf{X}_1\right)^{-1} \\ &= tr \sum \mathbf{x}_{i1} \mathbf{x}_{i1}^T \left(\mathbf{X}_1^T \mathbf{X}_1\right)^{-1}\end{aligned}$$

burada $\sum \mathbf{x}_{i1} \mathbf{x}_{i1}^T = \left(\mathbf{X}_1^T \mathbf{X}_1\right)$ olduğundan, standardize toplam ortalama karesel hatanın varyans bileşeni,

$$\sum \frac{V(\hat{y}_{i1})}{\sigma^2} = tr \sum \mathbf{x}_{i1} \mathbf{x}_{i1}^T \left(\mathbf{X}_1^T \mathbf{X}_1\right)^{-1} = tr \mathbf{I}_p = p \tag{7.12b}$$

elde edilir. Varyans bileşeninin, σ^2 hariç, parametre sayısına eşit olduğu görülmektedir. Eşitlik (7.12b) nin sağ tarafı aday model için izdüşüm matrisinin köşegen elemanlarının toplamını verir.

İkinci aşamada eşitlik (7.11) in sapma bileşeni ele alınsın. Bu bileşen eşitlik (7.6) dikkate alınarak tahminlenebilir. Daha önce belirtildiği gibi eksik belirlenmiş bir model için s^2 sapmalı olup sapma miktarı,

$$\sum \frac{[Sapma(\hat{y}_i)]^2}{n-p}$$

şeklindeydi. Sonuç olarak, σ^2 eğer bilinseydi, eşitlik (7.11) in tahmini,

$$C_p = p + \frac{\left(s^2 - \sigma^2\right)(n-p)}{\sigma^2} \tag{7.13a}$$

elde edilir. Bu tahmin Mallows un C_p istatistiği olarak adlandırılır. C_p istatistiğini ifade etmenin farklı yolları vardır. En çok kullanılanlarından biri de $KT(e)$ değerine göre yazmaktır. Eğer σ^2 nin bağımsız bir tahmini $\hat{\sigma}^2$ elde edilebiliyorsa, C_p istatistiği modeller arasında kıyaslamada kullanılabilecek çok faydalı bir kriterdir. Böyle bir tahminin mevcut olduğu durumlarda, p parametreli bir model için,

$$C_p = p + \frac{\left(s^2 - \hat{\sigma}^2\right)(n-p)}{\hat{\sigma}^2} \qquad (7.13b)$$

yazılabilir. En küçük C_p değerine sahip model uygun model olarak seçilir.

7.7 EN İYİ ALT KÜME İÇİN SEÇİM KRİTERLERİ VE DURDURMA KURALLARI

En iyi alt küme için seçim kriteri olarak kullanılan istatistiklerin pek çoğu artık kareler toplamlarının monoton fonksiyonudur. Bu nedenle özdeş alt küme modelleri verirler. Bununla birlikte seçilen kriterin farklı alt kümelerin seçilmesini önerdiği durumlarla da karşılaşılabilir. Böyle durumlarda alt küme modelleri arasındaki farklılıkların etkisi ortaya konulabilir. Bu durum özellikle daha ilerideki çalışmalarda kullanılmak amacıyla birkaç modeli karşılaştırarak tanımlamak açısından oldukça faydalıdır. Hemen, hemen tüm seçim prosedürlerinde ortak olarak kullanılan beş kriter kısaca açıklanacaktır. Bu kriterler;

1. Çoklu belirlilik katsayısı R^2.
2. Artık kareler ortalaması s^2.
3. Düzeltilmiş belirlilik katsayısı R_a^2.
4. Mallow'un *Cp* istatistiği
5. Anlamlılık seviyeleri, α.
6. Bilgi kriterleri *AIC* ve *SBC*.

şeklindedir. Kriterlerin açıklanmasına geçilmeden önce seçim yöntemlerinde önemli bir rolü olan sona erdirme kuralları incelenecektir.

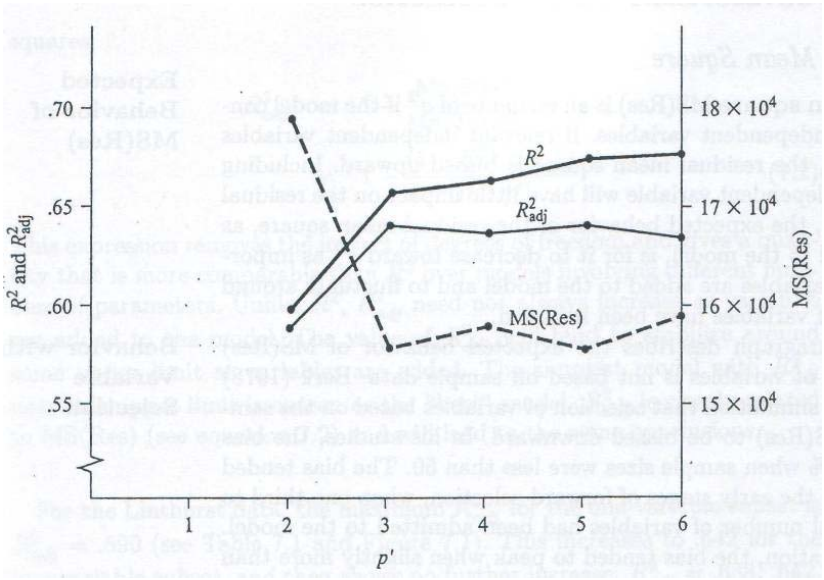
Adımsal (Stepwise) regresyon seçim metotları için hazırlanan bilgisayar programları genellikle seçim sürecini sona erdirecek bir kriter içerirler. *İleri doğru* (forward) *seçim yönteminde* ortak kriter, modelin mevcut artık kareler ortalaması ile aday değişkeni oluşturduğu artık kareler ortalamasının birbirine oranının indirgenmesine dayanır. Bu kriter, bir kritik *F*-testi değeri veya eşdeğer olarak bir kritik anlamlılık seviyesi değerine göre ifade edilir. Burada gerçekleştirilen *F*-testi, dikkate alınan değişkenin kısmı kareler toplamlarının *F*-testidir. Modele giriş kriterini sağlayan değişken kalmadığında ileri doğru seçim işlemi sona erer. Bu *F*-testi, klasik anlamlılık testinden çok bir sona erdirme kuralı olarak algılanmalıdır.

Geriye doğru (backward) *eleme yöntemi* için sona erdirme kuralı modeldeki kalan değişkenlerin en küçük kısmi kareler toplamlarının *F*-testidir. Geriye doğru eleme yöntemi, modelde kalan tüm değişkenler modelde kalma kriterlerini karşılanması durumunda sona erer. Adımsal regresyon seçim yöntemi sona erdirme kuralı hem ileri doğru seçim hem de geriye doğru eleme kriterlerini kullanır. Bu değişken seçim süreci, modeldeki tüm değişkenler modelde kalma kriterlerini karşılandığında ve

dışarıda modele girme kriterlerini karşılayan değişken kalmayınca sona erer. Bir değişkenin modele girmesi için kullanılan kriterin, değişkenin elenmesi için kullanılan kriterle aynı olması gerekli değildir. Seçim işlemini değişkenlerin daha büyük sayıdaki alt setlerini dikkate almak açısından daha güçlü hale getirebilmek için, giriş kriterin daha esnek olması avantajlıdır.

