T.C.
İSTANBUL ÜNİVERSİTESİ
SOSYAL BİLİMLER ENSTİTÜSÜ
EKONOMETRİ ANABİLİM DALI

DOĞRUSAL REGRESYON MODELLERİNDE BAĞIMSIZ DEĞİŞKEN SEÇİM YÖNTEM VE ÖLÇÜTLERİ

(Yüksek Lisans Tezi)

Hazırlayan
Nuran Eyyuboğlu

Danışman
Prof. Dr. Mehmet GENCELİ

İSTANBUL 1990

GİRİŞ

Bağımsız değişkenlerin seçimi regresyon analizinin kullanıldığı bütün alanlarda ve ekonometride önemli bir problemdir. Bu çalışmada bağımsız değişken seçiminde yaygın olarak kullanılan yöntem ve ölçütler incelenecektir.

Birinci bölümde genel olarak model kavramı üzerinde durulmuştur.

İkinci bölümde doğrusal regresyon modelinde bağımsız değişken seçim yöntemleri ve ölçütleri ele alınmıştır. Seçim yöntemleri olası bütün regresyon çözümleri ve sırasal seçimler olarak iki ana başlıkta toplanarak , çözüm süreçleri verilmiştir. Olası bütün regresyon çözümleri değerlendirilmesinde kullanılan karar ölçütlerinden R^2_p, MSE_p, S_p, C_p ayrı ayrı açıklanarak bir örnek üzerinde uygulanmıştır. Yine bu bölümde sırasal seçim yöntemleri de incelenerek aynı örneğe uygulanmıştır.

Üçüncü bölümde karar ölçütleri arasındaki karşılıklı ilişkiler araştırılmıştır. Seçim yöntemleri karşılaştırılarak bu yöntemlerin birbirlerine göre zayıf ve üstün yönleri belirtilmiştir.

MODEL KAVRAMI

Varolan bir yapıyı tanımak, onu analiz etmek ve gelecekteki şeklini kestirebilmek amacı ve gereği (ve merakı), bu yapı üzerinde gözlem ve deney yapma şansının olmadığı durumlarda yapının küçük bir benzerini; modelini kurma gereğini ortaya getirmiştir.

Model, karmaşık bir yapıda olan gerçek yaşamın, kantitatif yöntemlerle basitleştirilmiş ve daha kolay analiz edilebilir duruma getirilmiş bir anlatımı olarak tanımlanabilir.

Model kurarken ilk aşama incelenecek konunun belirlenmesidir. Konu saptandıktan sonra yapılacak iş, varolan yapıda birbirini etkileyen büyüklükleri ortaya koymaktır. Gerçek yaşamın karmaşık ve ayrıntılı yapısı göz önüne alındığında bu büyüklüklerin; değişkenlerin sayısının hiç de az olmayacağı ortaya çıkar. Öte yandan, bir modeli gerekenden fazla sayıda değişken ile kurmak, modelin kolay kavranılabilir ve kullanılabilir olma özelliğini yitirmesine neden olacaktır. Bunun yanı sıra gerekenden az sayıda değişken ile kurulmuş bir modelin gerçek yapı hakkında vereceği bilgi de yetersiz, ya da sınırlı kalacaktır.

Bu gerçekler göz önüne alındığında kurulacak her modelde şu iki amaç birbiri ile çelişecektir:

1- Basit olmak

2- Bol ayrıntı vermek

Bu durumun çözümü, basit, ancak varolan yapıyı olabildiğince ayrıntıları ile yansıtan "en uygun" sayıda ve kompozisyonda bağımsız değişken içeren "en iyi" modeli aramaktır. En uygun sayı ve birleşimde değişken içeren model, tek ve değişmez değildir. Diğer bir ifade ile, varolan yapıyı anlatacak birden çok model kurulabilir. O halde, kurulacak model saptanırken belirleyici olan nokta, model kurmadaki amaç olacaktır.

2. DOĞRUSAL REGRESYON MODELİNDE BAĞIMSIZ DEĞİŞKEN SEÇİM YÖNTEMLERİ ve KARAR ÖLÇÜTLERİ

Bu bölümde, değişken seçim süreçlerinden hiçbirinin değişmez en iyi bağımsız değişken kümesini vermediği ve aslında bu kümenin tek olmadığı bilgisi altında, seçim yöntem ve ölçütlerinden yaygın olarak kullanılanlar incelenecektir. Seçim yöntemleri,

1) Olası bütün regresyon çözümleri,
2) Sırasal seçimler

olmak üzere 2 ana başlıkta toplanır. Olası bütün regresyon modelleri çözüm sonuçları değerlendirilirken "Karar Ölçütleri" kullanılır. Bu ölçütler, sırasal seçim yöntemlerinde, değişken alımında "durulması" gereken aşamayı da belirleyebilmektedir. Bu çalışmada R^2_p, MSE_p, S_p ve C_p ölçütleri incelenecektir.

n-gözlem sayısını,

k-parametre sayısını

göstermek üzere genel regresyon modelinin matris terimlerle anlatımı;

$$Y_{nx1}=X_{nxk}\beta_{kx1}+u \qquad (2.1)$$

şeklindedir.

p-modele giren değişken sayısını,

r-model dışında kalan bağımsız değişken sayısını gösterir ise genel regresyon modeli;

$$Y=X_p\beta_p+X_r\beta_r+u \qquad (2.2)$$

şeklinde düzenlenebilir.

X_p, X bağımsız değişken matrisi içinden seçilmiş, p sayıda,(p<=k),değişken içerir (nxp) boyutunda bir matristir. Bu alt bağımsız değişken kümesine ait parametre tahmin vektörü ;

$$b_p=(X'_pX_p)X_pY \qquad (2.3)$$

şeklindedir.

2.1. OLASI BÜTÜN REGRESYON ÇÖZÜMLERİ

Bu yöntem, bağımsız değişken listesine alınmış tüm X'ler için olası bütün modellerin çözümünü içerir (2^k-1 sayıda). Olası bütün regresyon çözümlerini yaparak en iyi X kümesini aramak bilgisayar kullanımını ve artan değişken sayısına bağlı olarak da fazla zaman ayrılmasını gerektirmektedir.

