S Ü Fen Ed Fak Fen Derg Sayı 31 (2008) 97-107, KONYA

# Doğrusal Regresyon Çözümlemesinde Çoklu Bağlantı Probleminin Sapan Değer İçeren Küçük Örneklemlerde İncelenmesi

Özgül VUPA*, Özlem GÜRÜNLÜ ALMA

Dokuz Eylül Üniversitesi (DEÜ), Fen-Edebiyat Fakültesi, İstatistik Bölümü, Buca, İzmir

**Özet:** Ridge Regresyon (RR) ve Temel Bileşenler Regresyon (TBR) Çözümlemesi, çoklu doğrusal bağlantı (ÇDB) probleminin varlığında kullanılan regresyon çözümlemesindeki yanlı kestirim yöntemlerindendir. Çoklu bağlantı probleminin varlığı, regresyon katsayılarının En Küçük Kareler (EKK) kestiriminde önemli etkilere sahiptir. Bu problemin en önemli etkisi, regresyon katsayılarının en küçük kareler tahminlerinin büyük varyansa sahip olmalarına neden olmasıdır. Ayrıca veri setinin sapan değer içermesi durumunda bu veri setinin yapısı ve çözüm yaklaşımıda değişmektedir. Bu çalışmada çoklu doğrusal bağlantı probleminin sapan değer içeren küçük örneklemli veri kümelerinde ne zaman ve nasıl oluştuğunun, sonuçlarının ne olduğunun ve bu çoklu doğrusal bağlantı probleminin çözümünün hangi yöntemlerle giderileceğinin tartışılması amaçlanmıştır.

**Anahtar Kelimeler:** Çoklu Bağlantı Problemi, En Küçük Kareler Yöntemi, Ridge Regresyon Çözümlemesi, Sapan Değer, Temel Bileşenler RegresyonÇözümlemesi

# Investigation Of Multıcollinearity Problem In Small Samples Included Outlier Value In Linear Regression Analysis

**Abstract:** Ridge Regression and Principal Components Regression Analysis are biased estimation techniques and these techniques are used in the presence the problem of multicolinearity in regression analysis. The presence of multicolinearity has a number of potentially serious effects on the least squares estimates of regression coefficients. The most important effect of this problem is that it causes high variances in the estimation of regression coefficients. The purpose of this study is to determine how and when the problem of multicollinearity happens in the small samples that include the specific outlier ratios, what the results of this problem are and which methods are used to eliminate the solution of this multicollinearity problem.

**Keywords:** Multicollinearity Problem, Least Square Method, Ridge Regression, Outlier, Principal Component Regression

* E-mail: ozgul.vupa@deu.edu.tr

## 1. Giriş

ÇDB (multicollinearity), X tasarım matrisinin kolonlarının doğrusal bağımlılığı şeklinde tanımlanabilir. Bağımsız değişkenler arasında ÇDB problemi; verilerin hatalı olmasından, yanlış veri toplama yönteminden, kitle veya modelde yapılan kısıtlamalardan, modelin tanımlanmasında ve model seçiminde yapılan hatalardan kaynaklanabilir. Veri setindeki bu problemler, regresyon katsayılarına ait En Küçük Kareler (EKK) tahmincilerinin varyans ve kovaryans değerlerinin büyük olmasına, buna bağlı olarak da regresyon modeline dayanan yorumların hatalı olmasına neden olur. Ancak, katsayılara ait varyans ve kovaryans değerlerinin büyük oluşunun tek nedeni de veri setinde ÇDB probleminin olması değildir. Bağımsız değişkenler arasında ÇDB probleminin olup olmadığının belirlenmesi için birçok yöntem önerilmiştir.

İncelenecek olan veri setlerinde ÇDB probleminin haricinde, veri setleri her zaman geçerli veriler içermeyebilir. Gerçek hayatta; verinin elde edilişi sırasında, veri üzerinde geçekleştirilecek olan işlemlerde; verinin yanlış girilmesi, kopyalanması, dönüştürülmesi gibi, verinin elde ediliş yönteminde veya verinin ölçümü sırasında yapılan yanlışlıklar nedeniyle veri setleri hatalı olabilir. Bu durumlarda veri kümesinde verilerden bir kısmının diğerlerinden çok farklı özelliğe sahip olduğu görülür. Literatürde bu tip veriler sapan değer olarak ifade edilmektedir. Çözümleme için elde edilmiş bir veri setinde bazı verilerin dağılımının diğerlerinin sahip olduğu dağılımdan farklı bir dağılıma sahip olması durumu örnek olarak gösterilebilir.