7.7.1 Belirlilik Katsayısı

Amaç *y*'deki değişkenliğin olabildiğince büyük kısmını açıklayabilen bir model seçmektedir. Modele bağımsız değişken eklendiğinde R^2 değeri azalmayacağı için, tüm bağımsız değişkenleri içeren model zorunlu olarak maksimum R^2 değerini verecektir. Modeldeki değişken sayısına karşı R^2 değerlerinin grafiği, en önemli değişkenler göz önüne alındığında maksimum R^2 yakınında düzleşen dik şekilde yukarı eğimli bir eğridir, bkz. Şekil 7.1. Model oluşturma amacıyla R^2'nin kullanılması durumunda, ilave değişkenin R^2'de oluşturduğu artış ile modelin yapısındaki karmaşıklığın artıp artmadığı değerlendirilmelidir. En iyi alt küme hacmi eğrinin düzleşmeye başladığı kıvrım yerine yakın bir noktada seçilir.



Şekil 7.1.Rawlings sf 221

R^2 istatistiğinin kullanımı aşamaları aşağıda verilmiştir.

1.Toplam değişken sayısından bir fazla (*r*+1) adet set oluşturulur.

Küme A sadece ortalama değeri içerir, model: $E(y)=\beta_0$

Küme B sadece tek değişkenli modeli içerir, model: $E(y)=\beta_0+\beta_1 x_1$

Küme C sadece çift değişkenli modeli içerir, model: $E(y)=\beta_0+\beta_1 x_1+\beta_2 x_2$

Böylece (*r*+1) adet küme elde edilir.

2 . Her bir küme için R^2 değerleri sıraya konur.

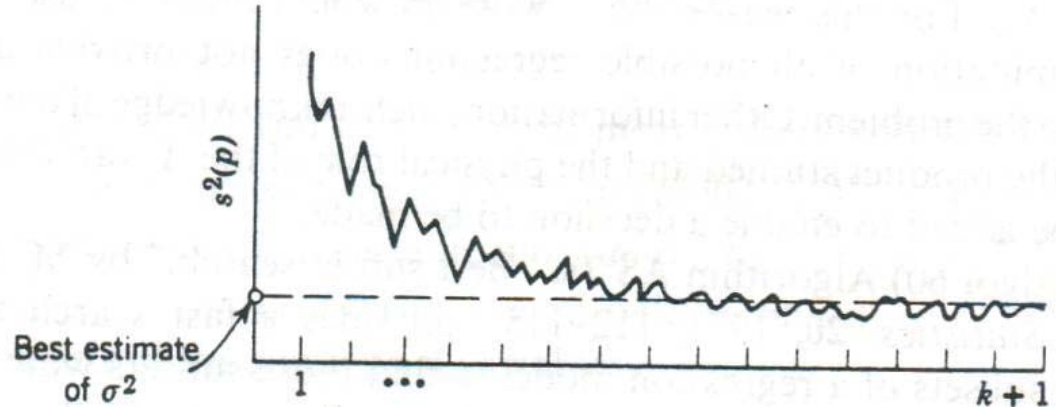
3. Her bir setteki büyük R^2 değerine sahip denklemler, içerdikleri değişkenlerin tutarlılığı açısından incelenir.

Her hangi bir küme içinde en büyük iki R^2 değeri arasındaki farkın çok az olması durumunda, bir sonraki küme için R^2 de oluşacak artış miktarının az olması beklenir. Bunun nedeni kalan değişkenlerle modeldeki değişkenler arasındaki korelasyon yapısının yüksek olmasıdır.

7.7.2 Artık Kareler Ortalaması

Artık kareler ortalaması, tüm bağımsız değişkenlerin modelde içerilmesi durumunda σ^2'nin bir tahminidir. Eğer uygun bağımsız değişkenler ihmal edilmiş ise artan bir sapma gözlenir. Önemsiz bir bağımsız değişkenin modelde içerilmesinin artık kareler ortalaması üzerindeki etkisi önemsizdir. Modele değişkenler eklendiğinde artık kareler ortalamasını iki farklı davranışta bulunabilir; önemli bağımsız değişkenler modele ilave edilmiş ise σ^2'ye doğru azalır, tüm uygun değişkenler modele girmiş ise σ^2 etrafında dalgalanır. Artık kareler ortalamasının bu davranışları değişken seçiminin örnek verileri üzerine kurulmadığı durumlar için geçerlidir.

Berk (1978)yapmış olduğu simülasyon çalışmasında, örnek verisine dayanan değişken seçimi de artık kareler ortalamasındaki sapmanın azalan yönde olduğunu göstermiştir. Çalışmasında, örnek hacminin 50'den az olduğu durumlarda sapmanın %25'den fazla olduğunu belirtmiştir. İleri doğru seçim yönteminde toplam değişken sayısının 1/3'ü ile ½'si arasındaki değişken modele dahil edildiğinde sapmanın değeri en üst noktaya ulaşmaktadır. Geriye doğru eleme yönteminde ise, değişkenlerin yarısından biraz fazlası modelden elendiğinde sapma değeri en üst noktaya ulaşır. Bu sonuçlara göre, bir değişken seçim prosedüründe değişkenleri eklendikçe s^2 değeri seçim işleminin orta aşamalarında σ^2 değerinin biraz altına düşecektir, tam modele yaklaşıldıkça tekrar σ^2 değerine yaklaşacaktır. Fakat bu büyüklükteki bir sapmanın (özellikle sapmanın ciddi olmadığı küçük örneklerde) s^2'nin değişken sayısına karşılık grafiğinde fark edilmesi pek olası değildir, bkz Şekil 7.2.



Şekil 7.2 Draper sf 298

Değişkenler modele eklendikçe elde edilen artık kareler ortalamaları modelin önemli bağımsız değişkenleri içerip içermediğinin yorumlanmasında kullanılabilir. Bağımsız değişken sayısının fazla ve tekrarlı gözlemlerin olduğu büyük problemlerde, artık kareler toplamının modeldeki parametre sayısına karşı plotu eğrinin düzleştiğini gösterecektir.

Modeldeki potansiyel değişen sayısı büyük olduğunda, örneğin r>10 ve veri noktaları sayısının da r'den çok büyük olması durumunda, örneğin 5r ile 10r gibi, s_p^2 plotu oldukça bilgilendirici olacaktır. Bununla birlikte küçük örneklerde bu durumla karşılaşılması beklenen bir durum değildir.