Çözüm Aşamaları:

1- p bağımsız değişken sayısını göstermek üzere, p=1, p=2, ..., p=k tane X içeren olası bütün modeller çözülür.

2- Eşit sayıda bağımsız değişken içeren modeller kendi aralarında bir karar ölçütüne göre sıralanır. Söz konusu karar ölçütleri daha sonra irdelenecektir.

3- Bu ölçüt gereğine göre her alt kümeden en iyi sonucu veren model ya da modeller seçilir.

4- Yine aynı karar ölçütüne göre sıralanan modellerden bu ölçüte göre, en iyi küme final modeli olarak seçilir.

2.2. DOĞRUSAL REGRESYON MODELİNDE BAĞIMSIZ DEĞİŞKEN SEÇİM ÖLÇÜTLERİ

Bu bölümde bağımsız değişken seçim ölçütlerinden (karar ölçütleri), yaygın olarak kullanılan R^2_p ve $\bar{R}^2_p$, MSE_p, S_p ve C_p incelenecektir.

2.2.1. R^2_p ÖLÇÜTÜ

En yaygın kullanılan ölçüt R^2_p ölçütüdür.Ölçüt öz olarak, Y deki değişkenliğin X ler tarafından açıklanabilirliği olarak tanımlanan çoklu belirginlik katsayısının değerlendirilmesini içerir. (1)

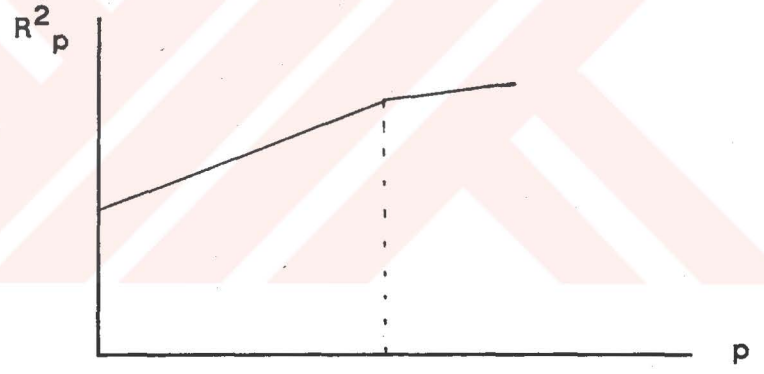
$$R^2_p = \frac{SSR_p}{SST} = 1 - \frac{SSEp}{SST} \qquad (2.4)$$

---

(1) J.Neter & W.Wasserman, Applied Linear Statistical Models, IRWIN, 1974, s.376

Bağımsız değişken listesinde bulunan tüm X ler modele alındığında R^2_p nin en büyük değer alacağı açıktır. Regresyon modeline her alınan bağımsız değişken modelin Y yi açıklayabilirliğini pozitif olarak etkileyeceğine ya da varolan durumu değiştirmeyeceğine göre, seçim ölçütü olarak R^2_p yi kullanmaktaki amaç bu ölçütü en büyük yapan X kümesini bulmak değildir.

R^2_p nin p ye; değişken sayısına göre grafiği çizildiğinde ortaya başlangıçta dik olarak artan ve bir noktadan sonra (knee) eğimi azalan bir eğri çıkmaktadır (2). Bu nokta genellikle bu ölçüte göre seçilecek en iyi kümeye karşılık gelen değişken sayısını vermektedir.



Şekil(1)

Bu noktada, modele alınacak bir fazla değişkenin modele katkısının önemli olmayacağına karar verilir.

Bazı yazarlar R^2_p ye alternatif olarak düzeltilmiş R^2_p nin kullanılmasını önermektedirler.

$$\bar{R}^2_p = 1 - \frac{n(1-R^2_p)}{n-p} \qquad (2.5)$$

---

(2) R.R. Hocking, The Analysis & Selection of Variables In Linear Regression , Biometrics, 32, 1976, s.15.

Bu ölçüte göre en iyi model R^2_p si en yüksek olan modeldir.

2.2.2. MSE_p ÖLÇÜTÜ

R^2_p modeldeki değişken sayısını dikkate almadığı ve p artarken R^2_p asla azalmayacağına göre kareli ortalama hata, MSE_p bir ölçüt olarak önerilmiştir (3).

$$MSE_p = \frac{SSE_p}{n-p} \qquad (2.6)$$

Bu ölçüte göre MSE_p değeri en az olan bağımsız değişken kümesi en iyi model olarak seçilir.

2.2.3. S_p ÖLÇÜTÜ

Bu ölçüt de gerçek ve tahmin edilen Y değerleri arasındaki farkı en az yapan X kümesini bulmayı amaçlar.

$$S_p = \frac{SSE_p}{(n-p)(n-p-2)} = \frac{MSE_p}{n-p-2} \qquad (2.7)$$

En küçük Sp değerine karşılık gelen X kümesi en iyi model olarak seçilir.

---

(3) J.Neter & W.Wasserman, a.g.e., s.379.

2.2.4. C_p ÖLÇÜTÜ

Bu ölçüt ilk kez 1964 yılında Mallow tarafından önerilmiştir ve Mallow ölçütü (MC) olarak da anılmaktadır.(4)

Ölçüt öz olarak standartlaştırılmış ortalama kareli hatanın değerlendirilmesini içerir.

$\hat{Y}p = Xpbp$ iken ölçüt,

$$\Gamma_p = \frac{1}{\sigma^2} [E(\hat{Y}_p) - X\beta]' [E(\hat{Y}_p) - X\beta] + p \quad (2.8)$$

şeklindedir.