Bağımsız değişkenler arasında tespit edilen ÇDB probleminin çözümü için modeli yeniden tanımlamak, konu ile ilgili ek veriler toplamak ve EKK yöntemi yerine Ridge Regresyon (RR) veya Temel Bileşenler Regresyon (TBR) gibi başka yöntemlerin uygulanması literatürde birçok çalışmada önerilmektedir. RR çözümlemesinde, EKK tahminlerine küçük bir yanlılık sabiti eklenerek varyansların küçültülmesiyle gerçekleştirilir. Böylece daha anlamlı sonuçlar elde edilir. Yanlı tahmin yöntemlerinden biri olan RR'nun doğrusallık, varyans homojenliği ve bağımsızlık gibi varsayımları EKK'in varsayımlarıyla aynıdır. Ancak yanlı tahmin yöntemlerinde güven aralıkları hesaplanmadığından normallik varsayımı yapılmamaktadır (Krishnan, 2008). TBR çözümlemesinde ise birbirinden bağımsız bileşenler türetilir.

Bu çalışmada çoklu doğrusal bağlantının veri kümesinde ne zaman ve nasıl oluştuğunun, sonuçlarının ne olduğunun ve bu ÇDB probleminin çözümünün hangi yöntemlerle giderileceğinin tartışılması amaçlanmıştır. Ayrıca ÇDB problemi içeren küçük örneklemli veri setinin, belirli güven düzeylerindeve belirli oranlarda sapan değer içermesi durumları da Minitab 11.0 istatistiksel paket programında hesaplanıp EKK, RR ve TBR ile hesaplanmış regresyon modellerine ait parametrelerin karşılaştırılması amaçlanmıştır.

## 2. Materyal ve Method

Bu çalışmada ÇDB probleminin saptanması, sonuçları ve çözüm yaklaşımları için simülasyon uygulaması Minitab 11.0 ve NCSS paket programlarında hazırlanmış ve EKK, RR TBR çözümlemeleri ile sonuçlandırılmıştır.

### 2.1 Çoklu Doğrusal Bağlantı Probleminin Saptanması

Bir veri setinde çoklu doğrusal bağlantı probleminin saptanmasında kullanılan birçok yöntem vardır (Gujarati, 1995). Bu yöntemlerden ilki basit korelasyon matrisinin incelenmesidir. İki bağımsız değişken arasındaki basit korelasyon katsayısı oldukça anlamlı ($r > \%75$) ise, bu durum ÇDB problemine yol açabilir. Bu katsayının büyüklüğüne rağmen, istatistiksel olarak bulunan anlamlı korelasyonların her zaman ÇDB problemine yol açmadığı göz ardı edilmemelidir. Ayrıca modele yeni bağımsız değişkenler eklendiğinde, $R^2$'deki değişimlerin incelenmesi ile ÇDB saptanabilir (Krishnan, 2008). $R^2$'de önemli bir gelişme sağlanamazsa, ÇDB problemi ortaya çıkmış olabilir. Kısmi korelasyon katsayılarının incelenmesi ile de ÇDB saptanabilir. İki değişken arasındaki basit korelasyon katsayısı anlamlı, fakat kısmi korelasyon katsayısı anlamsız ise bu durum ÇDB problemi için bir işaret olabilir. Ama şu da unutulmamalıdır ki kısmi korelasyon katsayılarının incelenmesi yaklaşımı her zaman etkili

olmamaktadır. Başka bir deyişle, kısmi korelasyon katsayılarının yüksek olması durumunda bile ÇDB problemi olabilmektedir. Literatürde çok fazla kullanılan varyans artırıcı faktörün (VIF = Variance Inflation Factor) kullanılması ile ÇDB saptanabilir (Roso, 2005). Değişkenler arasında ilişki yoksa ($R^2$ = 0 olacağından) VIF 1'e eşittir. Bağımlı ve bağımsız değişkenler arasında tam bir ilişki varsa ($R^2$ = 1 olacağından) VIF sonsuz olacaktır. $R^2$, % 90 ise, VIF 10 (1/(1-0.9) = 10) olarak elde edilir (Albayrak, 2007). Birkes ve Dodge'e (1993) göre VIF değeri 10'a eşit veya daha büyükse anlamlı ÇDB problemi söz konusudur. ÇDB probleminin saptanmasında kullanılan bir diğer yöntem, bağımsız değişkenler için tolerans değerinin (TV = Tolerance Value = 1- $R^2$) hesaplanmasıdır. Burada daha küçük tolerans değeri, daha büyük VIF değeri demektir. Kullanılan yöntemlerden biri de, yardımcı regresyon eşitliklerinden yararlanarak F değerlerinin hesaplanmasıdır. Çoklu doğrusal bağlantının saptanmasında kullanılan son yöntem ise, koşul sayısı (CN = Condition Number = [Maksimum Özdeğer / Minimum Özdeğer]) veya koşul endeksi (CI = Condition Index = [Maksimum Özdeğer / Minimum Özdeğer]$^{1/2}$) değerlerinin bulunmasıdır (Roso, 2005). CI değeri 10 ile 30 (CN değeri 100 ile 1000) arasında ise orta ve 30'dan (CI değeri 1000'den) büyükse çok güçlü ÇDB problemi olduğunu gösterir (Gujarati, 1995).

