HAFTA 14

10.4. Çoklu Bağıntı (Multicollinearity)

Regresyon modeli,

$$Y_i = \alpha_0 + \alpha_1 x_{1,i} + \alpha_2 x_{2,i} + ... + \alpha_p x_{p,i} + e_i \ , \ i = 1, 2, 3, ..., n$$

şeklinde verilmiş olsun. Açıklayıcı değişkenlerden en az biri, diğer açıklayıcı değişkenlerin bir lineer birleşimi ise *çoklu bağıntı* problemi vardır. Bu durumda normal denklemlerin çözümü ya yoktur, ya da birden fazla çözümü vardır. Normal denklemler her zaman tutarlı (Tanım (8.3.1) anlamında tutarlılık) olduğundan, çoklu bağıntı probleminin olması halinde normal denklemlerin birden fazla çözümü vardır. Regresyon modeli, $\underset{\sim}{y} = X \underset{\sim}{\beta} + \underset{\sim}{e}$ şeklinde yazıldığında, X matrisinin kolonlarından enaz biri diğer kolonların lineer birleşimi olarak yazılabiliyorsa, $X'X$ matrisi singülerdir. Yani, $X'X$ matrisinin determinantı sıfırdır. Bazen, $X'X$ matrisinin determinantı sıfıra çok yakın olabilir.

Açıklayıcı değişkenlerin sayısı az olduğunda, açıklayıcı değişkenler arasında bir ilişki varsa bunun ortaya çıkarılması kolaydır. Ancak, regresyonda açıklayıcı değişkenlerin sayısı fazla ise bunun görülmesi kolay olmayabilir. Çoklu bağıntının varlığını ortaya çıkaran bir çok gösterge vardır. Bunlardan bazıları aşağıda özetlenmiştir.

i) En basit şekilde, iki açıklayıcı değişken arasındaki korelasyon katsayısı 1'e yakın olabilir. Böyle bir durumda, çoklu bağıntıdan şüphelenilebilir.

ii) Regresyon modelinin anlamlılığı konusunda bilgi veren F istatistiği ile katsayılara ilişkin $t-$istatistikleri uyumlu olmayabilir. Örneğin, F istatistiği anlamlı olmasına rağmen, $t-$istatistiklerinin hepsi anlamlı olmayabilir (Örnek (10.3.1) deki gibi). Böyle bir durum çoklu bağıntıya işaret eder.

iii) Regresyon katsayıları için oluşturulan güven aralıkları oldukça geniş ise açıklayıcı değişkenler arasında çoklu bağıntı olabilir.

iv) Bir bağımsız değişkenin modele eklenmesi ya da çıkarılması halinde, regresyon katsayılarında büyük değişiklikler oluyorsa çoklu bağıntıdan şüphelenilebilir (Örnek (10.3.1) de $X_3 = X_1 X_2$ değişkeninin modele katıldığında sonuçları karşılaştırınız).

Çoklu bağıntının derecesi ve varlığı ile ilgili literatürde bir çok ölçüt bulunmaktadır. Bunlardan bazılarını açıklamak için aşağıdaki üç açıklayıcı değişkenli,

$$Y_i = \alpha_0 + \alpha_1 x_{1,i} + \alpha_2 x_{2,i} + \alpha_3 x_{3,i} + e_i \ , \ i = 1, 2, 3, ..., n$$

regresyon modeli ele alınmıştır. Model, $\underset{\sim}{y} = X\underset{\sim}{\beta} + \underset{\sim}{e}$ olarak yazıldığında, $X'X$ matrisinin özdeğerlerini $k = 1,2,3,4$ için λ_k ile gösterelim. Bazen, kesim noktasına göre ortogonalliği sağlamak için X matrisi merkezileştirilir (ortalamaları çıkartılarak). Buna göre, X matrisinin *koşul sayısı(condition number),* $K(X) = \lambda_{\max} / \lambda_{\min}$ değeri veya buna benzer olarak *koşul indeksi(condition index)* $k = 1,2,3,4$ için $\delta_k = \lambda_{\max} / \lambda_k$ değerleri hesaplanır. Buradaki, en büyük koşul indeksi aynı zamanda koşul sayısıdır. Koşul indeksi bazen $\delta_k = \sqrt{\lambda_{\max} / \lambda_k}$ olarak da verilir. Buna göre, koşul indeksi 10 civarında ise, zayıf bir bağımlılıktan söz edilebilir. Eğer koşul indeksi 30 ile 100 arasında ise, basitten güçlüye doğru çoklu bağıntı vardır. Koşul indeksi, 100 ün üzerinde ise, ciddi çoklu bağıntı problemi vardır (Rawlings, Pantula ve Dickey, 1998, sayfa 371).