$[E(\hat{Y}p) - X\beta]' [E(\hat{Y}p) - X\beta]$ model k sayıda bağımsız değişken ile çözülecek iken p sayıda bağımsız değişken ile sınırlandırılmış olmasından kaynaklanan yanlı terimdir (5). SSB ile gösterilirse,

$$\Gamma_p = SSB/\sigma^2 + p \quad (2.9)$$

şeklinde olacaktır.

$$E(SSE_p) = (n-p)\sigma^2 + SSB \quad (2.10)$$

olduğundan;

$$\Gamma_p = E(SSEp) / \sigma^2 - n+2p \quad (2.11)$$

---

(4) Takeshi Amemiya, Selection of Regressors, International Economic Review, Vol.24, No2, 1980, s.163.
(5) R.R. Hocking, a.g.e., s.18

şeklinde yazılabilir. Γ_p nin kestirimi C_p ile gösterilir ve

$$C_p = SSE_p/\hat{\sigma}^2 + 2p-n \quad (2.12)$$

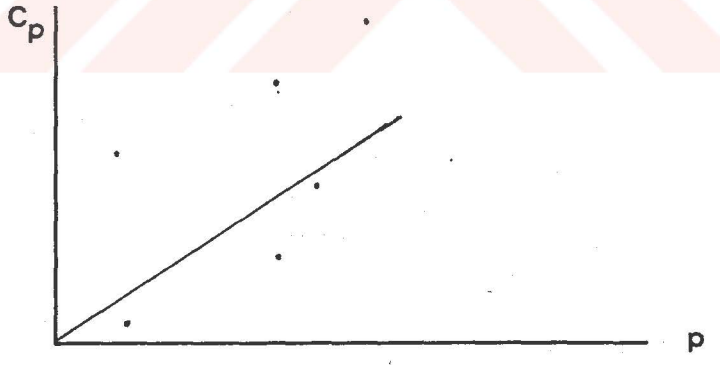
şeklindedir.

Eğer seçilmiş olan p adet bağımsız değişken en iyi bağımsız değişken kümesi ise, başka bir deyişle Y yi en iyi açıklayan X ler kümesi ise yanlı terim 0 a yaklaşacak ve

$$C_p = (n-p)\hat{\sigma}^2/\hat{\sigma}^2 + 2p-n = p \quad (2.13)$$

olacaktır. Bu özellik en iyi X kümesini belirlemekte C_p nin diğer ölçütlere göre bir üstünlüğüdür.

C_p ve p değerleri bir grafik üzerinde gösterildiğinde noktaların az yanlılıkla $C_p = p$ doğrusu çevresine düştüğü görülecektir.



Şekil - 2

2.2.5. OLASI BÜTÜN REGRESYON ÇÖZÜMLERİ İÇİN UYGULAMA

İncelenen yöntem ve ölçütler aşağıdaki veriler üzerinde uygulanmıştır.

Y	X_1	X_2	X_3	X_4
39	3	8	21	12
46	2	9	18	18
48	3	11	17	19
51	4	10	22	21
56	5	12	24	15
54	4	13	26	18
52	3	11	29	20
58	8	7	21	23
62	11	14	17	16

1.) Olası Bütün Regresyon Çözümleri İçin Model Seçimi

Bütün olası regresyon çözümleri için R^2_p, R^2_p, S_p ve C_p değerleri aşağıdadır.

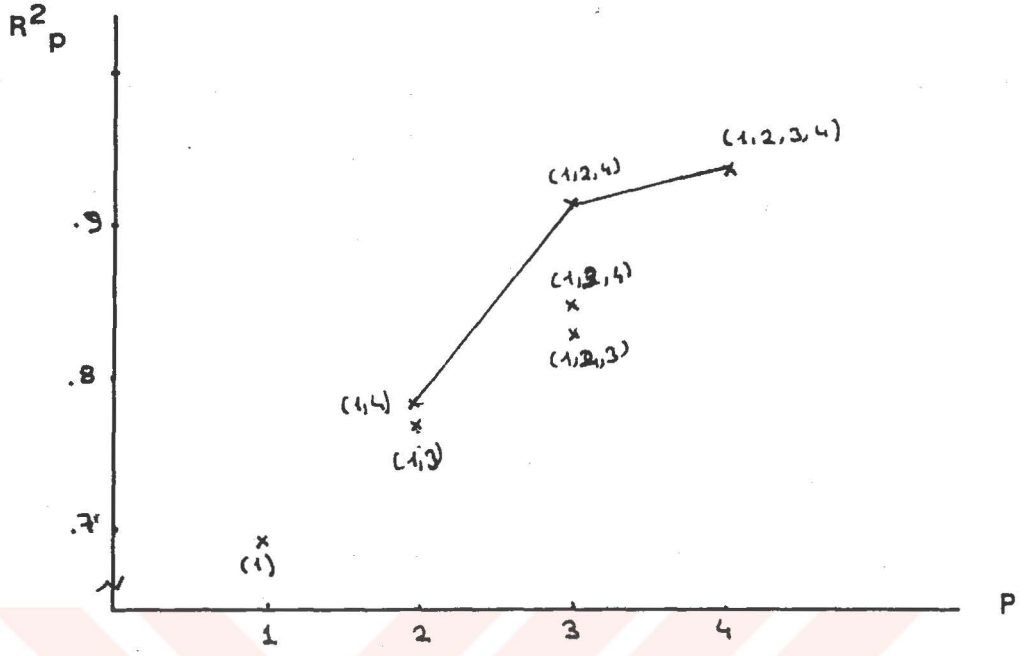
X	P	R^2_p	R^2_p	MSE_p	S_p	C_p
X_1	1	0.69	0.65	16.8	2.1	14.0
X_2	1	0.29	0.21	38.5	5.1	42.1
X_3	1	0.02	0.12	54	7.7	63.2
X_4	1	0.12	0.01	14.9	6.9	5.6
X_1, X_2	2	0.76	0.69	14.9	2.5	11.0
X_1, X_3	2	0.77	0.71	14.1	2.4	10.1
X_1, X_4	2	0.78	0.71	13.7	2.3	9.6
X_2, X_3	2	0.29	0.09	44.0	7.3	44.1
X_2, X_4	2	0.53	0.40	29.2	48.7	27.3
X_3, X_4	2	0.12	0.12	54.8	9.1	56.6
X_1, X_2, X_3	3	0.82	0.72	13.4	2.7	9.09
X_1, X_2, X_4	3	0.91	0.86	6.8	1.4	2.6
X_2, X_3, X_4	3	0.54	0.31	33.5	6.7	28.7
X_3, X_4, X_1	3	0.84	0.76	11.7	2.3	7.4
X_1, X_2, X_3, X_4	4	0.93	0.87	6.14	1.5	3.0

a) R^2_p Ölçütüne Göre Seçim

Eşit sayıda X içeren kümelerin arasında R^2_p ye göre seçilmiş modeller aşağıdadır.