ÇDB probleminin saptanmasında kullanılan ve yukarıda açıklanan yaklaşımlardan her biri belirli dezavantajlara sahiptir. Ayrıca, hangi durumda hangi yaklaşımın kullanılabileceği konusunda da bir öneride bulunulamamaktadır (Albayrak, 2007).

### 2.2 Çoklu Doğrusal Bağlantı Probleminin Sonuçları

Regresyon çözümlemesinde ÇDB problemi bazı sonuçlara yol açar (Orhunbilge, 2000). ÇDB durumunda regresyon katsayıları anlamsız ve bu katsayıların standart hataları da büyük olmaktadır. Bu da ÇDB halinde regresyon katsayılarının varyans ve kovaryanslarını artırmaktadır. Modelin belirtme katsayısı $R^2$ değeri yüksek, ancak bağımsız değişkenlerden hiçbiri veya çok azı t testine göre anlamlı olmaktadır. Bir başka deyişle katsayılar önemli ama model geçersiz de denilebilir. İlgili bağımsız değişkenlerin bağımlı değişkenle olan ilişkilerinin yönü, hipotezlerdeki beklentilerle çelişmektedir. Bağımsız değişkenler birbiriyle bağlantılı ise, bunlardan bazılarının modelden çıkartılması gerekebilir. Fakat asıl sorun hangi değişken(ler)in çıkartılacağıdır. Modelden yanlış bir değişkenin çıkartılması, modelin hatalı tanımlanmasına yol açmaktadır. Bu yüzden değişkenlerin seçilmesi ve buna bağlı olarak modelin kurulması regresyon çözümlemesinde önemli bir yer tutar.

### 2.3 Çoklu Doğrusal Bağlantı Probleminin Çözümü

ÇDB problemi bir takım yöntemler kullanılarak çözülebilmektedir (Orhunbilge, 2000). Örneğin bir veya daha çok bağımsız değişken modelden çıkartılabilir. Fakat asıl sorun "hangi değişken(ler)in modelden çıkartılacağıdır". Dikkatli olunmazsa böyle bir yaklaşım, modelin yanlış kurulmasına neden olabilir. İkinci bir çözüm yaklaşımı olarak farklar alınarak değişkenler dönüştürülebilir. Fakat böyle bir dönüşüm hatalar arasında otokorelasyon problemine yol açabilir. Ayrıca bazen yeni gözlem değerlerinin elde edilmesiyle ÇDB problemi ortadan kaldırılabilir. Fakat her zaman örneklem büyüklüğünü artırmak mümkün olmayabilir. ÇDB probleminin diğer bir çözümü ise birbiriyle ilişkili olan iki değişken yerine bu iki değişkenin toplamının (tek bir değişken olarak) alınmasıdır. ÇDB probleminin son çözümü olarak da EKK yönteminin düzeltilmiş şekli olan ve yanlı standartlaştırılmış regresyon katsayılarını tahmin eden "Ridge Regresyon" veya birbirinden bağımsız bileşenler türeten "Temel Bileşenler Regresyon Çözümlemesi" yöntemlerinin kullanılmasıdır.