7.7.3 Düzeltilmiş Belirlilik Katsayısı

Düzeltilmiş belirlilik katsayısı R_a^2, çoklu belirlilik katsayısının serbestlik derecesi ile ölçeklenmiş halidir. Bu istatistik kareler toplamından daha uygun olan kareler ortalamasının bir oranını içerir. Bölüm 4 deki eşitlik (4.62) den hatırlanacağı gibi bu ifade serbestlik derecesinin etkisini ortadan kaldıran ve farklı sayıda parametre içeren modellerin karşılaştırılmalarında R^2'ye göre daha iyi sonuç verir. R^2'den farklı olarak R_a^2 değeri modele değişken eklendiğinde her zaman artmaz. R_a^2 değeri değişkenler ilave edildikçe bir üst limitte durağan hale gelir. Bu üst limite yakın R_a^2 değerli en basit model en iyi model olarak seçilir. R_a^2 istatistiği s^2 ile ilişkili olduğu için sonuçlar benzerdir.

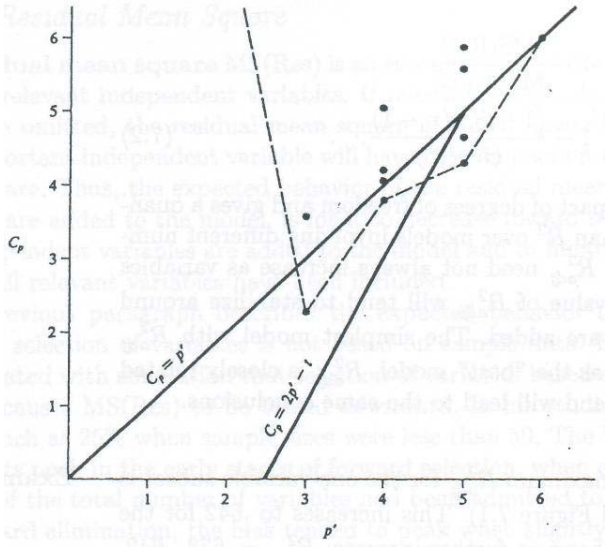
7.7.4 Mallow'un C_p İstatistiği

C_p istatistiği mevcut veri seti için tahmininin standardize toplam ortalama karesel hatasının bir tahminini ifade eder. C_p istatistiği ve C_p grafiği ilk olarak Mallow (1973) tarafından tanımlanmıştır.

$$C_p = \frac{KT(e)_k}{s^2} + 2p - n \tag{7.14}$$

Burada $KT(e)_k$, k-değişkenli alt modelden elde edilen artık kareler toplamıdır. s^2 ise tüm bağımsız değişkenleri içeren modele göre elde edilen σ^2'nin bir tahminidir. Eğer p parametreli bir model yeterli ise bu durumda uyum yetersizliği ortaya çıkmayacağı için $E[KT(e)]=(n-p)\sigma^2$. $E(s^2)=\sigma^2$ olduğundan sonuç olarak $E(C_p)=p$ bulunur. Önemli bir değişkenin modelden elenmesi durumunda Önemli bir değişkenin modelden elenmesi durumunda artık kareler toplamı, $(n-p)$ σ^2'nin tahmini artı pozitif bir değere eşittir. Bu pozitif değer elenen değişkenin yaptığı katkıyı belirtir. Sonuçta bu gibi durumlarda C_p değerinin p değerinden büyük olması beklenir.

C_p grafiği, C_p'yi daha iyi alt modeller için p'nin fonksiyonu olarak ifade ettiği için, alt küme büyüklüğünün seçilmesinde uygun bir metottur. Her bir alt set büyüklüğü için minimum C_p, p değeri küçük olduğunda p'den oldukça büyük olacaktır. Önemli değişkenler modele eklendikçe p değerine doğru azalacak ve en sonunda p değeri çevresinde küçük değişimler gösterecek veya p değerinin altına inecektir. Tam modelin artık kareler ortalaması s^2 olarak ele alındığında tam model için C_p değeri p'ye eşit olacaktır. C_p'nin p değerine yakın bir değeri, σ^2'nin tahmini olan s^2'de küçük bir sapma olduğunun göstergesidir. Bu yorum C_p'nin paydasındaki s^2'nin, σ^2nin sapmasız bir tahmini olduğunu varsaymaktadır. Unutulmamalıdır ki, tam modelden elde edilen s^2'nin σ^2'nin sapmasız bir tahmini olabilmesi sadece tam modelin tüm uygun değişkenleri içermesi durumunda gerçekleşebilir.



Şekil 7.3 Rawlings

En iyi model C_p grafiğinin incelenmesinden sonra seçilir, bkz. Şekil 7.3. Seçim işlemi, yaklaşık olarak p değerine eşit düşük C_p değerlerine sahip bir modelin araştırılmasıdır. Bu seçimin kesin ve açık olarak yapılmadığı durumlarda, kişisel değerlendirmeler aşağıdaki iki durumdan birini tercih eder:

Durum 1. Sapmalı bir denklem: Bu denklem büyük bir $KT(e)_p$ değerine sahip olduğundan gerçek verileri iyi bir şekilde temsil edemez. Fakat gerçek modele göre daha küçük bir C_p değerine sahiptir. Diğer bir deyişle ortalama karesel hatası daha küçüktür.

Durum 2. Daha fazla parametreli bir denklem: Bu denklem $C_p=p$ olduğundan gerçek verilere daha iyi uyum sağlar fakat gerçek modele göre ortalama karesel hatası daha büyüktür.

Başka bir deyişle küçük modeller küçük C_p değerine sahip olup daha büyük modellerin C_p değeri modelin p değerine daha yakındır.

C_p'nin kullanılmasında farklı kriterler geliştirilmiştir. Mallows (1973) ,küçük C_p değerli ve p değerine yakın C_p değerli modellerin araştırmanın ileri aşamalarında kullanılmasını önermiştir. Hocking (1976) modelin kestirim için yada parametre tahminlemesi için kullanılmasına bağlı olarak iki farklı kriter tanımlamıştır. Kestirim modelleri için $C_p \leq p$ kriterini kullanmıştır, (Durum 1). Parametre tahminlemesinde ise tahminlerdeki aşarı sapmayı engellemek için modelden elenecek değişken sayısının mümkün olduğunca az olmasını önermiş ve $C_p \leq 2p-r$ seçim kriterini kullanmıştır, (Durum 2) burada r tam modeldeki parametre sayısıdır.

7.7.5 Alt Set Büyüklüğü Seçimi İçin "Anlamlılık Seviyeleri"

Adımsal değişken seçim metotlarında *anlamlılık seviyelerinin* tüm alt küme büyüklüklerini ele almadan önce seçim sürecini sona erdirecek şekilde seçilmeleri durumunda bu yaklaşım bir alt küme büyüklüğü seçim kriteri olarak kullanılabilir. Bendel ve Afifi (1977), ileri doğru seçim yöntemi için birkaç sona erdirme kuralını karşılaştırmış ve sabit bir anlamlılık seviyesi üzerine kurulmuş ardışık F-testinin oldukça uygun olduğunu göstermişlerdir. Optimum anlamlılık seviyesi 0.15 ile 0.25 arasında değişmektedir. Ardışık F-testi her ne kadar kriterlerin en iyisi olmasa da α= 0.15 için $n-p \leq 20$ olduğu

durumlarda yaklaşık en iyi kriterdir. Eğer α yaklaşık olarak 0.20 olduğunda n-p≥40 durumu için ardışık F-testi yerine C_p istatistiği tercih edilmektedir. Benzer sonuçlar, değişkenlerin önem sırasının bir ön bilgi olarak mevcut olması durumunda Kennedy ve Bancroft (1971) tarafından da elde edilmiştir. Bu çalışmada ileri doğru seçim yöntemi için anlamlılık seviyesinin 0.25, geriye doğru eleme yöntemi için 0.10 olması gerektiği belirtilmiştir. Bendel ve Afifi geriye doğru eleme yönteminde anlamlılık seviyesinin seçimi için bir yorum yapmamışlardır. Alt setin büyüklüğünün seçiminde sona erdirme kuralı ile ilgili sonuçlar özet olarak aşağıda sunulmuştur.