Bir diğer çoklu bağıntı ölçütü ise, çoklu bağıntı indeksi olarak bilinen, *mci* indeksidir. $p' = Iz(X'X)$ olmak üzere, $X'X$ matrisinin özdeğerleri λ_k olsun. $\lambda_{p'}$ en küçük özdeğeri göstermek üzere,

$$mci = \sum_{j=1}^{p'} (\lambda_{p'} / \lambda_j)$$

çoklu bağıntı indeksi tanımlanır. *mci* nin değeri 1 civarında ise güçlü bir çoklu bağıntı vardır. Bu değer 2 den büyük ise, çoklu bağıntı yoktur veya zayıftır (Rawlings, Pantula ve Dickey, 1998, sayfa 371).

Çoklu bağıntı için diğer bir ölçüt ise, *varyans şişirme katsayısı* (variance inflation factor) *VIF* dir. X tasarım matrisinin kolonları merkezileştirildikten sonra (bunları X_i^* ile gösterelim) X_i^* ın diğer açıklayıcı değişkenler üzerine regresyonundan elde edilen çoklu belirleme katsayıları R_i^2 olsun. Buradan açıklayıcı değişkenler için varyans şişirme katsayıları,

$$VIF_i = 1/(1 - R_i^2)$$

olarak tanımlanır. Eğer, X matrisinin kolonları birbirlerine ortogonal ise, R_i^2 değerleri sıfırdır. Bu durumda *VIF* değerleri 1 olur. Böyle bir durumda çoklu bağıntı yoktur. VIF_i değerleri 10 dan büyük ise, X_i açıklayıcı değişkeni ile diğer açıklayıcı değişkenler arasında ciddi bir bağımlılık vardır (Rawlings, Pantula ve Dickey, 1998, sayfa 373).

Açıklayıcı değişkenler arasındaki korelasyonlar,

X_1 ile X_2 arsındaki korelasyon : $r_{12} = \dfrac{\sum_{i=1}^{n} (x_{1,i} - \bar{x}_1)(x_{2,i} - \bar{x}_2)}{\sqrt{\sum_{i=1}^{n} (x_{1,i} - \bar{x}_1)^2 \sum_{i=1}^{n} (x_{2,i} - \bar{x}_2)^2}}$

X_1 ile X_3 arsındaki korelasyon : $r_{13} = \dfrac{\sum_{i=1}^{n}(x_{1,i}-\bar{x}_1)(x_{3,i}-\bar{x}_3)}{\sqrt{\sum_{i=1}^{n}(x_{1,i}-\bar{x}_1)^2 \sum_{i=1}^{n}(x_{3,i}-\bar{x}_3)^2}}$

X_2 ile X_3 arsındaki korelasyon : $r_{23} = \dfrac{\sum_{i=1}^{n}(x_{2,i}-\bar{x}_2)(x_{3,i}-\bar{x}_3)}{\sqrt{\sum_{i=1}^{n}(x_{2,i}-\bar{x}_2)^2 \sum_{i=1}^{n}(x_{3,i}-\bar{x}_3)^2}}$

şeklinde hesaplanır. Örneğin, açıklayıcı değişkenler arasındaki korelasyonlar

$$\begin{bmatrix} 1 & r_{12} & r_{13} \\ r_{21} & 1 & r_{23} \\ r_{31} & r_{32} & 1 \end{bmatrix} = \begin{bmatrix} 1.000 & 0.995 & 0.250 \\ 0.995 & 1.000 & 0.352 \\ 0.250 & 0.352 & 1.000 \end{bmatrix}$$

gibi elde edilmiş ise, X_1 ve X_2 değişkenleri arasında yüksek bir korelasyon gözlenmektedir. Dolayısı ile, çoklu bağıntı probleminden söz edilebilir. Böyle bir durumda, regresyon analizi için lineer ilişkinin gözlendiği açıklayıcı değişkenlerden bir tanesini modelden atmak en basit çözümdür (Mickey, Dunn ve Clark, (2004), sayfa 324). Açıklayıcı değişkenler arasındaki ikili korelasyonlara bakılarak çoklu bağıntının varlığı için sezgisel bir karar verilebilir. Açıklayıcı değişkenlerin ikili saçılım grafiklerine bakılarak da çoklu bağıntı hakkında bir ön bilgi elde edilebilir.