#### 2.3.1 Ridge Regresyon Yöntemi

EKK yöntemi ile yapılan tahminler bu yöntemin varsayımlarını sağlanması durumunda yansız olmaktadır. ÇDB halinde ise, regresyon katsayılarının varyans ve kovaryansları artar. Bu durumunda herhangi bir bağımsız değişken modelden çıkartıldığında veya modele eklendiğinde kısmi regresyon katsayılarında anlamlı değişimler olur. Ayrıca ÇDB halinde kısmi regresyon katsayılarının işaretleri gözlenenden veya beklenenden farklı da olabilir. Yani çoklu doğrusal bağlantılı verilerde hesaplanan standartlaştırılmış regresyon katsayıları kararlılığını kaybeder

(Faden, 1978). RR yöntemi, bu tahminlere küçük bir yanlılık sabiti ekleyerek varyansı azaltmaya yardım eder (Hoerl & Kennard, 1970, Roso, 2005). Genelde, varyans-kovaryans matrisinin köşegen değerlerine küçük bir yanlılık sabiti (k) ilave etmenin dışında, RR ile EKK yöntemlerinin işleyişi aynıdır. Bir başka deyişle RR ile bir taraftan tahminlerin varyansı azaltılmakta, diğer taraftan ise bu katsayı (k) oranında yanlı tahminler elde etmektedir. Böylece yansız tahminlerle yüksek varyans veya yanlı tahminlerle düşük varyans gibi iki durum söz konusu olur.

RR yöntemi X tasarım matrisinin kolonları arasında doğrusal bir bağımlılık ve $X'X$ matrisinin singüler olması durumunda kullanılır. $X'X$ matrisi singüler olması tersi olması anlamındadır. RR yöntemde ilk yapılacak işlem bağımsız değişkenlerin standartlaştırılmasıdır. Standartlaştırma işlemi bağımsız değişkenlerin kendi ortalamalarından farklarının alınıp kendi standart sapmalarına bölünmesi ile sağlanır. Standartlaştırılmamış değişkenlerin bulunduğu model yani EKK regresyon modeli $y_i = \beta_0 + \beta_1 x_{1i} + ... + \beta_p x_{pi} + \varepsilon_i$ şeklindedir ( $i = 1,2,...,n$ ). Burada Y bağımlı değişken, $X_1, X_2, ..., X_p$ bağımsız değişkenler, $\beta_1, \beta_2, ..., \beta_p$ parametreler ve $\varepsilon_i$ gözlenemeyen hata terimleridir. Matris notasyonu ile EKK regresyon modeli $\underline{Y} = X\underline{\beta} + \underline{e}$ ve regresyon katsayıları $\hat{\underline{\beta}} = (X'X)^{-1}X'\underline{Y}$ ile gösterilir. Eğer standartlaştırılmamış regresyon modeli standartlaştırılırsa aşağıdaki eşitlik elde edilir.

$$Y_i = \mu + \gamma_1 z_{1i} + \gamma_2 z_{2i} + ... + \gamma_p z_{pi} + \varepsilon_i \qquad (1)$$

Matris notasyonu ile standartlaştırılmış bu model ile gösterilir ve bu matris notasyonuna göre matrisler aşağıdaki gibi yazılabilir.

$$\underline{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} \quad \underline{1} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \quad Z = \begin{bmatrix} z_{11} & z_{12} & \cdots & z_{1p} \\ z_{21} & z_{21} & \cdots & z_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ z_{n1} & z_{n2} & \cdots & z_{np} \end{bmatrix} \quad \underline{e} = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix} \quad \underline{\gamma} = \begin{bmatrix} \gamma_1 \\ \gamma_2 \\ \vdots \\ \gamma_n \end{bmatrix} \qquad (2)$$

$\underline{Y} = \underline{1}\mu + Z\underline{\gamma} + \underline{e}$ ile gösterilen matris notasyonundaki $\mu$ ve $\underline{\gamma}$ parametrelerinin EKK tahmin edicileri sırasıyla $\hat{\mu}_{EKK} = \overline{Y}$ ve $\hat{\underline{\gamma}}_{EKK} = (Z'Z)^{-1}Z'\underline{Y}$ ile gösterilebilir. $Z'Z$ matrisinin singüler olması durumunda ise $\mu$ ve $\underline{\gamma}$ parametrelerinin Ridge tahmin edicileri de sırasıyla $\hat{\mu}_{RR} = \overline{Y}$ ve $\hat{\underline{\gamma}}_{RR} = (Z'Z + kI)^{-1}Z'\underline{Y}$ gibi yazılabilir. Bir başka deyişle RR analizinde korelasyon matrisinin köşegen değerlerine küçük bir yanlılık sabiti eklenerek, yanlı standartlaştırılmış regresyon katsayıları hesaplanmış olur. Burada k, Ridge parametresidir $k = p\dfrac{\hat{\sigma}^2_{EKK}}{\left\|\hat{\underline{\gamma}}_{EKK}\right\|^2}$ gibi hesaplanır.