P	R^2_p	Modeldeki Bağımsız Değişken
4	0.93	X_1, X_2, X_3, X_4
3	0.91	X_1, X_2, X_4
3	0.84	X_1, X_3, X_4
3	0.82	X_1, X_2, X_3
2	0.78	X_1, X_4
2	0.77	X_1, X_3
1	0.69	X_1

R^2_p nin p ye göre grafiği aşağıdadır.



Şekil - 3

Grafikten de görüldüğü gibi "diz" (knee) $R^2_{p=}$ 0.91 ve p=3 olduğu noktadadır. Bu ölçüte göre

$$Y= 15.6 +1.4X_1 + 1.25X_2 + 0.8X_4$$

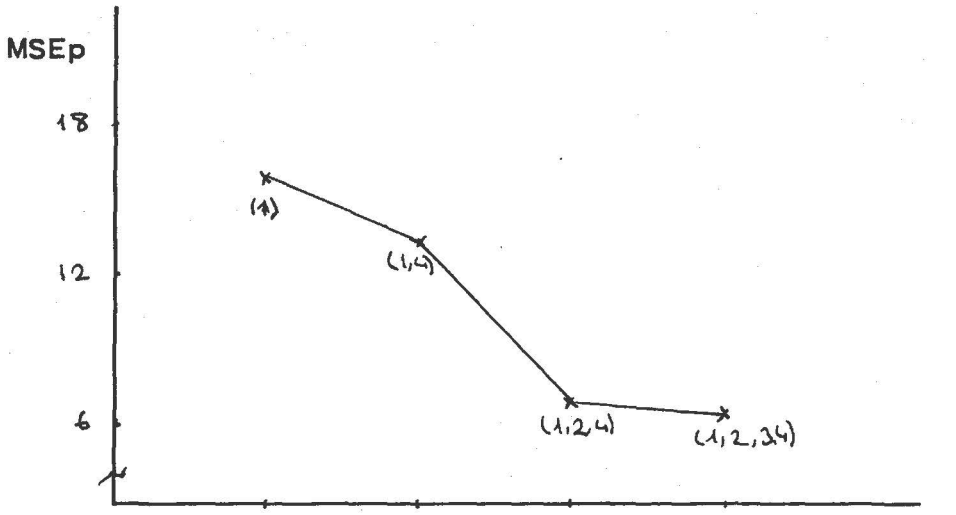
en iyi model olarak seçilir.

Düzeltilmiş R^2_p si en büyük olan model de aynı modeldir.

b) MSE_p Ölçütüne Göre Seçim

P	R^2_p	Modeldeki Bağımsız Değişken
4	6.14	X_1, X_2, X_3, X_4
3	6.7	X_1, X_2, X_4
2	13.7	X_1, X_4
1	16.8	X_1

MSE_p nin p ye göre grafiği aşağıdadır.



Şekil - 4

En uygun model MSE_p nin en küçük olduğu $Y= f(X_1,X_2,X_4)$ modelidir. X_3 ün modele katkısının önemli olmadığına karar verilebilir;

$$Y=15.6 +1.4X_1 + 1.25X_2 + 0.8X_4$$

c) S_p Ölçütüne Göre Seçim

S_p ye göre seçilmiş modeller aşağıdadır

P	S_p	Modeldeki Bağımsız Değişken
4	1.5	X_1,X_2,X_3,X_4
3	1.4	X_1,X_2,X_4
3	2.3	X_1,X_2,X_3
2	2.3	X_1,X_3
1	2.3	X_1

Bu ölçüte göre en iyi model yine,

$$Y=15.6 +1.4X_1 + 1.25X_2 + 0.8X_4$$

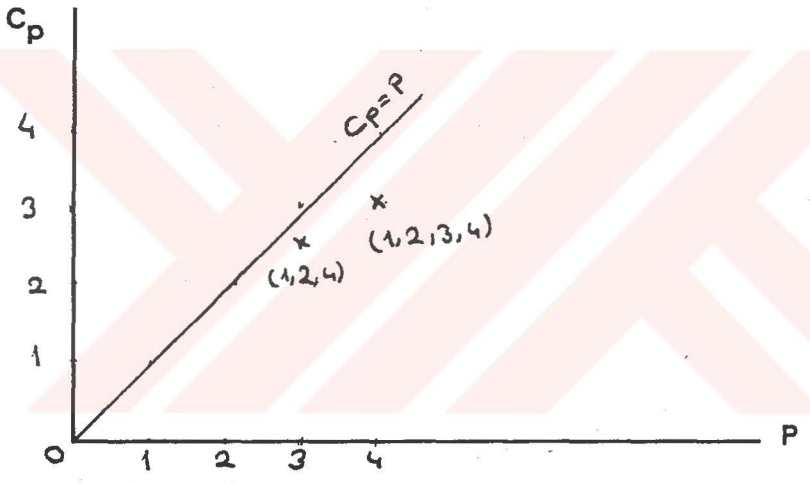
olmaktadır.

d) C_p Ölçütüne Göre Çözüm

C_p ölçütüne göre seçilmiş modeller aşağıdadır.

P	C_p	Modeldeki Bağımsız Değişken
4	3	X_1, X_2, X_3, X_4
3	2.6	X_1, X_2, X_4
2	9.6	X_1, X_4
1	14	X_1

C_p ölçütüne göre de en iyi model X_1, X_2, X_4 değişkenlerini içeren küme olarak görünmektedir. C_p nin p ye göre grafiği ve değerlerin C_p=p doğrusu etrafındaki dağılımları aşağıdadır.



Şekil - 5

p=3 iken C_p=2.6 değerini almıştır. k sayıda bağimsız değişken var iken yalnız p sayıda X ile model kurulmasından ortaya çıkan hatayı en küçük yapan küme (X_1, X_2, X_4) değişkenlerini içermektedir.

