21. Yüzyılda Fen ve Teknik / Science And Technıque in The 21st Century
Cilt / Volume 1, Sayı / Issue 22, Kış / Winter
Derleme Makalesi / Review Article
Sayfa /Page 65-74

Tarımsal Araştırmalarda Kullanılan Çoklu Doğrusal Regresyonda Değişken Seçimi

Selection of Variables in Multiple Linear Regression Used in Agricultural Research

Volkan KARADAVUT[1] Mehmet Emin YAZICI[2] Ufuk KARADAVUT[3]

Öz:

Çoklu dorusal regresyon analizinde, bağımlı değişkeni en iyi açıklayabilecek model denklemini, kullanım amacına göre elde etmek içim, mevcut olan bağımsız değişkenler arasından seçim yapmak en önemli aşamadır. Değişken seçimi olarak bilinen bu aşamada, öncelikle alt küme modellerini bulmak ve bir alt kümenin diğerlerinden daha iyi olduğuna karar vermektir. Bu çalışmada uygulamada en çok kullanılan değişken seçim yöntemlerinden ileriye doğru seçim, geriye doğru seçim, adımsal seçim ve olası bütün alt kümeler yöntemleri anlatılmıştır. Bu çalışmada, çoklu doğrusal regresyon analizinde değişken seçim yöntemlerini bakla bitkisi (*Vicia faba* L.)'nden alınan verilerle uygulamasını yaparak araştırmacıların bilgisine sunmaktır.

Anahtar sözcükler: Değişken seçimi, Regresyon analizi, Bakla, Tarım

Abstract:

In multiple linear regression analysis, to make a choice among the present independent variable in order to form the model equation which can best explain the dependent variable as fitting the aim of its usage is one of the most important steps of multiple linear regression analysis. At this step known as variable choice. The aim of finding the models or worse than another one. In this study, it explained that the most useful in application forward selection, backward selection, stepwise selection and selection of all likelihood subsets. The aim of this study is to introduce the variable choice methods used some values observed from the Faba bean plants (*Vicia faba* L.) to researchers studying in agricultural research.

Keywords: Choice of variable, Regression analysis, Faba bean, Agriculture

[1] Cumhuriyet Üniversitesi Timur Karabal Meslek Yüksek Okulu, Suşehri-Sivas, Sorumlu yazar; vkaradavut@hotmail.com
[2] Ahi Evran Üniversitesi Fen Bilimleri Enstitüsü, Cacabey Yerleşkesi, Kırşehir
[3] Ahi Evran Üniversitesi Ziraat Fakültesi, Cacabey Yerleşkesi, Kırşehir

Makale Bilgisi /Article İnfo
Geliş / Received: 21.07.2014 – Kabul Accepted: 10.12.2014

GİRİŞ

Tarımsal araştırmaların çoğunda çoklu regresyon analizi sıklıkla kullanılmaktadır. Tarımsal araştırmalarda çok fazla karakterim ölçümü yapılmakta, zaman ve emek harcanmaktadır. Özellikle ıslah çalışması yapan enstitülerde bu iş binlerce parselde , binlerce hattan veri almak olduğunda işler daha da zorlaşmakta ve hatta belli bir noktadan sonra imkansızlaşabilmektedir. Bu nedenle değişken seçimi ön plana çıkmakta ve bağımlı değişkeni en iyi şekilde açıklayabilecek en az bağımsız değişkenin tespit edilmesi yoluna gidilmektedir.

Değişken seçiminde en önemli problem, seçilen alt kümenin diğerinden daha iyi olup olmadığına karar verebilmek için belirli bir ölçütün gerekliliğidir (İpek, 2002). Çoklu doğrusal regresyon analizinde uygun alt küme ya da kümeler üzerinde karar vermek için farklı ölçütler bulunmaktadır. Bu ölçütleri kullanırken modelin amacının dikkate alınması gerekmektedir. En çok kullanılan değişken seçim ölçütleri şunlardır; 1) İleriye doğru seçim, 2) Geriye doğru seçim, 3) Adımsal regresyon, 4) Her bir değişkenin ya da değişkenlerin ayrı ayrı regresyonun yapılmasından oluşan olası tüm alt kümeler yöntemidir.