Ridge parametresinin hesaplanmasındaki $\hat{\sigma}^2_{EKK}$, $\sigma^2$'nin EKK tahmin edicisi ve I, $p \times p$ boyutlu birim matrisidir (Birkes ve Dodge, 1993, Roso, 2005). Ayrıca $\left\|\hat{\underline{\gamma}}_{EKK}\right\| = \sqrt{\sum (\hat{\underline{\gamma}}_{EKK})^2_i}$ , i = 1, 2, ... , p şeklindedir. Ridge parametresi olan k, 1'den küçük pozitif sayısal bir değerdir (genelde $k \leq 0.3$). k değeri, 1'e yaklaştıkça tahminlerin yanlılığı artmakta, fakat varyansları azalmaktadır.

### 2.3.2 Temel Bileşenler Regresyon Çözümlemesi

Temel bileşenler (TB) çözümlemesi, orijinal p değişkenin varyans yapısını daha az sayıda ve bu değişkenlerin doğrusal bileşenleri olan yeni değişkenlerle ifade etme yöntemidir. Aralarında korelasyon bulunan p sayıda değişkenin açıkladığı yapıyı, aralarında korelasyon bulunmayan ve sayıca orijinal değişken sayısından daha az sayıda ($p > k$) orijinal değişkenlerin doğrusal bileşenleri olan değişkenlerle ifade etme yöntemine temel bileşenler çözümlemesi denir. Temel bileşenler çözümlemesinin amacı veri indirgemesi ve tahminlemesini yapmaktır (Foong, 2007).

Temel bileşenler çözümlemesinde değişkenler standartlaştırıldığından 7. eşitlikteki $(X'X)$ ifadesi R'e eşit olur. Burada R, bağımsız değişkenler için korelasyon matrisini göstermektedir. TB çözümlemesini gerçekleştirmek için bağımsız değişkenler temel bileşenlere dönüştürülür. Bu matematiksel olarak $(X'X) = PDP' = Z'Z$ şeklinde yazılır. Burada TB modelini tanımlayan D, $X'X$ özdeğerlerinin köşegen matrisini; P, $X'X$ özvektör matrisini ve Z, veri matrisini göstermektedir. Burada P ortogonal olduğundan $P'P = I$ 'dır (Foong, 2007). Böylece orijinal değişkenlerin (X) ağırlıklı ortalamalarını ifade eden yeni değişkenler (Z) türetilmektedir. Bu durum, regresyon çözümlemesine başlamadan önce değişkenlerin logaritmik veya karekök dönüşümlerinin alınmasından farklı bir şey değildir. Bu yeni değişkenler temel bileşen olduğu için, bu bileşenler arasındaki korelasyonlar sıfırdır. $X_1$, $X_2$, $X_3$ gibi değişkenlerle çözümlemeye başlanmışsa, $Z_1$, $Z_2$ ve $Z_3$ gibi dönüştürülmüş değişkenler elde edilmektedir. Çok küçük özdeğerler hesaplanması durumunda çok güçlü çoklu doğrusal bağlantı problemi ortaya çıkar. Bu sorunun üstesinden gelebilmek için düşük özdeğerlere denk bileşenler çözümlemeden çıkartılmaktadır. Böylece Y değişkeni bağımlı ve $Z_1$ ve $Z_2$ bileşenleri ise bağımsız değişkenler olarak alınan modelde artık çoklu doğrusal bağlantı söz konusu olmamaktadır. Daha sonra sonuçlar X ölçeğine geri dönüştürülerek B'nin tahminleri elde edilir. Bu tahminler yanlı olacaktır. Fakat bu yanlılığın büyüklüğü varyansın azaltılmasıyla dengeleneceği göz önüne alınmalıdır. Başka bir deyişle TB tahminlerinin hata kareleri ortalamasının EKK tahminlerinden daha küçük olması beklenmektedir (Albayrak, 2007).