Örnek 10.4.1 Örnek (10.3.1) deki regresyon modelini tekrar göz önüne alalım. Orada, $X_3 = X_1 X_2$ alındığında, üç açıklayıcı değişkenli regresyon modeli,

$$Y_i = \alpha_0 + \alpha_1 x_{1,i} + \alpha_2 x_{2,i} + \alpha_3 x_{3,i} + e_i \ , \ i = 1, 2, 3, ..., n$$

olarak ele alınmış, F istatistiği anlamlı olmasına rağmen, $t-$istatistiklerinin hiç biri anlamlı bulunamamıştı. Buna göre, çoklu bağıntının varlığından şüphelenilebilir. Veriler aşağıda tekrar yazılmıştır.

$X_1 = IQ$	105	110	120	116	122	130	114	102
$X_2 =$ Süre	10	12	6	13	16	8	20	15
$X_3 = X_1 X_2$	1050	1320	720	1508	1952	1040	2280	1530
Y	75	79	68	85	91	79	98	76

Bu regresyon modeline göre, ANOVA tablosu ve parametrelere ilişkin istatistiki sonuçlar da aşağıda tekrar yazılmıştır.

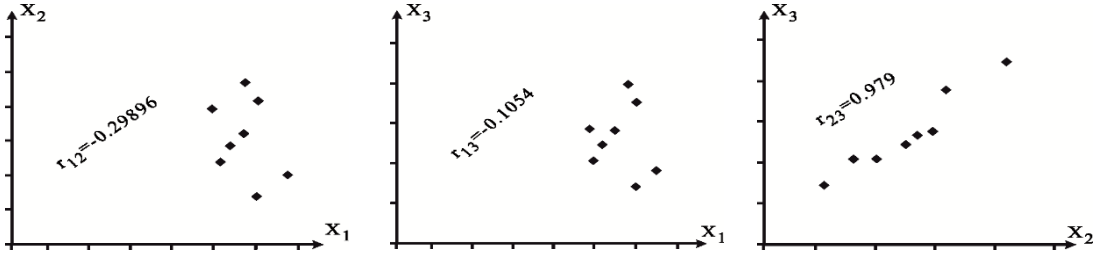
Kaynak	Serbestlik Derecesi	Kareler Toplamı	Ortalama Kareler Toplamı	F Değeri

Regresyon	3	610.81033	610.81033	26.217
Artıklar	4	31.06467	7.76617	
Toplam	7	641.875		

ANOVA tablosundan, $F = 26.217$ olarak gözlenmiş olup bu değer anlamlıdır. Oysa, aşağıdaki tabloda parametrelere ilişkin değerler göz önüne alındığında $t-$istatistiklerinin hiç biri anlamlı değildir.

Parametre	Tahmin	Stand. hata	$t: H_0: \alpha_i = 0$	%5 Kritik değer	Sonuç
α_1	-0.131170	0.45529954	-0.288	2.7667	Kabul
α_2	-4.111072	4.52430095	-0.909	2.7667	Kabul
α_3	0.053071	0.03858059	1.376	2.7667	Kabul

Açıklayıcı değişkenler arasındaki saçılım grafikleri aşağıda verilmiştir. Bu grafiklerden, X_2 ile X_3 arasında doğrusal bir ilişki göze çarpmaktadır.