Değişken seçiminde kullanılan teknikler temelde, bağımsız değişkenlerce bağımlı değişkenler için oluşturulacak tahmin ya da model yapısının bütününü temelde koruyarak, bağımsız değişkenin sayısını azaltmaktır. Değişkenler şu nedenlerden dolayı azaltılır (Miller, 1984); a) Maliyetleri azaltmak, b) Modele katkısı az olan değişkenleri çıkartmak, c) Bağımsız değişkenlerden bazıları yüksek derecede ilişkili olduklarından, regresyon katsayılarını küçük varyanslı olarak tahmin etmek. Weisberg (1980), değişen seçiminde iki temel konu olduğunu belirtmektedir. Bunlar, 1) Bir alt kümenin değişkenden daha iyi olup olmadığına karar vermek için özel bir ölçütün gerekliliği, 2) Ele alınması gerekli alt kümelerin sayılarının fazla olması durumunda hesaplamanın fazla zaman almasıdır.

Bütün bunların yanında regresyon modelinin tahmini amacıyla kullanımda çok sayıda değişkenin büyük bir tahmin varyansına neden olacağı, az sayıda değişkenin olması ise yanlı bir tahmin vereceği ve bu durumun alt küme seçimi için bir neden teşkil ettiği bilinmektedir (Düzgüneş ve ark., 1986). Seçim işlemi ve değişken sayısının azaltılması regresyon modelinin kullanım amacına ve verilerin yapısına bağlıdır. Eğer amaç, iç değer bulma ise daha az sayıda değişken istenir. Dış değer bulmak ise, etkili tüm değişkenlerin modelde bulunması tercih edilir (İpek, 2002). Bundan dolayı her amaç için farklı değişken seçimi kullanılması gerekmekte ve her amaç için farklı modeller oluşturulabilmektedir. Bu çalışmada amacımız tarımsal araştırmalarda çok değişkenli çalışıldığı için oldukça çok zaman ve masraf

yapılmaktadır. Bu nedenle değişken azaltma yöntemlerinden bazıları göstererek araştırmacıların bilgilendirilmesidir.

Materyal ve Metod

Yapılan çalışmada materyal olarak Hatay ekolojik koşullarında yetiştirilmiş olan bakla (*Vicia faba* L.) bitkisine ait veriler kullanılmıştır. Dört farklı değişken alınarak uygulamanın daha iyi anlaşılması için basitleştirilmeye çalışılmıştır. Bitkilerden bitki boyu (X_1), bitkide dane verimi (X_2), bitkide dal sayısı (X_3) ve ilk bakla yüksekliği (X_4) ölçülerek değerlendirmeye alınmıştır. Bitkiler ekim ayı içersinde ekilmişler ve mayıs ayının son haftasında hasat edilmişlerdir. Ekim 30 cm sıra arası ve 15 cm sıra üzeri mesafesine göre el ile yapılmıştır. Ekimden önce nodozite faaliyetinin teşvik edilmesi için 2.5 kg/da saf azot ve 5 kg/da hesabıyla fosfor (P_20_5) verilmiştir. Ekim işleminden sonra iki kez el ile yabancı ot mücadelesi yapılmış ve bu işlem için ilaç kullanılmamıştır. Elde edilen verilerden değişken seçim işlemleri şu şekilde yapılmıştır;

DEĞİŞKEN SEÇİMİ

1.1. Adımsal Yöntemler

a) İleriye doğru seçim: Bu yöntemde bağımsız değişkenlerin bağımlı değişken ile korelasyonları bulunur. İlk adımda bağımlı değişkenle en yüksek korelasyonu olan bağımsız değişken denkleme alınır. Modelin F testi yapılır ve önemli bulunursa, bağımlı değişkenle en yüksek korelasyona sahip değişken denklemme alınır. Son giren bağımsız değişken için $H_0 : \beta_i = 0$ hipotezinin F testi yapılır. F testi;

$$F_{(i)} = \frac{HKT_{(p)} - HKT_{(p+(i))}}{\hat{\sigma}^2_{(p+(i))}}$$

eşitliği ile hesaplanır. Burada HKT_p = p değişkenli modelin hata kareler toplamı, $HKT_{(p+(i)}$= i. ci değişken elendikten sonra p+1 sayıdaki değişken üzerinden bulunan hata kareler toplamı, $\hat{\sigma}$ = p+(i) değişkenli modelin varyansıdır. Test sonucunda bulunan $F_{(i)}$ değeri ile $F_{(T)(\alpha:1,\ n-p)}$ değeri karşılaştırılır. $F_{(i)} > F_{(T)}$ ise H_0 red edilir. Böylece i. değişken denkleme dahil edilir.