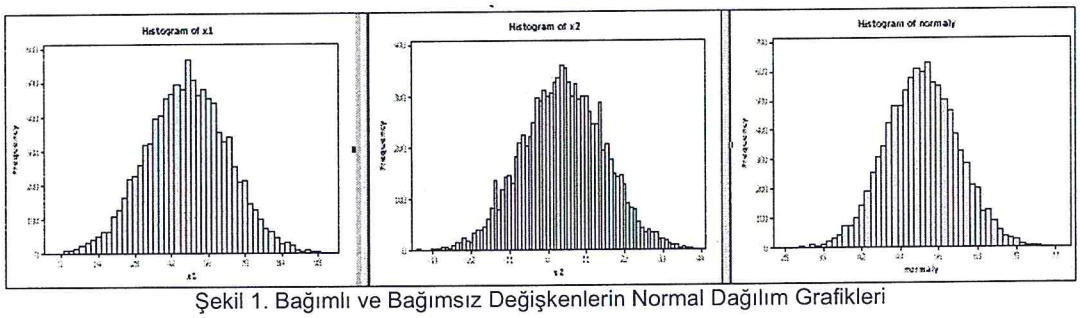
Matematiksel olarak ifade edildiğinde regresyon katsayılarının tahmini $\underline{\hat{A}} = (Z'Z)^{-1}Z'\underline{Y} = D^{-1}Z'\underline{Y}$ şeklinde hesaplanır (Foong, 2007). Bu eşitlik ile EKK regresyonu elde edilir. Böylece $\hat{A}$ ve $\hat{B}$ gibi iki regresyon katsayıları seti arasındaki ilişkiler ise $\underline{\hat{A}} = P\underline{\hat{B}}$ ve $\underline{\hat{B}} = P\underline{\hat{A}}$ yazılmaktadır. A'nın ilgili elementi sıfıra eşitlenerek ilgili TB çözümlemesinden çıkartılabilir. Son aşamada ise $\underline{\hat{B}} = P\underline{\hat{A}}$ eşitliği kullanılarak hesaplanan katsayılar orijinal ölçeğine dönüştürülmektedir.

Son olarak, RR yönteminde k yanlılık sabitinin seçiminde yaşanan belirsizliğin aksine, TB çözümlemesinde çıkarılacak olan temel bileşenlerin sayısı göreceli olarak daha kesindir.

### 2.4 Simülasyon Verisinin Tasarımı

Kitleye ait regresyon çözümlemesinde kullanılacak model denklemi $Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon_i$ biçimindedir. Bu model denkleminde yer alan bağımlı ve bağımsız değişkenler ile hata terimleri Minitab 11.0 paket programı kullanılarak simülasyon çalışmasıyla türetilmiştir. Kitle için oluşturulan bu bağımsız değişkenler $X_1 \sim N(47.87,15.53^2)$ ve $X_2 \sim N(4.49,11.71^2)$ parametreleri ile normal dağılıma sahiptir. Aynı zamanda hata terimleri de $\varepsilon_i \sim N(0,1)$ standart normal dağılıma sahip olacak şekilde oluşturulmuştur. Tüm bağımsız değişkenler ve hata terimleri Minitab paket programında 10000 adet üretilmiş olup yapılan çözümlemeler için gerekli örneklemler bu verilerden rasgele olacak şekilde seçilmiştir.

Türetilen bağımsız değişkenlerle ÇDR modeline ait regresyon katsayıları $\beta_0 = 0$, $\beta_1 = 0,982$ ve $\beta_2 = 1,28$ olacak şekilde belirlenmiş ve model denklemi de $Y \sim N(52.92,6.5)$ normal dağılımlı olup $Y_{normal} = 0,982X_1 + 1,28X_2$ regresyon modeliyle ifade edilmiştir. Türetilen bağımlı ve bağımsız değişkenlerin normal dağılımdan geldiğini gösteren grafikler Şekil 1' de verilmiştir.




Şekil 1. Bağımlı ve Bağımsız Değişkenlerin Normal Dağılım Grafikleri

Çoklu doğrusal bağlantının olabilmesi için bağımsız değişkenler arasında önemli derecede güçlü bir ilişkinin olması gerekmektedir. Türetilen bağımsız değişkenler arasındaki bu güçlü ilişkinin derecesini ve yönünü gösteren ilişki (pearson korelasyon) matrisi Tablo 1'de verilmektedir.