Şekil 10.4.1 Örnek (10.4.1) de verilen açıklayıcı değişkenler arasındaki serpilme grafikleri

Tasarım matrisi,

$$X = \begin{bmatrix} 1 & 105 & 10 & 1050 \\ 1 & 110 & 12 & 1320 \\ 1 & 120 & 6 & 720 \\ 1 & 116 & 13 & 1508 \\ 1 & 122 & 16 & 1952 \\ 1 & 130 & 8 & 1040 \\ 1 & 114 & 20 & 2280 \\ 1 & 102 & 15 & 1530 \end{bmatrix}, \quad X'X = \begin{bmatrix} 8 & 919 & 100 & 11400 \\ 919 & 106165 & 11400 & 1306102 \\ 100 & 11400 & 1394 & 158366 \\ 11400 & 1306102 & 158366 & 18068568 \end{bmatrix}$$

olup, $X'X$ matrisinin özdeğerleri,

$$\underset{\sim}{\lambda} = (18164437,\ 11692.179,\ 5.8100784,\ 0.0026387)'$$

olarak hesaplanmıştır. X matrisi ortalamalar çıkartılarak merkezileştirildiğinde özdeğerler,

$$\underset{\sim}{\lambda} = (1823712.7,\ 593.83781,\ 8,\ 0.3758126)'$$

şeklindedir. Buna göre, $X'X$ matrisine ve merkezileştirilmiş $X'X$ matrislerine göre hesaplanan $\delta_k = \sqrt{\lambda_{\max} / \lambda_k}$ koşul indeksi değerleri aşağıdadır.

$X'X$ matrisi			Merkezileştirilmiş $X'X$ matrisi		
Temel Bileşen	Özdeğer	Koşul İndeksi	Temel Bileşen	Özdeğer	Koşul İndeksi
1	18164437	1.00	1	1823712.7	1.00
2	11692.179	1553.5	2	593.83781	3071.06
3	5.8100784	3126366.9	3	8.00	227964.0
4	0.0026387	6883858339	4	0.3758126	4852718.35

Merkezileştirilmiş $X'X$ matrisi için hesaplanan $\delta_k = \sqrt{\lambda_{\max} / \lambda_k}$ koşul indeksi değerleri oldukça büyük olduğundan çoklu bağıntı vardır.

Şimdi de çoklu bağıntı için başka bir ölçüt olan *varyans şişirme katsayılarını (VIF)* hesaplayalım. Bunun için X_1 in X_2 ve X_3 üzerine regresyonundan $R_1^2 = 0.9370$, X_2 nin X_1 ve X_3 üzerine regresyonundan $R_2^2 = 0.9974$ ve X_3 ün X_1 ve X_2 üzerine regresyonundan, $R_3^2 = 0.9971$ değerleri hesaplanmıştır. Her üç durumda da R_i^2 değerleri oldukça yüksektir. Herbir açıklayıcı değişken için varyans şişirme katsayıları hesaplanarak değerleri aşağıda verilmiştir.

VIF değerleri oldukça yüksek olduğundan (bir tanesinin bile 10 dan büyük olması yeterli) açıklayıcı değişkenler arasında ciddi bir çoklu bağıntı problemi vardır⊕

Açıklayıcı değişken	R_i^2	$VIF_i = 1/(1-R_i^2)$
X_1	$R_1^2 = 0.9370$	$VIF_1 = 15.873$
X_2	$R_2^2 = 0.9974$	$VIF_2 = 384.62$
X_3	$R_3^2 = 0.9971$	$VIF_3 = 344.83$

Çoklu bağıntının gözlenmesi durumunda yapılması gereken en basit şey, o açıklayıcı değişkeni (yukarıdaki örnekte X_3 değişkeni) modelden atmaktır. Yukarıdaki örnekte, modelden X_3 değişkeni çıkartıldığında belirlenen problemlerin çoğunun ortadan kalktığı görülmektedir (Örnek (10.3.1)). Bazen, açıklayıcı değişkenin modelden atılması başka istatistiki sorunlara neden olabilir. Böyle bir durumda ise yanlı tahmin tekniklerine başvurulur. Bu tekniklerden en popüler olanı "*Ridge Regresyonu"*dur.

Ridge regresyonu, $X'X$ matrisinin singüler olması halinde, $(X'X + k\,I)$ nonsingüler olacak şekilde pozitif küçük bir k sayısının seçimine dayanır. p açıklayıcı değişkenin olduğu regresyon modeli,

$$Y_i = \alpha_0 + \alpha_1 x_{1,i} + \alpha_2 x_{2,i} + ... + \alpha_p x_{p,i} + e_i \ , \ i = 1, 2, 3, ..., n$$

olarak verilmiş olsun. X tasarım matrisinin standartlaştırılmış kolonlarına (kesim noktası hariç) $j = 1, 2, ..., p$ için z_j diyelim. Bu yeni açıklayıcı değişkenlere göre regresyon modeli,

$$Y_i = \alpha_0 + \alpha_1 z_{1,i} + \alpha_2 z_{2,i} + ... + \alpha_p z_{p,i} + e_i \ , \ i = 1, 2, 3, ..., n$$