Eğer H_0 kabul edilirse işleme son verilir ve bir önceki p değişkenli model en iyi alt küme denklemi olarak karar verilir.

b) Geriye doğru çıkarma: Bu yöntem ileriye doğru seçimin tam tersidir. Yani bütün bağımsız değişkenlerin bulunduğu denklemden bağımsız değişkenlerin tek tek çıkartılmasıyla yapılır. Her bir adımda denklemden her bir bağımsız değişken için F testi yapılır. F testi;

$$F_{(i)} = \frac{HKT_{(p-(i))} - HKT_{(p)}}{\hat{\sigma}^2_{(p)}}$$

eşitliği ile hesaplanır. Burada HKT_p = p değişkenli modelin hata kareler toplamı, $HKT_{(p-(i)}$= i. ci değişken çıkartıldıktan sonra p-1 sayıdaki değişken üzerinden bulunan hata kareler toplamı, $\hat{\sigma}$ = p bağımsız değişkenli modelin varyansıdır. Test sonucunda bulunan $F_{(i)}$ değeri ile $F_{(T)(\alpha:1, n-p)}$ değeri karşılaştırılır. $F_{(i)} < F_{(T)}$ ise i. değişken denklemden çıkarılır. $F_{(i)} > F_{(T)}$ ise işleme son verilir. Böylece i. değişkenin de bulunduğu alt küme denkleminin en iyi model olduğu kararı verilir.

c) Adımsal (Stepwise) regresyon: Denklem seçme yöntemlerinden en çok kullanılanı Adımsal regresyon yöntemidir.İleriye doğru seçim ve geriye doğru çıkarma yöntemlerinin bileşimidir. Bu yöntemin ilk adımında bağımlı değişkenle en yüksek korelasyonlu bağımsız değişken denkleme alınır. Her adımda kısmi F testi yapılır. Ancak, denklem içindeki bütün bağımsız değişkenler için ayrı ayrı kısmi F değeri hesaplanır. Öyleki, bir önceki adımda önemli bulunan bir değişken bir sonraki adımda önemsiz bulunabilir. Her adımda denkleme hangi yeni değişkenin katılması gerektiğini yine kısmi korelasyon katsayısı ile belirlenir. Bu işlemler böylece sürdürülür. Yöntemin sona ermesi ileriye doğru seçim yönteminde olduğu gibidir (Montgomery ve Pack, 1982).

1.2. Tüm Olası Alt Kümeler Tekniği:

k sayıda bağımsız değişken olduğunda, bu değişkenlerin bütün kombinasyonlarını temsil eden 2^k-1 tane olası alt küme vardır. Bu kümelerden hangisinin en iyi alt küme olduğunu belirlemek için; Hata kareler ortalaması, Çoklu ve Düzeltilmiş belirleme katsayıları ile Standartlaştırılmış toplam hata kareler ortalaması (Mollows C_p istatistiği) kullanılmaktadır.

1) Hata kareler ortalaması: p sayıdaki değişkenin (bağımlı değişken dahil) denklem için hata kareler ortalaması (HKO_p)**,**

$$HKO_p = \frac{HKT_p}{(n-p)}$$

eşitliği ile bulunmaktadır. Bütün denklemleri hata kareler ortalamaları bulunduktan sonra alt küme modelin seçimi için iki yol bulunmaktadır. Bunlar;

- En küçük HKO'na karşılık gelen alt küme seçilir,
- Her p büyüklüğü için en küçük HKO'nı veren alt küme bulunur.

2) Çoklu ve düzeltilmiş belirleme katsayıları: Her bir alt küme denklemlerinden,

$$R_p^2 = 1 - \frac{HKT_p}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

yardımıyla 2^k-1 sayıda çoklu belirleme katsayıları bulunur. Her p büyüklüğü için en büyük R^2 değerini veren alt küme seçilir. Bazı durumlarda R^2 nin sürekli artan bir değer olmasından dolayı seçenek olarak;

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

eşitliği ile bulunan düzeltilmiş çoklu belirleme katsayısı en iyi alt kümeyi belirlemede kullanılır. Bu ölçüte göre, iyi modeller $\bar{R}_p^2$'nin büyük değere sahip olan alt kümeler olmaktadır (Wisberg, 1980).