1. n-p çok büyük olmadıkça tüm bağımsız değişkenlerin kullanılması oldukça kötü bir kuraldır.
2. Eğer n-p≤10 ise sona erdirme kurallarının pek çoğu yetersiz kalmaktadır. C_p istatistiği de buna dahildir, n-p≥40 için tavsiye edilmektedir.
3. (r-p)adet değişkenin uyum yetersizliği testi, anlamlılık seviyesi kullanılsın ya da kullanılmasın genellikle yetersiz bir sona erdirme kuralıdır.
4. Belirlilik katsayısının sapmasız bir versiyonun kullanılması da n-p büyük olmadıkça yetersiz kalmaktadır. Bu durum muhtemelen R_a^2, R^2 ve s^2'nin de yetersiz olabileceğini belirtir.

Bu kısımdaki açıklamalardan da görüldüğü gibi C_p istatistiği ve anlamlılık seviyeleri alt set büyüklüğünün seçiminde kullanılan en gözde kriterdir. C_p istatistiği küçük ve orta büyüklükteki örnek setlerinde zayıf kalmaktadır. Orta büyüklükteki örnek setleri için anlamlılık seviyesi, C_p 'ye oranla biraz daha iyi sonuçlar vermektedir. Küçük örneklerdeki C_p nin zayıf performansı genel bir yetersizlik olarak algılanmalıdır. Çünkü n-p≤10 için bütün kriterlerin performansı düşüktür.

7.7.6 Bilgi Kriterleri AIC ve SBC

Akaike (1969) bilgi Kriteri AIC;

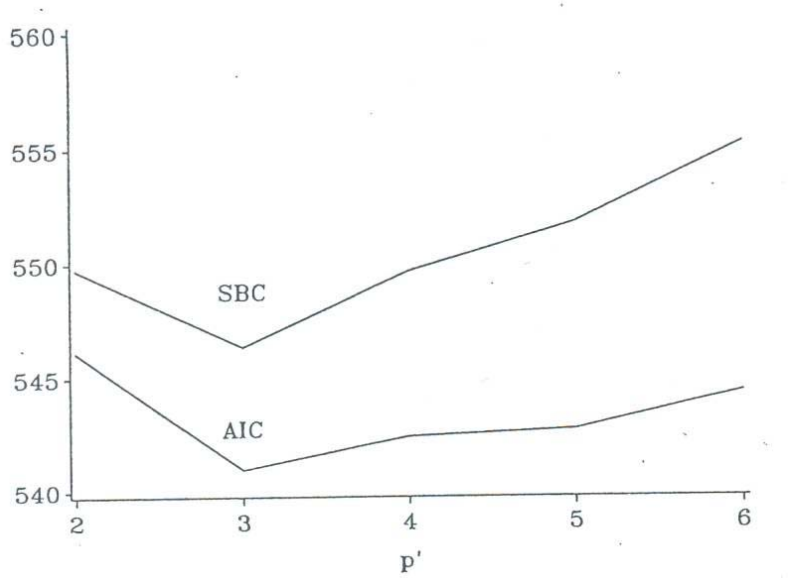
$$AIC_p = n\ln\left[KT(e)_p\right] + 2p - n\ln(n) \tag{7.15}$$

eşitliği ile tanımlanmıştır. Modeldeki bağımsız değişken sayısı arttıkça $KT(e)$ azalacağından AIC'deki ilk terim p'ye bağlı olarak azalacaktır. Bununla birlikte ikinci terim p'ye bağlı olarak artacaktır. Bu durum modeldeki artan parametre sayısının oluşturduğu bir ceza olarak düşünülebilir. Sonuç olarak AIC uyum hassasiyetinde oluşan iyileşme ile bu uyumu sağlamak için kullanılan parametre sayısı arasında bir denge oluşturmaya odaklanmıştır. AIC_p değerlerinin p değerine karşı bir grafiği oluşturulur ve minimum AIC değerine ait p değeri alt set büyüklüğü olarak belirlenir, bkz Şekil 7.4.

AIC geniş bir kullanıma sahip olmakla birlikte, belirlenen alt set büyüklükleri genellikle gerçek modelden daha fazla terim içermektedir. AIC'nin daha daha fazla sayıda bağımsız değişken içeren modellere yönelmesi nedeniyle birkaç alternatif kriter geliştirilmiştir. Alternatif kriterlerden biri Schwarz (1978) Bayesyan Kriteri olup,

$$SBC_p = n\ln\left[KT(e)_p\right] + p\ln(n) - n\ln(n) \tag{7.16}$$

eşitliği ile verilmiştir. İki kriter arasındaki temel fark ikinci terimde *SBC*'nin ln(n) çarpanını kullanmasıdır. Sonuçta parametre sayısının artması durumunda ortaya çıkan ceza *AIC*'ye göre daha fazla olmaktadır. Uygun alt set büyüklüğü minimum *SBC* değerini veren p değeri olarak belirlenir, bkz Şekil 7.4.



Şekil 7.4 Rawlings sf226.

7.8 EN İYİ REGRESYON DENKLEMİNİN SEÇİMİ

Bu bölümde buraya kadar genel olarak model kurma süreci ile ilgilenildi. Bu kısım ise denklemdeki değişkenleri seçmek için kullanılan bazı istatistiksel prosedürlere ayrılmıştır. Bir y yanıtı için $x_1, x_2,...,x_k$ bağımsız veya kestirici değişkenlerine göre doğrusal regresyon denkleminin oluşturulmak istendiği varsayılsın. Değişken seçimi için Kısım 7.6 'da verilen özelliklerin yanı sıra seçim sırasında dikkate alınması gereken birbirinin zıttı iki genel kriter vardır:

Bu kriterler arasında oluşturulan uzlaşma *en iyi regresyon denkleminin seçimi* olarak adlandırılır. Bu amacı gerçekleştirebilecek her hangi bir problem için prosedür sayısı birden fazladır. Eğer bir gözlemlerin gerçek rassal varyansının σ^2, değeri bilinseydi en iyi regresyon denkleminin seçimi çok daha kolay olacaktı. En iyi regresyon denkleminin seçiminde kullanılan bazı yöntemler aşağıda verilmiştir.

1. Tüm mümkün regresyonlar
2. En iyi alt set regresyonu
3. Geri doğru eleme
4. İleri doğru seçim yöntemi
5. Adımsal regresyon

Bu kısımda yukarıda bahsedilen değişken seçim metotlarından kısaca açıklanacaktır.