Tablo 1. Bağımsız Değişkenler Arasındaki İlişki Katsayı Değerleri

| | Normal Y | $x_1$ |
|---|---|---|
| $x_1$ | 0.009 | |
| p değeri | 0.387 | |
| $x_2$ | -0.011 | -0.980 |
| p değeri | 0.279 | 0.000 |

Tablo 1 incelendiğinde bağımsız değişkenler arasında güçlü bir ilişkinin olduğu (-0,98), bağımlı ve bağımsız değişkenler arasında ise anlamlı bir ilişkinin olmadığı görülmektedir. Ancak bu durum bağımsız değişkenlerin bağımlı değişkeni iyi bir şekilde açıklayamadığından değil ÇDB probleminin varlığından kaynaklanmaktadır. Kitleye ait oluşturulan ÇDR modeline ait belirtme katsayısının değeri $R^2 = 0{,}99$ 'dır. Aynı zamanda bu değerin çok yüksek olması, regresyon katsayılarının anlamlı ($p < 0.05$), modelin geçerli ($p < 0.05$) ve modele ait varyans değerinin ($s = 7.19$) çok küçük olduğu saptanmıştır. Ayrıca bu modelin VIF değeride 16.6 olarak bulunmuştur. Böylece kitle için türetilen verilerin ÇDB çözümlemesi için kullanılması uygun görülmüştür.

ÇDR modelinde bağımlı değişkenin sapan değer içermesi durumunun araştırıldığı bu çalışmada, bağımsız değişkenler arasındaki ilişkinin derecesinin sabit kalacak şekilde bağımlı Y değişkeni de belirli oranlarda sapan değer içerecek şekilde kirletilmiştir. Bu kirletme işlemi aşağıda belirtilen adımlarla gerçekleştirilmiştir.

**Adım 1:** Bağımlı değişkenin sapan değer içerdiği durumda ÇDB problemi incelendiği bu çalışmada örneklemin içerdiği sapan değer miktarı, çekilecek örneklem sayısı Rousseeuw ve Leroy'in 1987 yılında yapmış olduğu çalışma göz önünde bulundurularak belirlenmiştir. Rousseeuw ve Leroy (1987), sapan değer içermeyen en az bir alt kümenin büyük bir olasılıkla seçilebileceği bir yöntem önermişlerdir. Veri kümesinin kirlilik oranı $\varepsilon$ olarak gösterilirse, n/p oranı $1-[1-(1-\varepsilon)^n]^k$ ifadenin değerine çok yaklaşacaktır.

**Adım 2:** Adım 1 de verilen ifade yardımıyla kirlenme oranının $\varepsilon$ olduğu bir veriden n birimlik k tane alt kümeler çektiğimizde bunlardan en az birinin sapan değer içermeyen gözlemlerden oluşma olasılığı hesaplanmıştır. Örneğin kirlenme oranının $\varepsilon$ = %50 olduğu bir veriden 15 birimlik alt kümeler çektiğimizde bunlardan en az birinin sapan değer içermeyen gözlemlerden oluşma olasılığını 0,95 olması için çekmemiz gereken 15 birimlik alt kümelerin sayısı 98'dir.

**Adım 3:** Adım 1 ve Adım 2'deki bilgilere dayanarak çekilen kümelerin sapan değer içermeyen gözlemlerden oluşma olasılığı %95 ve %99 olarak belirlenmiştir.

**Adım 4:** Çekilen küçük örneklem genişliğinin 20 olmasına karar verilmiştir.

**Adım 5:** Çekilecek 20 birimlik örneklemlerin içerdiği sapan değer yüzdelerinin ε = %5 ve ε = %10 olması durumları incelenmiştir.

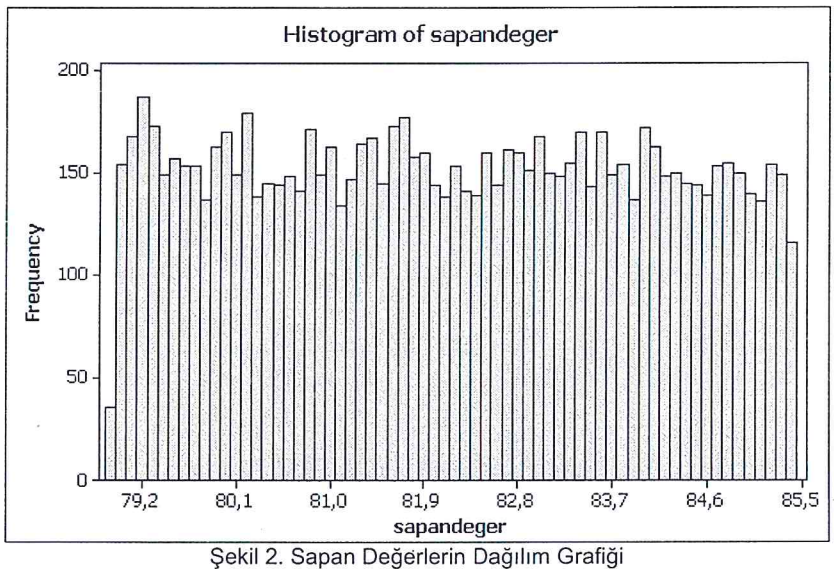
**Adım 6:** Rousseeuw'un belirtmiş olduğu 1 formulu için gerekli olan tüm parametreler belirlendikten sonra araştırma için üzerinde çalışılacak örneklem sayıları belirlenmiştir. Örneğin; örneklem büyüklüğü 20 birim olan bir veri kümesinin %5'lik sapan değer içerdiği durumda bir başka deyişle sapan değer miktarının 1 olduğu, %95 güven düzeyinde çekilecek örneklem sayısının $1-[1-(1-\varepsilon)^n]^k = 0{,}95$, $1-[1-(1-0{,}05)^{20}]^k = 0{,}95$ ifadeleri ile k'nın 7 olduğu hesaplanmıştır.