şeklinde yazılır. Bu durumda, $\hat{\alpha}_0 = \bar{Y}_n$ olup model, $\underset{\sim}{y} = \underset{\sim}{1}\alpha_0 + Z\underset{\sim}{\beta} + \underset{\sim}{e}$ şeklinde yazıldığında, $\underset{\sim}{\beta}$ nın ridge tahmin edicisi $k \geq 0$ için,

$$\underset{\sim}{\tilde{\beta}}(k) = (Z'Z + k\,I)^{-1} Z'\,\underset{\sim}{y}$$

şeklindedir. Burada, $k = 0$ için standart en küçük kareler tahmin edicisi elde edilir. Yukarıda verilen Ridge regresyon tahmin edicisinin varyans-kovaryans matrisi,

$$Var(\underset{\sim}{\tilde{\beta}}(k)) = (Z'Z + k\,I)^{-1}(Z'\,Z)(Z'Z + k\,I)^{-1}\sigma^2$$

şeklindedir. Burada amaç, hata kareler ortalaması (MSE) en küçük olacak şekilde k sayısının seçilmesidir. Genel olarak, k sıfıra yakın bir sayıdır. k nin değeri arttıkça, $\underset{\sim}{\tilde{\beta}}(k)$ nin değerleri değişir. k arttıkça $VIF_{\max}$, $Var(\underset{\sim}{\tilde{\beta}}(k))$ matrisinin diagonal elemanları, $Iz(Var(\underset{\sim}{\tilde{\beta}}(k)))$, $(\underset{\sim}{\tilde{\beta}}(k)'\underset{\sim}{\tilde{\beta}}(k))^{1/2}$ ve R^2 değerleri azalır. Yani, bu değerler k değerleri ile ters orantılıdır. Diğer taraftan, k arttıkça $\underset{\sim}{\tilde{\beta}}(k)$ nin yanlılığı artar. $k = 0$ için hata kareler ortalaması s^2 ve kesim noktası haricindeki parametrelerin sayısı p olmak üzere, k nin seçimi için $k = p\,s^2 / (\underset{\sim}{\tilde{\beta}}(0)'\underset{\sim}{\tilde{\beta}}(0))$ önerilmektedir (Rawlings, 1988, sayfa 339). Ridge regresyon tahmin edicisi $\underset{\sim}{\tilde{\beta}}(k)$ ile standart en küçük kareler tahmin edicisi $\underset{\sim}{\hat{\beta}}$ arasında,

$$\underset{\sim}{\tilde{\beta}}(k) = \left(Z'Z + k\,I\right)^{-1} Z'\,Z\,\underset{\sim}{\hat{\beta}}$$

şeklinde bir ilişki vardır. Bu bölümü bitirmeden önce, çoklu bağıntının tespiti ve ridge regresyonu ile ilgili üç açıklayıcı değişkenli aşağıdaki regresyon modelini ele alalım. Veriler SAS da rasgele üretilen hata terimleri ile oluşturulmuştur.

Örnek 10.4.2 Üç açıklayıcı değişkenli regresyon modeli

$$Y_i = \alpha_0 + \alpha_1 x_{1,i} + \alpha_2 x_{2,i} + \alpha_3 x_{3,i} + e_i \ , \ i = 1, 2, 3, ..., n$$

olarak verilmiş olsun. Bu modele uygun olduğu düşünülen veriler aşağıdadır.

i	X_1	X_2	X_3	Y
1	**10**	**-4**	**5**	10.37
2	11	-4	4	10.62
3	12	-3	3	10.71

4	13	-2	2	11.26
5	14	-1	1	10.70
6	15	0	2	11.14
7	16	1	3	12.46
8	17	2	4	12.09
9	18	3	5	15.23
10	19	4	6	11.23
11	**10**	**-4**	**5**	11.35
12	11	-4	4	12.34
13	12	-3	3	11.59
14	13	-2	2	11.38
15	14	-1	1	11.71
16	15	0	2	14.25
17	16	1	3	12.64
18	17	2	4	12.27
19	18	3	5	15.29
20	19	4	6	13.89