7.8.1Tüm Mümkün Regresyonlar

Bilgisayar yardımı olmadan bu prosedürün uygulanması hemen, hemen imkansızdır. Bu prosedür ilk olarak, x_0 ve her hangi sayıdaki $x_1...,x_r$ değişkenlerini içeren tüm mümkün regresyon denklemlerinin

uyumunu gerektirir. x_0 terimi daima denklemde bulunduğu için uyumu yapılması gereken denklem sayısı 2^r adettir. Görüldüğü gibi r sayısı büyük ise uyumu yapılması gereken denklem sayısı da oldukça büyük olacaktır. Her bir regresyon denklemi R^2, s^2 ve C_p istatistiklerine göre değerlendirilir. Daha önceden de belirtildiği gibi bu istatistikler birbiriyle ilişkilidir. Kullanılabilecek en iyi denklemin hangisi olduğunun seçimi gözlenenmiş örnekler değerlendirilerek gerçekleştirilir.

7.8.2 En İyi Alt set Regresyonu

Kullanılan seçim yöntemlerinden biri de en iyi alt setlerin bulunmasıdır. Bu yöntemde, tüm mümkün alt setler hesaplanmadan her bir alt set büyüklüğündeki en iyi alt setler tanımlanır. Bu metotlar EKK'nın temel özelliği olan, modelden bir değişken eksiltiğinde artık kareler toplamı azalmaz, mantığına dayanmaktadır. Farklı alt set modellerinin artık kareler toplamlarının karşılaştırılması yöntemi ile diğer alt setlerin hesaplanma ihtiyacı ortadan kaldırılır. Örneğin ,iki değişkenli bir alt setin artık kareler toplamı bir üç değişkenli modelinkinden daha az ise bu üç değişkenli modelin aynı iki değişkeni içeren alt setlerinin hiçbiri hesaplanmaz,çünkü onların artık kareler toplamı iki modelinkinden daha büyük olacaktır.

Değişken sayısının fazla olduğu problemlerde bilgisayar programı olmaksızın bu yöntemin kullanılması imkansızdır. Bu amaçla SAS, SPSS, MINITAB gibi programlar kullanılabilir. Amaç en iyi K adet alt kümenin belirlenmesinde, maksimum R^2, maksimum R_a^2 ve Mallows C_p istatistikleri kullanılabilir. Araştırmacı istediği en iyi alt set sayısını K ve kriteri belirler. Daha sonra program tek değişkenli en iyi K alt seti, iki değişkenli en iyi K alt seti ve sonunda tüm değişkenleri içeren en iyi tek alt seti elde eder. Eğer seçilen K değeri bir alt setin mevcut denklem sayısını aşıyorsa tüm mümkün denklemler incelenir. Bu yöntemin kullanılmasının bazı olumsuz yönleri aşağıda verilmiştir.

1. Yöntem en iyi K alt set denklemlerinde gereğinden fazla sayıda kestirici değişken içerilmesine olanak vermektedir.
2. Eğer K çok küçük seçilmiş ise uyumu yapılabilecek en duyarlı denklem en iyi K alt kümede içerilmeyebilir.

Eğer en iyi civarındaki denklemlerin incelenmesi isteniyor ise adımsal regresyon metodu ile birlikte kullanılması tavsiye edilir.

7.8.3 Geriye Doğru Eleme Metodu

Geriye doğru eleme metodu, sadece belirli sayıda değişken içeren en iyi modelleri incelemek ile ilgilendiği için tüm mümkün regresyon metoduna göre daha ekonomiktir Bu prosedürün temel adımları aşağıda tanımlanmıştır.

1. Tüm değişkenleri içeren bir regresyon denklemi hesaplanır.
2. Her bir kestirici değişken için o değişkene, regresyon denklemine giren en son değişken muamelesi yapılarak kısmi F-testi değeri hesaplanır.
3. F_L ile gösterilen en düşük kısmi F-test değeri, önceden seçilen F_0 anlamlılık seviyesi ile karşılaştırılır.

a) Eğer $F_L<F_0$ ise x_i değişkeni denklemden çıkarılır ve regresyon denklemi kalan değişkenlerle Adım 2'den tekrar hesaplanır
b) Eğer $F_L>F_0$ ise değişken regresyon denkleminde kalır ve hesaplanır.

Geriye doğru eleme metodu ile tüm değişkenleri içeren modelden başlayarak alt set modelleri seçer. Bu seçim için her bir işlemde regresyon kareler toplamında en az artışı oluşturan değişken elenir. Yukarıda belirtilen F_0 gibi bir sona erdirme kuralının dikkate alınması durumunda metot, en iyi alt küme modeli bir değişkeni içerinceye kadar çalışır.

Bu metotda ilk olarak tüm değişkenleri içeren model için EKK denklemi bulunur. $\mathbf{X}^T\mathbf{X}$ matrisi tekil olmadığı için elde edilen artık hata varyansı σ^2'nin iyi bir tahminini verir. Geriye doğru eleme yöntemi temelde bu asimtotik σ^2 değerinin büyüklüğünde gerçekçi bir artış oluşturanlar haricinde tüm gereksiz açıklayıcı değişkenlerin elenmesini esas alır.

İkinci aşamada en küçük kısmi F-değeri seçilir ve belirlenmiş α riskini baz alan F-kritik değeri ile karşılaştırılır. En küçük kısmi F-değerine sahip değişken x_j olsun. Hesaplanan F değeri kritik değerden küçük ise x_j değişkeni elenir. Daha sonra kalan değişkenlerin bulunduğu model için EKK denklemi elde edilir ve yukarıdaki işlem tekrarlanır.

Modelde kalan değişkenlerin kısmi F-değerleri kritik değeri aştığında için prosedür sona erer. Bu metot oldukça tatmin edicidir. Çünkü her hangi bir şeyi gözden kaçırmak istemeyen istatistikçiler en azından tüm değişkenleri içeren denklemi görmek isterler. Ayrıca tüm mümkün regresyon ile karşılaştırıldığında zaman açısından daha ekonomiktir. Bununla birlikte $\mathbf{X}^T\mathbf{X}$ matrisi yaklaşık tekil ise aşırı uyumlu denklem yuvarlama hatalarına bağlı olarak anlamlı olmayabilecektir. Bu metotda elenen değişken bir daha asla dikkate alınmaz. Elenmiş değişkenleri içeren tüm alternatif modeller incelenmezler.

7.8.4 İleri Doğru Seçim Metodu

İleri doğru seçim yöntemi, işlemin her aşamasında modele bir değişken ilave ederek alt set modellerini seçer. Metot bağımlı değişkendeki değişkenliğin en büyük kısmını açıklayan bağımsız değişkeni içinde bulunduran modelle başlar. Bu değişken Y ile en yüksek kısmi korelasyona sahiptir. Her bir adımda artık kareler toplamında en büyük azalışa neden olan değişken modele ilave edilir. Bu değişken mevcut modelin artıkları ile en yüksek korelasyona sahiptir. Eğer bir sona erdirme kriteri tanımlanmamış ise, ileri doğru seçim yöntemi tüm değişkenler modele girinceye kadar devam eder.