2.3. SIRASAL SEÇİM YÖNTEMLERİ

Olası bütün regresyon çözümlerini yapmanın ortaya çıkardığı zorluklardan ötürü sırasal seçim yöntemleri önerilmiştir.

Bu yöntemler öz olarak bir ölçüte ya da teste göre anlamlı olan değişkenlerin modele eklenmesini ya da anlamsız olan değişkenin model dışı bırakılmasını içerir.

Sırasal seçim yöntemleri,

1)İleri doğru sıralama,

2)Geriye doğru sıralama,

3)Geriye doğru eleme,

4)İleri doğru seçim ve

5)Adım adım regresyon (Stepwise) yöntemi olarak sıralanır.

Bu seçim yöntemleri aşağıdaki teorem (6) ile ilintilidir.

Teorem 1:Hata paylarının normal dağıldığı ,$E(u)=0$ ve $\sigma^2_u=\sigma^2$ olduğu varsayımları altında,

$$Y=\Sigma\beta_i X_i+u$$

genel doğrusal modeli verilmiş olsun.

$\beta_1, \beta_2, \ldots\ldots\ldots, \beta_{(k-p)}$, (k-p) sayıda X içeren kümeye ait parametreler iken,

$$H_0: \beta_1=\beta_2=\beta_3\ldots\ldots\ldots=\beta_{(k-p)} = 0$$

için varyans oranı;

$$F_s = \frac{SS_{(k-p)}}{k-p} \Big/ \frac{SST-SSR_k}{n-k} = \frac{SS_{(k-p)}/k-p}{MSE_k} \qquad (2.13)$$

(6) Mary L.Thompson, Selection of Variables In Multiple Regresion, International Statistical Review,46 1978, s.8.

şeklindedir.

Formüldeki terimler:

$SSR_k = b'X'Y$,

$SSR_p = b_p'X_p'Y$,

$SS_{k-p} = SSR_k - SSR_p$,

$SST = Y'Y$,

$MSE_k = SSE_k / n-k$,

$SSE_k = SST - SSR_k$,

şeklindedir.

2.3.1.İLERİ DOĞRU SIRALAMA

Bu yöntem bağımsız değişkenleri azalan önem sırası içinde sıralamayı içerir.(7)

Çözüm Aşamaları:

1)F_s oranı en küçük olan değişken en önemli bağımsız değişkendir. p=1 için ;

$$F_s = \frac{SS_{k-1}/(k-1)}{MSE_k}$$

şeklindedir.

Açıktır ki F_s oranı en düşük olan değişken, SS_{k-1} değeri en küçük ya da SSR_1 değeri en büyük olan değişkendir ($SS_{k-1} = SSR_k - SSR_1$).

2) Modele alınan her değişkenden sonra model dışında kalan, (k-p) sayıda parametre için H_0 sınanır. $F_s > F_c$ ise model dışında kalan parametrelerden en az birinin 0' dan farklı olduğuna karar verilerek seçime devam edilir.

---

(7) Mary L.Thompson, a.g.e., s.9.

3) Modele 2. sırada alınacak değişken ise daha önce seçilmiş 1. değişkenle birlikte en büyük SSR_2 değerini veren değişkendir.

4) 2. aşamada olduğu gibi model dışındaki parametrelerin anlamlılığı sınanır. Bu seçim yöntemine modele alınmayan (k-p) sayıda bağımsız değişkene ait parametrelerin önemsiz olduğuna karar verilene, yani $F_s<F_c$ olana dek devam edilir. $F_s<F_c$ olduğunda bu basamağa kadar alınmış p sayıda bağımsız değişken final modeli olarak seçilir.

2.3.2.GERİYE DOĞRU SIRALAMA

Bu yöntemde amaç, ileri doğru sıralama yönteminin tersine, modele girecek değişkenleri seçmek değil, model dışında kalacak değişkenleri saptamaktır.

Çözüm Aşamaları:

1) Birinci basamakta en küçük SS_1 değerini veren bağımsız değişken en önemsiz değişken olarak seçilir.($SS_1=SS_{k-(k-1)}=SSR_k-SSR_{k-1}$).

2) Bu aşamada 1. basamakta en önemsiz değişken olarak seçilen X için F_s değeri hesaplanır.

$SS_1=SS_k-SS_{k-1}$ iken

$$F_s = \frac{SS_1}{MSE_k} \qquad (2.15)$$

şeklindedir.$F_s<F_c$ ise bu değişkene ait parametre için H_0 kabul edilir; değişken modele alınmaz.

3) Bu aşamada ise daha önceki basamakta modelden çıkarılmış değişken ile birlikte en küçük SS_2 değerini veren değişken 2. önemsiz X olarak seçilir.

4) Bu basamakta seçilmiş X'ler için F_s değeri hesaplanır.

$$SS_{k-2}=SS_k-SS_{k-2} \text{ iken:}$$

$$F_s= \frac{SS_{k-2}/2}{MSE_k} \qquad (2.16)$$

$F_s<F_c$ ise bu iki değişken içinde H_0 kabul edilmektedir; değişkenler modele alınmaz.

Bu yöntem $F_s>F_c$ değerini aldığı basamakta durdurulur. Bu basamakta önemsiz değişken olarak seçilmiş değişkene ait parametreler için H_0 red edilir. Dolayısı ile final modeli bu değişkeni ve arta kalan X'leri içerir

2.3.3.GERİYE DOĞRU ELEME YÖNTEMİ

Bu yöntem olası bütün regresyon çözümlerinin gelişmiş bir şeklidir.Diğer sırasal seçim yöntemlerinden temel farkı, çözüm aşamasına k sayıda bağımsız değişkeni içeren modelle başlamasıdır .(8)

Çözüm Aşamaları:

1) k sayıda bağımsız değişken içeren tam model, (full model) çözülür.

2) Her X için kısmi F'ler:

$$F_s = SS_1/MSE_k \qquad (2.17)$$

değerleri hesaplanır.

3) F_s değeri en küçük bağımsız değişken, en önemsiz olarak seçilir.

4) Bu F_s değeri kritik F değeri ile karşılaştırılır. $F_s<F_c$ ise H_0 kabul edilerek değişken modelden çıkarılır.