3. Standartlaştırılmış toplam hata kareler ortalaması (Mollows C_p istatistiği): Mollows tarafından önerilen C_p ölçütü mevcut veri kümesi için her bir gözlem üzerinden tahmini hata kareler ortalaması toplamının standartlaştırılması olarak tanımlanır ve,

$$C_p = \frac{HKT_p}{\hat{\sigma}^2} + 2p + n$$

eşitliği ile bulunur. Burada verilen C_p ölçütü,tahminin varyans ve yanını birlikte içeren bir ölçüttür (ipek, 2002). Burada p bağımlı ve bağımsız değişkenlerin sayısı yani parametre sayısıdır. Tüm olası alt kümeler için bulunan C_p değerleri birbirleriyle karşılaştırılarak en iyi alt küme modeli seçilir. Bunun için iki yol önerilmektedir (Montgomery ve Pack, 1982);

- En küçük C_p değerini veren alt küme yada alt kümeler seçilir,
- C_p'nin p'ye karşı çözümleri yapılarak en iyi alt kümeler bulunur.

Araştırma Bulguları

Yapılan çalışmada önce ileriye doğru seçim işlemi yapılmıştır. Buna göre elde edilen sonuçlar Çizelge 1'de gösterilmektedir.

Çizelge 1. Değişken Seçiminde İleriye Doğru Seçim

Model	B	Std. Sapma	t	P	R^2	F
Sabit	80.648	4.861	16.65	0.000	52.5	9.95
X_1	1.6788	0.5221	3.16	0.012		
Sabit	51.671	2.560	20.19	0.000	97.8	174.57
X_1	1.4751	0.1236	11.93	0.000		
X_2	0.68559	0.05394	12.71	0.000		
Sabit	48.110	4.128	11.65	0.000	98.1	119.56
X_1	1.6670	0.2140	7.79	0.000		
X_2	0.6768	0.05391	12.55	0.000		
X_3	0.2126	0.1948	1.09	0.311		
Sabit	34.161	9.224	3.70	0.010	98.7	98.7
X_1	2.0402	0.2965	6.88	0.000		
X_2	0.76097	0.07023	10.84	0.000		
X_3	0.5808	0.2832	2.05	0.086		
X_4	0.10696	0.6480	1.65	0.150		

İleriye doğru seçim işleminde değişkenler birer birer denkleme alınırlar. Değişken alma işlemi belirleme katsayıları arasındaki farkın önemli çıktığı noktaya kadar devam etmektedir. Çizelge 1 incelendiğinde X_3 ve X_4 'ün denklemden atılması gerektiği

görülmektedir. Çünkü, olasılık (P) değerine baktığımızda her ikisinin de oldukça yüksek değerler aldıkları görülmektedir. 3 ve 4 değişkenli denklemlerde belirleme katsayısı yüksek olmasına rağmen önemlilik kontrolünde önemsiz çıkmaları nedeniyle denklemden çıkarılmışlardır. Buna göre ileriye doğru seçimde genel denklemimiz;

$$Y = 51.671 + 1.4751\ X_1 + 0.68559\ X_2$$

şekilde olmuştur.

Yapılan çalışmada yapılan geriye doğru seçim işleminden elde edilen sonuçlar Çizelge 2'de gösterilmektedir.

Çizelge 2. Değişken Seçiminde Geriye Doğru Seçim

Model	**B**	**Std. Sapma**	**t**	**P**	**R^2**	**F**
Sabit	34.161	9.224	3.70	0.010	98.7	112.44
X_1	2.0402	0.2965	6.88	0.000		
X_2	0.76097	0.07023	10.84	0.000		
X_3	0.5808	0.2832	2.05	0.086		
X_4	0.10696	0.6480	1.65	0.150		
Sabit	48.110	4.128	11.65	0.000	98.1	119.56
X_1	1.6670	0.2140	7.79	0.000		
X_2	0.6768	0.05391	12.55	0.000		
X_3	0.2126	0.1948	1.09	0.311		
Sabit	51.971	2.560	20.19	0.000	97.8	174.57
X_1	1.4751	0.1236	11.93	0.000		
X_2	0.68559	0.05394	12.71	0.000		
Sabit	80.648	4.861	16.65	0.000	52.5	9.95
X_1	1.6788	0.5221	3.16	0.012		

Çizelge 2 incelendiğinde geriye doğru seçimde bütün değişkenler denklemde kullanılmakta ve bire birer denklemden çıkarılmaktadır. Bu belirleme katsayılarını arasındaki farkın önemli olduğu noktaya kadar devam etmektedir. Buna göre X_4 ve X_3 değişkenleri denklemden atılmışlardır. Buna göre geriye doğru seçimde denklemimiz;

$$Y = 51.671 + 1.4751\ X_1 + 0.68559\ X_2$$

şekilde olmuştur.