Verilerden de görüleceği gibi birinci ve onbirinci değerler hariç, diğerleri arasında $X_2 = X_1 - 15$ şeklinde bir ilişki vardır. Dolayısı ile, X tasarım matrisinin kolonları arasında muhtemel bir lineer ilişkiden söz edilebilir. Böyle bir durumda, $X'X$ matrisi singüler olup, bu singülerliği kaldırmak için bu kolonlardan bir tanesi modelden atılabilir. Modelden bir değişkenin atılması istatistiki sonuç çıkarımlarda bazen sorun yaratabilir. Açıklayıcı değişkenlerin sayısı fazla ise, bu lineer ilişki burada olduğu gibi hemen görülemeyebilir.

Model: **MODELI**
Dependent Variable: Y

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	1	16.44466	16.44466	**11.935**	**0.0028**
Error	18	24.80182	1.37788		
C Total	19	41.24648			

Root MSE	1.17383	**R-square**	**0.3987**
Dep Mean	12.12600	Adj R-sq	0.3653
C.V.	9.68028		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
INTERCEP	1	7.548394	1.35079459	5.588	**0.0001**
X1	1	0.315697	0.09138262	3.455	**0.0028**

a) Y nin X ler üzerine regresyonlarından elde edilen SAS çıktıları yukarıdadır.

Model I: $Y_i = \alpha_0 + \alpha_1 x_{1,i} + e_i$, $i = 1, 2, 3, ..., n$ modeli için, ANOVA tablosu ve parametrelere ilişkin istatistiki sonuçlar yukarıdaki şekilde gözlenmiştir.

Bu modele göre, α_1 ve regresyon modeli anlamlıdır (F istatistiğinin değeri 11.935 dir). Oysa, R^2 nin değeri oldukça küçüktür.

Model II: $Y_i = \alpha_0 + \alpha_1 x_{1,i} + \alpha_2 x_{2,i} + e_i$, $i = 1, 2, 3, ..., n$ modeli göz önüne alındığında, ANOVA tablosu ile parametrelere ilişkin istatistiki sonuçlar da aşağıdadır.

Modele, X_2 açıklayıcı değişkeni eklendiğinde, birinci modelde anlamlı gözüken α_1 parametresi ikinci modelde anlamlı değildir. Bu regresyon modeli de anlamlı (hesaplanan F istatistiğinin değeri 5.678) gözükmesine rağmen, her iki parametrenin anlamsız olduğu ($t-$ istatistiklerinden) görünmektedir.

Model: **MODEL II**
Dependent Variable: Y

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	2	16.51772	8.25886	**5.678**	**0.0129**
Error	17	24.72876	1.45463		
C Total	19	41.24648			

Root MSE	1.20608	R-square	**0.4005**
Dep Mean	12.12600	Adj R-sq	0.3299
C.V.	9.94624		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
INTERCEP	1	10.881667	14.93735328	0.728	0.4762
X1	1	0.092333	1.00103968	0.092	**0.9276**
X2	1	0.236250	1.05412424	0.224	**0.8253**

Model III : $Y_i = \alpha_0 + \alpha_1 x_{1,i} + \alpha_2 x_{2,i} + \alpha_3 x_{3,i} + e_i$, $i = 1, 2, 3, ..., n$ modeli ele alınsın. Buna göre, ANOVA tablosu ve ilgili istatistikler aşağıdadır.

Model: **MODEL III**
Dependent Variable: Y

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	3	17.15632	5.71877	**3.798**	**0.0313**
Error	16	24.09016	1.50563		
C Total	19	41.24648			

Root MSE	1.22704	R-square	**0.4159**
Dep Mean	12.12600	Adj R-sq	0.3064
C.V.	10.11911		