---

(8) N.Draper & H.Smith, Applied Regression Analysis , Wiley, 1966, s.162.

5) $F_s > F_c$ olan bağımsız değişken bulunana kadar modelden değişken çıkarmaya devam edilir. q. basamak için F_s değeri şöyledir:(9)

$$F_s=(n-k-q-1)SS_1/SST-SSR_{k-q-1} \quad (2.18)$$

2.3.4.İLERİ DOĞRU SEÇİM YÖNTEMİ

İleri doğru seçim yöntemi, ileri doğru sıralamada oldugu gibi, karşılık gelen parametre için H_0 hipotezinin kabul edileceği X değişkenini bulana dek modele değişken alınmasını içerir.

Çözüm Aşamaları:

1) İleri doğru seçim yöntemi korelasyon matrisinin hesaplanması ile başlar.

2) Y ile korelasyonu en yüksek olan X modele alınacak ilk değişken olarak seçilir.

3) Bu değişken için kısmi F_s değeri hesaplanarak, karşılık gelen parametre için H_0 varsayımı sınanır.

4) Bu aşamada modele ilk alınan değişkenle birlikte, Y ile korelasyonu en büyük olan X modele alınacak 2. değişken olarak seçilir.

5) Bu değişkene karşılık gelen parametrenin anlamlılığı sınanır. $F_s > F_c$ ise modele alınacak 3. değişken için kısmi korelasyon katsayıları araştırılır.

q. basamak için F_s oranı aşağıdaki gibi yazılır: (10)

---

(9) Mary L.Thompson, a.g.e., s.10.

(10) Mary L.Thompson, a.g.e., s.11.

$$F_s = \frac{SS_1}{(SST-SSR_q)/(n-q)} \quad (2.19)$$

$F_s<F_c$ olduğu basamakta modele son giren ve model dışında kalan değişkenlerin etkisiz olduğuna karar verilerek bir önceki basamaktaki model, final modeli olarak seçilir.

2.3.5.ADIM ADIM REGRESYON (STEPWISE REGRESYON)

Stepwise regresyon ileri doğru seçim yönteminin daha gelişmiş şeklidir. Bu gelişme, modele alınacak her yeni değişkeni sınamanın yanı sıra, modele daha önce alınmış değişkenlerin sanki o basamakta modele alınıyorlarmış gibi,yeniden test edilmesindedir. Diğer tüm yöntemlerden farklı olarak stepwise regresyonda modele alınmış olan herhangi bir değişken, ileri basamakların birinde model dışı kalabilir; başka bir deyişle modelden "düşer". Böylece bağımsız değişkenlerin her birinin modele etkilerinin yanı sıra diğer değişkenlerle etkileşimleri de sınanır.

Çözüm Aşamaları:

1) Y ile korelasyonu en büyük olan X modele ilk alınacak olan bağımsız değişkendir.

2) Bu X değişkenine ait kismi F_s değeri hesaplanır.

3) $F_s > F_c$ ise, bu değişkenle birlikte en büyük kısmi korelasyon katsayısına sahip değişken modele alınarak 2. aşamada olduğu gibi F_s hesaplanır.

4) $F_s>F_c$ ise değişken modele alınır.Bu adıma kadar çözüm aşamaları, ileri doğru regresyon ile aynıdır.

5) Bu aşamada ise modele 1. basmakta alınmış X için, modele yeni giriyormuş gibi H_0 yeniden sınanır.

6) Bu X için hala $F_s > F_c$ ise değişken modelde kalır. Tersi durumda modelden çıkarılır; değişken modelden düşer.

7) Süreç geride kalan değişkenlerin önemsiz olduğuna karar verilen aşamaya, Fs < Fc olana kadar devam eder.Bu aşamadaki model final modeli olarak seçilir.

Bu yöntem, modele alınacak değişkeni ve daha önceki basamaklarda alınmış değişkenleri de sınaması ve hesaplama kolaylığı bakımından pek çok uygulayıcı ve yazar tarafından var olan yöntemlerin en iyisi olarak kabul edilmektedir. SPSS, MINITAB ve MICROSTA paket programları stepwise ile çözüm yapmaktadır.

2.3.6.SIRASAL ÇÖZÜM YÖNTEMLERİ İÇİN UYGULAMA

1)İleri Doğru Sıralama

İleri doğru sıralama yönteminin aşamaları aşağıdadır:

1) $SSR(X_1)=303.5$, $SSR(X_2)=130.3$, $SSR(X_3)=0.95$, $SSR(X_4)=54.1$

X_1 en önemli değişken olarak secilir.

2) H_0: $\beta_2 = \beta_3 = \beta_4 = 0$

H_1: β_2 , β_3 , β_4' ten en az biri 0'dan farkli.

$$Fs = \frac{SS_3/3}{MSE_4} = \frac{353.6/3}{6.14} = 19.1$$

$F_{c,0,05:3,6}$:4.76 < F_s H_0 red.

3) $SSR(X_3X_1)=339.5$, $SSR(X_3X_2)=130.3$, $SSR(X_3X_4)=54.1$

X_1, 2. önemli değişken olarak modele alınır.

4) H_0: $\beta_2 = \beta_4 = 0$

H_1: β_2 ve β_4 'ten en az biri 0' dan farklidir.

$F_s = SS_2/2 = 277.4/2 = 22.6$

MSE_4 6.14

$F_{c,0.05;2,6} = 5.14 < F_s$ H_0 red.

5) $SSR(X_3X_1X_2)=357.9$, $SSR(X_1X_3X_4)=368.9$

X_4 3. önemli değişken olarak seçilir.

6) H_0: $\beta_2 = 0$

H_1: β_2 0'dan farklı.

$$F_s = \frac{SS_1/1}{MSE_4} = \frac{39.5}{6.14} = 6.4$$

$F_{c,0.05;1,6}=5.99 < F_s$ H_0 red; X_2 de modele alınabilir.

Tam model geçerli kabul edilir.

2)Geriye Doğru Sıralama

Bu yöntem için çözüm aşağıdadır.

1) $SS_1(X_1)=170.4$, $SS_1(X_2)=39.5$, $SS_1(X_3)=10.1$,$SS_1(X_4)=49.8$

X_3 en önemsiz değişken olarak seçilir.