Yapılan çalışmada yapılan Adımsal (Stepwise) seçim işleminden elde edilen sonuçlar Çizelge 3'de gösterilmektedir.

Çizelge 3. Değişken Seçiminde Adımsal (Stepwise) Seçim

Model	**B**	**Std. Sapma**	**t**	**P**	**R^2**	**F**
Sabit	51.971	2.560	20.19	0.000	97.8	174.57
X_1	1.4751	0.1236	11.93	0.000		
X_2	0.68559	0.05394	12.71	0.000		

Yapılan adımsal seçim işleminde bilgisayarımız bize yalnızca belirlenen sonuçları vermektedir. Bu sonuçlar Çizelge 3' te gösterilmektedir. Buna göre X_1 ve X_2 değişkenleri denklemde kalabilmişlerdir. Buna göre adımsal seçimde denklemimiz;

$$Y = 51.671 + 1.4751\ X_1 + 0.68559\ X_2$$

şekilde olmuştur.

Çizelge 3. Değişken seçiminde mümkün olan bütün alt kümeler yöntemi ile seçim

Model	Değişken sayısı	HKO	R^2	$\bar{R}^2$	C_p
1	X_1	119.63	52.5	47.2	220.1
1	X_2	84.97	57.9	46.2	195.7
1	X_3	212.65	53.8	44.1	193.8
1	X_4	84.12	55.0	47.9	206.4
2	$X_1\ X_2$	6.28	97.8	97.2	5.7
2	$X_1\ X_3$	133.84	96.3	94.8	8.1
2	$X_1\ X_4$	9.67	98.1	92.6	9.5
2	$X_2\ X_3$	48.67	81.5	76.9	83.5
2	$X_2\ X_4$	98.74	91.6	90.4	13.6
2	$X_3\ X_4$	22.58	93.7	89.6	11.8
3	$X_1\ X_2\ X_3$	7.89	98.1	97.3	6.2
3	$X_1\ X_2\ X_4$	8.56	95.6	93.5	7.4
3	$X_1\ X_3\ X_4$	8.61	91.1	90.8	9.6
3	$X_2\ X_3\ X_4$	13.54	94.7	91.7	8.3
4	$X_1\ X_2\ X_3\ X_4$	7.96	98.7	97.8	5.3

Tartışma Ve Sonuç

Yapılan çalışmada değişken sayımız yarı yarıya azalmıştır. Bunun en önemli faydası tarlaya çıkan araştırmacıların fazlaca önemli olmayan değişkenleri ölçmelerine gerek kalmadan belli sayıda ki belirli özellikleri ölçmelerinin yeterli olacağıdır. Buna göre hem zamandan büyük ölçüde kazanım sağlanabilecek hem de masraflar azaltılabilecektir. Çünkü, özellikle araştırma enstitülerinde çok geniş alanlarda çalışılmakta ve çok sayıda teknik personel bu konuda emek harcamaktadır. Çok sayıda değişkenin ölçülmeye çalışılması bazen ölçümlerin dikkatsiz yapılması, farklı kişilerin ölçmesi nedeniyle ölçüm hatasının olması gibi sakıncaları da beraberinde getirebilmektedir. Değişken sayısının azaltılması ile bu tür olumsuzluklar en aza indirilebilecektir. Bu çalışma sonucunda adımsal değişken seçim yönteminin hem daha hızlı ve hem de daha kolay olması nedeniyle öncelikli olarak değerlendirilmesinin uygun olacağı sonucuna varılmıştır.

Kaynaklar

Draper, N.R.; Smith, H. 1981. Applied Regression Analysis. Wiley, New York.

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

İpek, O. 2003. Çoklu doğrusal regresyonda değişken seçimi. Kara Harp Okulu Dergisi, İstanbul.

Miller, A.J. 1984. Selection of subsets regression variables. J. R. Stat. Society Ser. A, Part 3.

Montgomery, D.C.; Peck, E. A. 1982. Introductionto linear regression analysis. Wiley, New York.

Wiseberg, S. 1980. Applied linear regression. Wiley, New York.