7.8.5 Adımsal Regresyon Metodu

Geriye doğru eleme yöntemi tüm değişkenleri içeren en geniş denklemle başlar ve kullanılabilecek bir denklem elde edilinceye kadar değişkenler elenir. Adımsal regresyon seçim metodu ise tam ters yönden hareket ederek başlar. Denkleme giriş sıralaması ise kısmi korelasyon katsayısı kriteri ile gerçekleştirilir. Temel prosedür aşağıda tanıtılmıştır. İlk olarak y ile en yüksek korelasyona sahip değişken seçilir. Daha sonra birinci dereceden $y=f(x_j)$ doğrusal regresyon modeli kurularak değişkenin

anlamlı olup olmadığı kontrol edilir. Değişken anlamlı değil ise en iyi model olarak $y = \bar{y}$ modeli oluşturulur. Daha sonra model girecek ikinci bir değişken araştırılır. Bu aşama matematiksel olarak;

1. $y=f(x_1)$ regresyonundan elde edilen artıklar ya da
2. Her bir x_j için, $x_j=f(x_1)$ regresyonundan elde edilen artıklar arasındaki korelasyonların bulunmasına eş değerdir.

Daha sonra y ile en yüksek kısmi korelasyona sahip x_2 seçilir ve $y=f(x_1, x_2)$ regresyon denkleminin uyumu yapılır. Uyumu yapılan denklemin anlamlı olup olmadığı kontrol edilir, R^2 değerindeki gelişme belirlenir ve denklemdeki her iki değişkenin kısmi F-değerleri incelenir. Bu iki kısmi F-değerlerinden küçük olanı bir F yüzde noktası ile karşılaştırılır. Test sonucunun anlamlı olup olmamasına göre ilgili değişkenin denklemde kalması ya da elenmesine karar verilir. Mevcut denklemdeki en az faydalı değişkenin testi adımsal regresyon metodunun her adımında gerçekleştirilir. Daha önceki adımlarda modele girmeye en iyi aday olan bir değişken daha sonraki bir adımda, denklemdeki diğer değişkenlerle arasındaki ilişki nedeniyle gereksiz hale gelebilir. Bunu kontrol edebilmek için çalışmanın herhangi bir adımında denklemdeki en küçük kısmi F-değeri belirlenir ve önceden seçilmiş kritik F-değeri ile karşılaştırılır. Böyle bir yaklaşım, değişkenin model giriş aşamasını dikkate almaksızın modeldeki en önemsiz değişkenin bir değerlendirilmesini sağlar. Eğer test edilen değişkenin katkısı önemsiz ise modelden çıkarılır. Daha sonra modelde bulunmayan değişkenlerin en iyisinin kısmi F-testini aşıp modele girip giremeyeceğini kontrol edilir. Testi geçmesi durumunda modele girer ve tekrar modeldeki tüm değişkenler için kısmi F-değerleri test edilir. Eğer mevcut denklemdeki hiçbir değişken elenmiyor ve daha sonra en iyi aday değişken modele giremiyor ise süreç sona erer. Modele giren her değişkenin R^2 üzerindeki etkisi kayıt edilir.

Bu metot açıklanan tüm diğer metotlar içerisinde en uygun olarak göze çarpmaktadır. Bununla birlikte açıklanan metotların bazıları birlikte kullanılarak daha iyi sonuçlar elde edilmeye çalışılabilir. Çünkü ele alınan metotlar mutlak en iyi model değil kullanım için kabul edilebilir bir model seçmeye yöneliktir.

7.8.6 Metotların Birlikte Kullanılması

Öneri 1: Kabul ve ret kriterlerini belirleyerek adımsal regresyon ile başla. Seçim prosedürü sona erdiğinde modeldeki değişken sayısını r belirle, r değişkenli model için p değişkenli tüm mümkün regresyonlar içinden en iyi seti seç.

Değişkenlerin daha büyük bir setinin daha iyi bir sonuçlar verdiği durumlarda, adımsal regresyon metodu bu modeli incelemediği için önerilen yaklaşım tarafından ortaya çıkarılmayacaktır Önerilen prosedürün katkıları az fakat ilave hesaplamaları fazladır.

Öneri 2: Adımsal regresyonda daha az kısıtlayıcı (daha büyük α) kabul ve red seviyeleri kullanılarak birkaç değişkenin daha modele girmesinin sağlanmasıdır. Böyle bir yaklaşım, genel adımsal regresyon yaklaşımında modele giremeyen bazı değişkenlerin incelenmesini ve farklı bir modelin elde edilmesini sağlar.

Alternatif modelin elde edilmesi oldukça faydalı olmakla birlikte, kestirici değişkenler arasında yüksek korelasyonun mevcut olması problem yaratır.

7.9 MODELİN GEÇERLİ KILINMASI

Uyumu yapılan regresyon denkleminin geçerli kılınması kullanım amacına uygun etkin bir model olduğunun doğrulanması ya da kanıtlanması anlamına gelmektedir. Bu ifade sadece uyumu yapılan denklemin elde veri seti ile uyumlu olduğunun ispatlanması anlamına gelmez. Modelin geçerli kılınması, uyumu yapılan denklemin, uyumda kullanılan verilerden bağımsız bir veri seti ile etkinliğinin değerlendirilmesini gerektirir.

Model kurma aşaması, model için pek çok matematiksel form ve bağımsız değişken kombinasyonlarının araştırılmasını içerir. Parametrelerin tahminlenmesinde EKK yöntemi kullanıldığında seçilen modelin gözlenmiş verilerle uyumlu olması beklenen bir durumdur. Bunun ötesinde gerçek model bilinseydi, uyumu yapılan modelin örnek veri seti ile gerçek modele göre daha uyumlu olduğu görülürdü. Bu nedenle uyumu yapılan model uyumu yapılan gözlemlerden bağımsız bir veri seti ile geçerli kılınabilir. Bu amaca yönelik kullanılabilecek bazı geçerli kılma yöntemleri:

1. Kestirim değerlerinin ve parametre tahminlerinin fiziksel teori ile kıyaslanması.
2. Uyumu yapılan modelin benzetim (simülasyon) verileri ile doğrulanması.
3. Modelin yeni elde edilen verilerle geçerli kılınması.
4. Veri bölme-çapraz geçerli kılma (cross-validation) teknikleri ile doğrulama.

Yukarıda açıklanan yöntemlerden ilk ikisi araştırılan sistem ile ilgili ön bilgiye gereksinim duymaktadır. Üçüncüsü ise yeni verinin elde edilmesini gerektirdiğinden yaygın bir kullanıma sahip değildirler. Aşağıda bu üç yöntem kısaca tanıtıldıktan sonra en yaygın kullanılan çapraz-geçerli kılma yöntemi açıklanacaktır.