2) Ho : $\beta_3 = 0$,

H1 : β_3 0 dan farklı.

$F_s = SS_1$ 10.1 = 1.6

MSE_k 6.14

$F_{c,0.05;1,6}$: 5.99 > F_s H_0 kabul edilir.β_3 modele alınmaz.

3) $SS_2(X_3X_1)=173.8$, $SS_2(X_3X_2)=73.7$, $SS_2(X_3X_4)=68.2$

X_4 2. en önemsiz değişken olarak seçilir.

4) $F_s = \frac{SS_2/2}{MSE_4} = \frac{68.2/2}{6.14} = 5.55$

$H_0 : \beta_4 = \beta_3 = 0$,

$H_1 : \beta_4$ ve β_3 ten en az biri 0 dan farklı.

$F_{c,0.05;2,6}$:5.14 < Fs H_0 red.

X_4 ve diğer değişkenlerin önemli olduğuna karar verilir.

$Y = \beta_0 + \beta_4X_4 + \beta_2X_2 + \beta_3X_3$ modeli geçerlidir.

c) Geriye Doğru Eleme

Bu yöntem $Y=11.5\ 1.6X_1+1.1X_2+\ 0.29X_3+\ 0.8X_4$ modeli ile başlar.

1) $SS_1(X_1)=170.4, SS_1(X_2)=39.5, SS_1(X_3)=10.1, SS_4(X_4)=49.4$

X_3 modelden çıkarılacak değişken olarak seçilir (geriye doğru sıralamada olduğu gibi).

F_s = 1.6 hesaplanmıştı.

$F_s > F_c$ dolayısı ile $\beta_3 = 0$.

2) Bu basamakta $Y=15.6+1.4X_1+1.2X_2+0.8X_4$ modeli geçerlidir.

$SS(X_1)=94.6, SS(X_2)=267.3, SS(X_4)343.5.$

X_1 modelden çıkarılabilecek değişken olarak seçilir.

$F_s = \frac{SS_1}{MSE_k} = \frac{94.6}{6.14} = 15.4$

$H_0 : \beta_1=0,$

$H_1 : \beta_1$ 0 dan farklı.

$F_{c,0.05;1.7}$:5.59 > F_s H_0 red edilir.X_1 ve kalan değişkenler modele alınır," tam" model geçerlidir.

d) İleri Doğru Seçim

1) Korelasyon matrisi aşağıdadır:

	Y	X_1	X_2	X_3	X_4
Y	1	.8	.5	.0	.3
X_1		1	.4	.3	.0
X_2			1	.1	.2
X_3				1	.1
X_4					1

2)Y ile korelasyonu en yüksek X_1 dir.

H_0 : $\beta_1=0$,

H_1 : β_1 0 dan farklı.

$F_s = SS_1 / MSE_1 = 18.0$

$F_{c,0.05;1.7} = 5.99 < F_s$ H_0 red.

3)$r_{yx1.x2} = 0.47$, $r_{yx1.x3} = 0.52$, $r_{yx1.x4} = 0.55$

X_4 modele alınabilir.

H_0 : $\beta_4=0$,

H_1 : β_4 0 dan farklı.

$F_s = SS_1/MSE_2 = 2.8$

$F_{c,0.05;1.6} = 5.99 > F_s$ H_0 kabul edilir.

$Y=\beta_0+\beta_1X_1$ modeli geçerlidir.

Stepwise regresyon içinde aynı model geçerli olacaktır.

3.SEÇİM YÖNTEMLERİNİN KARŞILASTIRILMASI ve ÖLÇÜTLER ARASI İLİŞKİLER

Bu çalışmada bağımsız değişken seçim yöntemlerinden olası bütün regresyon çözümleri ve sırasal seçim yöntemleri incelendi.

Olası bütün regresyon çözümlerini yaparak en iyi modeli aramanın en büyük güçlüğü hiç kuşkusuz 2^k-1 sayıda eşitliği çözmektir. X sayısının az olduğu uygulamalarda bu yöntem ile seçime gidilebilir. Bu durum uygulayıcının olası bütün modelleri görerek seçime gitmesini sağlamaktadır.

Bu yöntemde, final modelini bulmak için bir karar ölçütü kullanılır. Model kurmaktaki amaca en uygun ölçüt seçimi ise subjektiftir. Burada karar ölçütleri arasındaki ilişkiler açıklanacaktır.

Cp ve R^2 ler arasındaki ilişki aşağıdaki gibi yazılabilir.(1)

$$C_p = \frac{(n-k)\ (1-R^2_p)}{1-R^2_k} + 2p - n \qquad (3.1)$$

ve

$$C_p = \frac{(n-p)\ (1-R^2_p)}{1-R^2_k} + 2p - n \qquad (3.2)$$

En küçük C_p ve en büyük R^2p ya da R^2p değerini veren model,uygulamada da olduğu gibi,aynı olabilir. Ancak (3.1) den görüldüğü gibi, R^2_p 'deki küçük artışlar, (n-k) çarpanından ötürü, C_p de büyük azalışlara yol açacaktır(2). Bu konuda yapılan çalışmalarda R^2_p'nin bazen önemli sayılacak değişkenleri model dışında bıraktığı görülmüştür.

---

(1) Mary L.Thompson, a.g.e., s.17.
(2) R.R. Hocking, a.g.e., s.21.

Ayrıca grafik ile analiz yapıldığında R^2_p'nin p' ye göre "diz" oluşturduğu noktayı görmek her zaman kolay değildir. Grafik üzerinde, C_p=p doğrusunun yardımı ile en iyi kümeyi seçmek ise daha kolaydır.

C_p ve MSE_p arasındaki ilişki aşağidaki gibi yazılabilir:(3)

$$C_p=(n-p)(MSE_p/\sigma^2-1) + p \quad (3.3)$$

MSEp ölçütune göre en iyi küme ;

i. MSEp'nin en küçük olduğu X kümesi,

ii.MSEp=MSE_k olan X kümesi ya da,

iii.Grafik üzerinde MSEp 'nin artmaya başladığı noktaya karşılik gelen X kümesidir.