****

Parameter Estimates

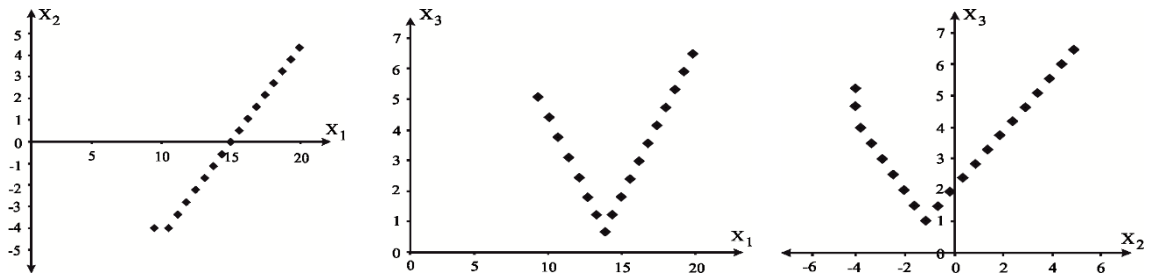
Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
INTERCEP	1	3.402083	19.04855747	0.179	0.8605
X1	1	0.556583	1.24312786	0.448	**0.6604**
X2	1	-0.279583	1.33322375	-0.210	**0.8365**
X3	1	0.154750	0.23761588	0.651	**0.5241**

Bu modele göre, regresyon modeli $\alpha = 0.05$ için anlamlı ($\alpha = 0.01$ için anlamlı değildir) olmasına rağmen, parametrelere ilişkin $t-$istatistiklerinin değerlerinden hepsi ayrı ayrı anlamlı değildir. Her üç model için, modele yeni bir açıklayıcı değişken eklendiğinde, R^2 değerinde küçük de olsa bir artış gözlenmektedir.

b) Açıklayıcı değişkenler arasındaki korelasyonlar hesaplanarak aşağıda verilmiştir.

$$\begin{bmatrix} 1 & r_{12} & r_{13} \\ r_{21} & 1 & r_{23} \\ r_{31} & r_{32} & 1 \end{bmatrix} = \begin{bmatrix} 1.000 & 0.996 & 0.290 \\ 0.996 & 1.000 & 0.342 \\ 0.290 & 0.342 & 1.000 \end{bmatrix}$$

Açıklayıcı değişkenler arasındaki grafiklerden (Şekil (10.4.2)) ve hesaplanan korelasyonlardan da görüldüğü gibi X_1 ile X_2 arasında güçlü bir ilişki göze çarpmaktadır. Dolayısı ile, $\underset{\sim}{y} = X \underset{\sim}{\beta} + \underset{\sim}{e}$ şeklinde verilen regresyon modeli için açıklayıcı değişkenler arasında muhtemel çoklu bağıntı ilişkisi vardır.



X_1 ile X_2 X_1 ile X_3 X_2 ile X_3

Şekil 10.4.2 Örnek (10.4.2) de verilen açıklayıcı değişkenlerin serpilme grafikleri

c) X_1 in X_2 ve X_3 üzerine regresyonundan R_1^2 ve VIF_1 değerleri $R_1^2 = 0.9941$ ve $VIF_1 = (1-R_1^2)^{-1} = 169.49$ olarak hesaplanmıştır. Benzer şekilde, X_2 nin X_1 ve X_3 üzerine regresyonundan $R_2^2 = 0.9943$ ve $VIF_2 = (1-R_2^2)^{-1} = 175.44$ elde edilmiş, son olarak X_3 ün X_1 ve X_2 açıklayıcı değişkenleri üzerine regresyonundan $R_3^2 = 0.4074$ ve $VIF_3 = (1-R_3^2)^{-1} = 1.69$ bulunmuştur. Varyans şişirme katsayılarının ayrı ayrı incelenmesi yerine (Neter, Wasserman ve Kutner (1985), sayfa 392) hesaplanan VIF katsayılarının ortalaması üzerinden çoklu bağıntı hakkında istatistiki sonuç çıkarım önerilmektedir. Eğer VIF lerin ortalaması anlamlı bir şekilde 1 den büyükse, çoklu bağıntı vardır. Bu örnek için

$$\overline{VIF} = \left(VIF_1 + VIF_2 + VIF_3\right)/3 = 115.54$$

olup bu değer 1 den oldukça büyüktür. Dolayısı ile, açıklayıcı değişkenler arasında anlamlı bir çoklu bağıntı vardır. Bütün göstergeler, açıklayıcı değişkenler arasında çoklu bağıntı probleminin varlığına işaret etmektedir.

d) Ridge regresyonu: X matrisinin standartlaştırılmış haline Z diyelim. X matrisinin kolonlarndan, ortalamaları çıkartılıp standart sapmalarına bölünerek Z matrisinin kolonları oluşturulmuştur. Ridge regresyon modeli, $\underset{\sim}{y} = \underset{\sim}{1}\beta_0 + Z\underset{\sim}{\beta} + \underset{\sim}{e}$ şeklinde yazıldığında, Z nin kolonları standartlaştırılmış olduğundan $\hat{\beta}_0 = \overline{Y}_n$ dır. Buradan, $\underset{\sim}{\beta}$ nın ridge tahmin edicilerinin k pozitif reel sayısına bağlı olarak, $\underset{\sim}{\tilde{\beta}}(k) = (Z'Z + k\,I)^{-1} Z'\,\underset{\sim}{y}$ şeklinde hesaplandığını biliyoruz.