Kestirim değerlerinin ve parametre tahminlerinin fiziksel teori ile kıyaslanması: Modelin kestirim değerlerinin ve parametre tahminlerinin kontrolü model seçimi gerçekleştirildikten sonra modelin temsil ettiği varsayılan sistem ile ilgili mevcut tüm ön bilgi kullanılarak gerçekleştirilmelidir. Bu ön bilgiler genellikle model parametrelerinin işaretleri ve/veya sayısal değerleri üzerinedir. Örneğin teorik olarak pozitif olması gereken bir parametre tahmininin negatif olarak tahminlenmesi uyumu yapılan modelin mevcut sistem için kullanılmasının uygun olmadığının önemli bir göstergesidir.

Uyumu yapılan modelin benzetim (simülasyon) verileri ile doğrulanması: Bazı araştırmalarda teorik bir model var olabilir. Fakat model uygulamalı kullanım için çok karmaşık yapıda olabilir. Bu teorik model kullanılarak benzetim çalışmaları ile doğrulama verileri türetilebilir. Uyumu yapılan regresyon modeli bu veri seti kullanılarak doğrulanabilir.

Modelin yeni elde edilen verilerle geçerli kılınması: Modelin doğrulanması konusunda diğer iyi bir yöntem modelin tahminleri ile karşılaştırılabilen yeni verilerin toplanmasıdır. Bir modeli geliştirmede ve parametrelerini tahmin etmede kullanılan matematiksel ve fiziksel varsayımların geçerliliği, eğer model yeni veriler ile doğru kestirimler üretmiyor ise sorgulamaya açıktır. Başarılabildiği durumlarda

yeni verilerin toplanması tüm model oluşturma süreci üzerinde tam bir kontrolün elde edilmesini sağlar. Bununla birlikte yeni verinin toplanması her zaman mümkün olmamaktadır.

Veri bölme-çapraz geçerli kılma (*cross-validation*) *teknikleri ile doğrulama*: Çapraz geçerli kılma yönteminde modelin bağımsız bir veri seti ile geçerli kılınmasını sağlamak için kullanılan iki temel yaklaşım vardır. İlk yaklaşımda veri seti ikiye ayrılır ve bir grup veri ile modelin uyumu yapılır. İkinci grup ile model geçerli kılınır. İkinci yaklaşımda ise tahminlenmiş $\hat{y}_i$ değeri ile y_i gözlemi arasındaki bağımsızlığı sağlayabilmek için $\hat{y}_i$ kestirim değerleri *i*-inci gözlem olmadan uyumu yapılan modellerden elde edilerek modelin geçerli kılınmasında bir test istatistiği kullanılır.

7.9.1 Veri Setinin Ayrılması

Araştırmacılar modelin kurulma aşamasındaki şartların kullanım aşamasında da geçerli olacağı hakkında genellikle bilgi sahibi değildirler. Bir modelin geçerli kılınmasında araştırmacı modelin olabildiğince gerçekçi olduğuna inanmalıdır. Modelin geçerli kılınması bu açıdan oldukça önemlidir. Pratikte veri setinin ayrılması modelin geçerli kılınması için çalışmalar yapılabilmesine imkan vermektedir. Veri seti aşağıda gösterildiği gibi ikiye ayrılır; ilk bölüm uyum örneği ikinci bölüm geçerli kılma örneğidir:

$$\left.\begin{matrix} y_1 & x_{11} & \cdots & x_{1k} \\ y_2 & x_{21} & \cdots & x_{2k} \\ \vdots & \vdots & \cdots & \vdots \\ y_{n_1} & x_{n_1 1} & \cdots & x_{n_1 k} \end{matrix}\right\} \textit{uyum örneği}$$

$$\left.\begin{matrix} y_{n_1+1} & x_{n_1+11} & \cdots & x_{n_1+1k} \\ y_{n_1+2} & x_{n_1+21} & \cdots & x_{n_1+2k} \\ \vdots & \vdots & \cdots & \vdots \\ y_{n_1+n_2} & x_{n_1+n_2 1} & \cdots & x_{n_1+n_2 k} \end{matrix}\right\} \textit{geçerli kılma örneği}$$

Uyum örneği kullanılarak aday modelin uyumu yapılır. İkinci aşamada aday model için geçerli kılma örneği kullanılarak parametre tahminleri ve kestirim yeteneği doğrulanır. Bu aşamada kestirim hatasına (artıklara) dayanan bazı istatistikler,

$$\sum_{n_1+1}^{n_1+n_2}\left(y_i-\hat{y}_i\right)^2 \text{ ve } \sum_{n_1+1}^{n_1+n_2}\left|y_i-\hat{y}_i\right|$$

hesaplanabilir. Bu istatistikler daha sonra aday modelleri kıyaslamada kullanılabilir. Bu yaklaşımın kullanılmasında karşılaşılabilecek temel bazı temel zorluklar da mevcuttur. Bunlardan en önemlisi veri setinin nasıl ayrıştırılacağıdır. Bu konuda genel kabul görmüş bir metot yoktur. Problem araştırma konusuna göre farklı şekillerde incelenmektedir. Örneğin, veriler zamana bağlı olarak elde edilmiş ise en son veriler geçerli kılma örneği olarak ayrılırlar ve böylece modelin kestirim yeteneği ile ilgili daha fazla bilgi edinilir. Farklı bir çalışmada ise bağımsız değişkenlerin belirli bir bölgesi için bağımlı değişkenin davranışının tahminlenmesi diğer bölgelere göre daha önemli olabilir. Önemli bölgeye düşen veri noktaları bu çalışma içim doğrulama örneğini oluştururlar. Diğer önemli bir problem ise bu iki örneğin hacimlerinin nasıl belirleneceğidir. Bu konuda da uygulanabilecek genel bir kural yoktur.

Gözlemlerin uyuma ayrılmış kısmı artıklar için yeterli bir serbestlik derecesi oluşturabilmelidir. Kullanılabilecek bir yaklaşıma göre $n \geq 2p+20$ olmalı ve bu kuralın sağlandığı durumlarda uyum ve geçerli kılma örnek hacimleri eşit olmalıdır.

En iyi alt kümenin seçiminde ya da model durağanlığı ve genel kestirim performansı çalışmalarında çapraz geçerli kılma mekanizması olarak veri setinin ayrılması yönteminin kullanılması durumunda, seçilen sonuç modele son aşamada bütün veri seti kullanılarak regresyon uyumu yapılır. Böylece veri setindeki tüm bilgi kullanılmış olur. Veri setinin ayrılmasındaki amaç modelin fonksiyonel yapısının ya da bağımsız değişkenlerin alt kümesinin araştırılmasıdır. Bununla birlikte parametrelerin son tahminlenmesi bütün veri seti kullanılarak gerçekleştirilmelidir.