(ii) ve (iii) ,den görülmektedir ki en iyi model MSEp/MSEk oranını 1' e yaklaştıran X kümesidir. Bu (3.3) te yerine konulursa, bu kümenin C_p=p koşulunu da sağladığı görülür.

$$RED_p=X_\beta-X_pb_p \quad (3.4)$$

ve

$$SSE_p=RED_p+SSE_k \quad (3.5)$$

iken C_p ve S_p ölçütleri aşağıdaki gibi yazılır.

$$C_p = RED_p/MSE_k + 2p-k \quad (3.6)$$

$$S_p = (RED_p + SSE_k)/(n-p)(n-p-2) \quad (3.7)$$

C_p ve S_p deki değişimin aynı olduğu açıktır.

R^2p ölçütüne alternatif olarak düzeltilmiş R^2_p önerilmiştir. R^2p ve MSE_p arasındaki ilişki aşağıdaki gibidir:

$$R^2p= 1 -[(n-1)MSEp /SST] \quad (3.8)$$

(3) R.R. Hocking, a.g.e., s.19.

MSE_p en küçük olduğunda R^2_p en büyük değeri alacaktır.

Mallow tanımlama amacına yönelik model çalışmalarında R^2_p'yi, ekstrapolasyon ve parametre tahmin amacına yönelik model çalışmalarında C_p 'yi önermiştir. MSEp ölçütü ise ;

$((n-k-1)/(n-p))MSE_k <= MSE_p <= ((n-k)/(n-p))MSE_k$

kısıtlarını sağlayan modeller için aynı amaca uygun kabul edilebilir.

Sırasal seçim yöntemlerinde ise çözüm süreci olası bütün regresyon çözümlerine göre daha kolaydır ve bir karar ölçütü kullanımı gerekli değildir. İleri doğru seçim yönteminde k sayıda X modele alınsa da yalnız k(k+1)/2 sayıda model çözümü gereklidir. Benzer şekilde geriye doğru eleme yönteminde de k sayıda regresyon eşitliği ve k(k+1)/2 sayıda SSE hesaplanır. Kuşkusuz bu 2^k-1 sayıda model çözmekten daha kolaydır. Ancak bu yöntemler diğer değişkenlere göre daha önemli X leri seçerken, uygulayıcıyı, amaca belki de daha uygun olabilecek farklı modelleri incelemekten alıkoymaktadır. Özellikle yapısal analiz amacına yönelik model çalışmalarında farklı modelleri bir arada görerek sonuca gitmek gereklidir.

İleri doğru seçim yönteminde 1. adımda en önemli değişken olarak modele alınan değişken geriye doğru eleme yönteminde önemsiz kabul edilerek model dışında kalabilmektedir. Bu, iki yöntem arası bir çelişkidir.

Stepwise regresyon ise her yeni değişkenin yanı sıra, modele daha önceki adımlarda alınmış değişkenleri de sınamaktadır. Bu, stepwise regresyonun diğerlerine göre üstünlüğüdür. Ancak bu yöntem k sayısının fazla olduğu ve X lerle Y lerin korelasyonlarının düşük olduğu durumlarda her zaman uygun çözüm yapamaz. Bunun bir örneği incelenen uygulamada görülmüştür. Yöntemin bu güçsüzlüğü kritik F değerini serbestlik derecelerinden bağımsız olan standart bir değer olarak belirlemekle giderilebilir. Paket programlarda bu standart değer $F_c=2$ olarak tanımlanmış-tır.

Stepwise regresyon X ler arasında korelasyonun yüksek olduğu; değişkenler arası çoklu doğrusal bağlantının bulduğu modellerde de güçsüzdür. Hocking, çoklu doğrusal baglantının bulunduğu modellerde en küçük kareler yöntemi yerine yanlı tahmin teknikleri kullanımını önermektedir.

SONUÇ

Sonuç olarak seçim süreçleri

i. fazla sayıda değişken içermeyen,

ii. ve buna rağmen, incelenen konu hakkında olabildiğince çok bilgi veren modeli aramaktır.

Çalışmada incelenen seçim yöntemlerinin birbirlerine göre zayıf ve üstün yönleri üçüncü bölümde belirtilmişti. Bir yöntemin güçsüzlüğünü diğer bir yöntemin üstünlüğünü kullanarak eleyen aşağıdaki çözüm süreci bir alternatif olarak önerilebilir:

1) Bağımsız değişken seçimine (2.3.5.) de incelenen stepwise regresyon yöntemi ile başlanır. Anımsanacağı gibi bu yöntem $F_s < F_c$ olana dek modele değişken alınmasını içeriyordu.

2) $F_s < F_c$ olan aşamada stepwise regresyon ile değişken alımı durdurulur.Bundan sonraki aşamalar için F_s ler F_c=2 değeri için sınanır. Bu koşulu sağlayan değişkenler de modele alınır.

3) $F_s < 2$ olduğunda stepwise regresyon ile değişken alımı durdurulur.

4) Bu aşamada, modelde bulunan X ler yeni bir alt bağımsız değişken listesi olarak kabul edilir ve bu aşamaya dek modelden "düşmüş" bağımsız değişkenler de bu listeye eklenir.

5) Bu yeni bağımsız değişken listesi için, (2.1) de incelendiği gibi, bütün regresyon çözümleri yapılarak uygun bir karar ölçütüne göre final modeli saptanır.

Böylece,

- 2^k-1 sayıda model çözülmeden ve
- önemli kabul edilen değişkenler için farklı model kompozisyonlarını da bir arada görerek değişken seçimine gidilir.

KAYNAKÇA

DRAPER N. & SMITH H., Applied Regression Analysis , Newyork, 1966

NETER J. & WASSWRMAN W., Applied Linear Statistical Models, Newyork, 1974

AMEMIYA Takeshi, Selection of Regressors, International Economic Review, Vol.24, No2, June, 1980

HOCKING R.R. , The Analysis & Selection of Variables In Linear Regression, Biometrics,32, 1-49, March 1976

THOMPSON Mary L., Selection of Variables In Multiple Regresion, International Statistical Review,46 1978 1-19.