Bu modele göre, regresyon kareler toplamı ile çoklu determinasyon katsayınının değerleri $SSR(k) = \underset{\sim}{\tilde{\beta}}'(k)\,Z'\,\underset{\sim}{y}$ ve $R^2(k) = SSR(k)/SST$ şeklinde olup, ridge regresyon tahmin değerlerini bulabilmek için SAS kodları bu örnek için aşağıda verilmiştir. k nin değişik değerleri için ridge tahminleri ile bu tahminlere karşılık gelen $R^2(k)$ değerleri tablo halinde aşağıdadır.

```
data a; proc iml;
 z={-1.52703    -1.28641     0.97468,-1.18769    -1.28641     0.32489,
    -0.84835    -0.92907    -0.32489,-0.50901    -0.57174    -0.97468,
    -0.16967    -0.21440    -1.62447, 0.16967     0.14293    -0.97468,
     0.50901     0.50027    -0.32489, 0.84835     0.85760     0.32489,
     1.18769     1.21494     0.97468, 1.52703     1.57227     1.62447,
    -1.52703    -1.28641     0.97468,-1.18769    -1.28641     0.32489,
    -0.84835    -0.92907    -0.32489,-0.50901    -0.57174    -0.97468,
    -0.16967    -0.21440    -1.62447, 0.16967     0.14293    -0.97468,
     0.50901     0.50027    -0.32489, 0.84835     0.85760     0.32489,
     1.18769     1.21494     0.97468, 1.52703     1.57227     1.62447};
  y= { 10.37, 10.62, 10.71, 11.26, 10.70, 11.14, 12.46, 12.09, 15.23, 11.23, 11.35,
       12.34,11.59,11.38,11.71,14.25,12.64,12.27,15.29, 13.89};
i={1 0 0 , 0 1 0 , 0 0 1}; k=0.2;  z1=t(z)*z; z2=inv(z1+k*i);  b1=z2*t(z)*y;
ssr=t(b1)*t(z)*y; sst=41.24648;  r_2=ssr/sst; print b1 r_2 ;  run;
```

$k = 0.2$ için ridge tahmin edicilerini bulmak için SAS kodları

k	$\tilde{\beta}_1(k)$	$\tilde{\beta}_2(k)$	$\tilde{\beta}_3(k)$	$R^2(k)$
$k = 0.00$	1.6444648	-0.7867600	0.2384082	0.4159643
0.02	1.3231211	-0.4599170	0.2195652	0.4147084
0.04	1.1368910	-0.2706830	0.2086365	0.4138964
0.05	1.0703313	-0.2031150	0.2047276	0.4135777
0.06	1.0153209	-0.1473100	0.2014952	0.4132970
0.08	0.9296870	-0.0605400	0.1964588	0.4128158
0.10	0.8660845	0.0037886	0.1927128	0.4120462
0.12	0.8169626	0.0533683	0.1898150	0.4120462
0.14	0.7778675	0.0927354	0.1875044	0.4117184
0.16	0.7460019	0.1247392	0.1856171	0.4114145
0.18	0.7195204	0.1512593	0.1840451	0.4111285
0.20	**0.6971567**	**0.1735855**	**0.1827143**	**0.4108563**
0.30	0.6227516	0.2471155	0.1782550	0.4096207

Parametrelerin ridge tahmin edicileri ile $R^2(k)$ değerleri yukarıdadır. Tablodan da görüldüğü gibi, k arttıkça $R^2(k)$ değerleri azalır. Diğer taraftan, ridge tahminleri ile standart en küçük kareler tahminleri arasında, $\tilde{\beta}(k) = (Z'Z + k\,I)^{-1} Z'\,Z\hat{\beta}$ şeklinde bir ilişki vardır⊕