7.9.2 PRESS İstatistiği

Model oluşturma çalışmalarının pek çoğunda geçerli kılma amacıyla veri setinin ayrılması pratik bir yaklaşım olmamaktadır. Geçerli kılma amacıyla kullanılabilecek önemli bir kriter PRESS istatistiğidir. Hatırlanacağı üzere geçerli kılmada kullanılacak $\hat{y}_i$ ile y_i değerlerinin birbirinden bağımsız olması istenmekteydi, diğer bir ifade ile bu değerlerden elde edilecek artıkların da,

$$u_i = y_i - \hat{y}_i \tag{7.17a}$$

birbirinden bağımsız olması istenir. Örnek içinden ilk gözlemin çıkartıldığı bir veri seti ele alınsın. Kalan n-1 gözlem kullanılarak aday modelin parametreleri tahminlensin. Daha sonra çıkarılan ilk gözlem veri setine dahil edilerek ikinci gözlem çıkarılsın ve tekrar regresyon uygulansın. Sırasıyla her defasında bir gözlem çıkarılarak bu işlem n defa uygulansın. Bu işlemlerden her defasında veriden çıkarılan gözlem için yanıt değeri tahminlenebilir. Bu değerler kullanılarak, PRESS artıkları,

$$e_{(i)} = y_i - \hat{y}_{i(i)} \tag{7.17b}$$

elde edilebilir. Alt indisteki parantez ilgili tahmin değeri elde edilirken kullanılmayan gözlemi tanımlar. PRESS artıkları y_i değerinden bağımsız olarak tahminlenen $\hat{y}_{i(i)}$ değerini kullanan gerçek kestirim hatasını tanımlamaktadırlar. Diğer bir deyişle y_i gözlemi hem model uyumunda hem de modelin değerlendirilmesinde kullanılmazlar. Bu işlemde $\hat{y}_{i(i)}$ değeri $x=x_i$ noktasındaki kestirilmiş değerdir. Fakat y_i gözlemi bu aşamada veri kümesinin dışında bırakıldığı için parametrelerin tahminlenmesinde kullanılmamışlardır. Fonksiyonel olarak,

$$\hat{y}_{i(i)} = \mathbf{x}_i^T \mathbf{b}_{(i)} \tag{7.18}$$

tanımlanır. Burada $\mathbf{b}_{(i)}$ i-inci gözlem kullanılmadan elde edilen parametre tahmin vektörüdür. Sonuç olarak her bir aday model n adet PRESS artığına sahiptir. Modelin değerlendirilmesinde kullanılacak PRESS (kestirim kareler toplamı) kriteri,

$$\begin{aligned} PRESS &= \sum \left(y_i - \hat{y}_{i(i)} \right)^2 \\ &= \sum e_{i(i)}^2 \end{aligned} \tag{7.19}$$

olup en küçük PRESS değerine sahip modelin geçerli kılınmış model olarak seçilir. PRESS artıkları kestirim yeteneğini belirten R^2 benzeri bir istatistik oluşturmak için,

$$R_P^2 = 1 - \frac{PRESS}{\sum (y_i - \bar{y})^2} \tag{7.20}$$

kullanılabilir.

Regresyon analizinde bireysel PRESS artıklarının başka önemli bir rolü de regresyon sonuçları üzerinde önemli etkisi olan gözlemlerin ya da veri noktalarının belirlenmesine yardımcı olmasıdır. Regresyon analizinde etkili gözlem ve veri noktaları Bölüm 11 de incelenecektir.

PRESS artıkları tekrarlı regresyon uygulanmasına gerek olmadan basit bir şekilde sıradan artıklar kullanılarak,

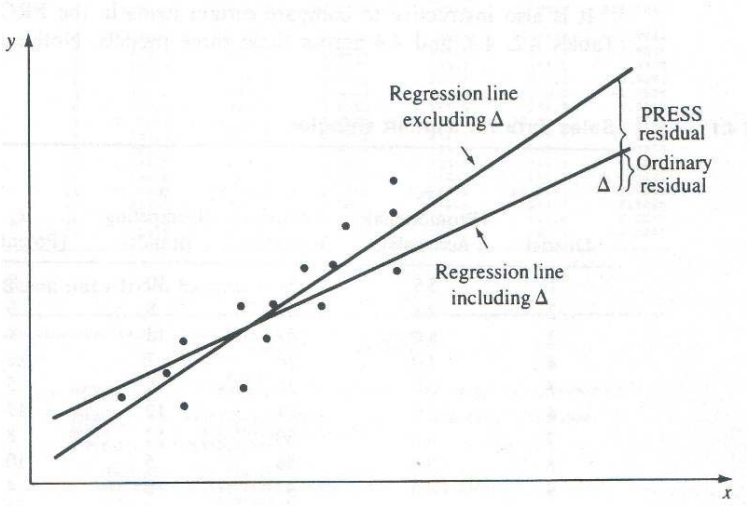
$$e_{i(i)} = \frac{y_i - \hat{y}_i}{1 - \mathbf{x}_i^T \left(\mathbf{X}^T \mathbf{X}\right)^{-1} \mathbf{x}_i} = \frac{e_i}{1 - h_{ii}} \tag{7.21}$$

elde edilebilir, bkz. Alıştırma 7.1. Sonuç olarak sıradan artıklar kullanılarak PRESS,

$$PRESS = \sum \left(\frac{e_i}{1 - h_{ii}} \right)^2 \tag{7.22}$$

formülü ile hesaplanabilir. Daha önce belirtildiği gibi eşitlik (7.21) sadece geçerli kılma amacıyla değil etkili gözlemlerin teşhis edilmesinde kullanılabilecek bir tanı istatistiğidir. Uyumu yapılmış değerler, parametre tahminleri, varyans-kovaryans matrisi ve diğer regresyon sonuçları için eşitlik (7.21) e benzer formüller Bölüm 11 de verilecektir. Eşitlik (7.21) in paydasındaki h_{ii} değeri izdüşüm matrisinin i-inci köşegen elemanı olup σ^2 hariç kestirim varyansını tanımlar. Kestirim yeteneğinin zayıf olduğu veri noktalarında h_{ii} değeri bire yakın olup güven aralığında göreceli olarak geniştir. Bu noktalar potansiyel yüksek etki noktalarıdır ve PRESS artıklarının değerleri de göreceli olarak büyüktür.

Sıradan artıklar, PRESS artıkları ve etkili veri noktası arasındaki ilişki basit regresyon kullanılarak Şekil 7.5 de Δ noktası dikkate alınarak açıklanmıştır. Eğer Δ gözlemi veri setinden çıkarılırsa b_0 ve b_1 değerleri oldukça etkilenecektir. Bu gözlemin izdüşüm matrisinin köşegendeki değeri de göreceli olarak büyüktür. Şekilden de görüldüğü gibi bu gözlemin PRESS artığı sıradan artık değerinden oldukça büyüktür.



Şekil 7.5 Myers 173

Göreli olarak büyük bir PRESS değerinin sebebi bir ya da birkaç büyük PRESS artığıdır. Onların kestirim hataları doğal olarak ihmal edilemez. Bununla birlikte araştırmacı bu gözlemlerin ağırlığını azaltmak istediğinde, alternatif olarak,

$$\sum \left| y_i - \hat{y}_{i(i)} \right| = \sum \left| e_{i(i)} \right| \tag{7.23}$$

kriterini kullanabilir. Bu ifade karesel olmadığı için büyük kestirim hatalarına sahip gözlemlerin PRESS içindeki ağırlığını azaltacaktır.