Yukarıda belirtilen adımlar gerçekleştirildiğinde ÇDR modelinin ÇDB problemi kirletilen bağımlı Y değişkeni üzerinden incelenmiştir. Yapılan tüm çözümlemelerde örneklem genişliği sabit olup 20'e eşittir. Buna bağlı olarak örneklem genişliğinin n = 20 olduğu farklı sapan değer yüzdeleri ve farklı güven düzeylerinde üzerinde çalışılacak örneklem sayıları belirlenmiş ve bunlar Tablo 2'de verilmiştir.

Tablo 2. %5 ve %10'luk Sapan Değer İçeren Örneklemlerden Birim Sayısına Göre Çekilecek Örneklem Sayısı ve Buna Bağlı Elde Edilen Sapan Değer Sayısı

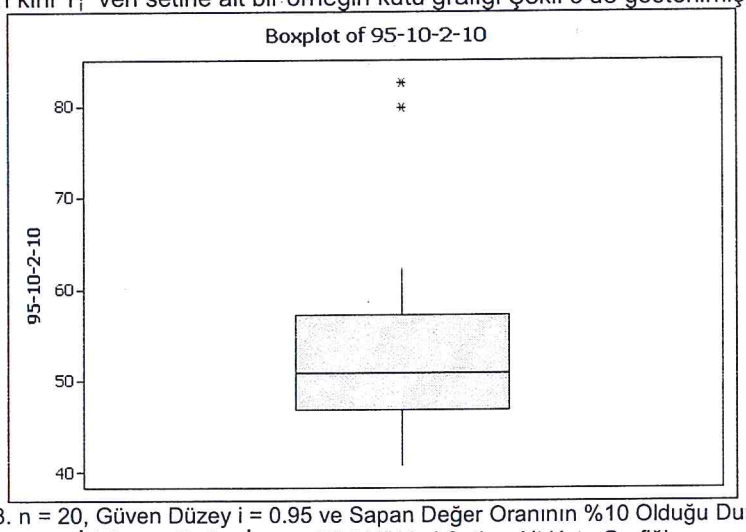
| Güven Düzeyi (1-α) | | | | | | | |
|---|---|---|---|---|---|---|---|
| %95 | | | | %99 | | | |
| Sapan Değer Yüzdesi | | | | Sapan Değer Yüzdesi | | | |
| %5 | | %10 | | %5 | | %10 | |
| Çekilecek Örneklem Sayısı | Sapan Değer Sayısı | Çekilecek Örneklem Sayısı | Sapan Değer Sayısı | Çekilecek Örneklem Sayısı | Sapan Değer Sayısı | Çekilecek Örneklem Sayısı | Sapan Değer Sayısı |
| 7 | 1 | 23 | 2 | 10 | 1 | 35 | 2 |

**Adım 7:** Üzerinde araştırma yapılacak örneklem sayılarının belirlenmesinden sonra her bir örneklemde yer alması gereken sapan değer sayısınca tekdüze dağılımdan gelen $Y_j \sim U(78.92,85.42)$ değerlerden yararlanarak "normal $Y_i$" veri setlerine dahil edilmesiyle sapan değer içeren toplam 75 adet "kirli $Y_i$" veri setleri oluşturulmuştur. Tekdüze dağılımdan gelen $Y_j$ değerlerinin grafiği Şekil 2'de verilmiştir.




Şekil 2. Sapan Değerlerin Dağılım Grafiği

n = 20, güven düzeyi = 0.95 ve sapan değer oranının %10 olduğu durumda iki sapan değer içeren "türetilen kirli $Y_i$" veri setine ait bir örneğin kutu grafiği Şekil 3'de gösterilmiştir.



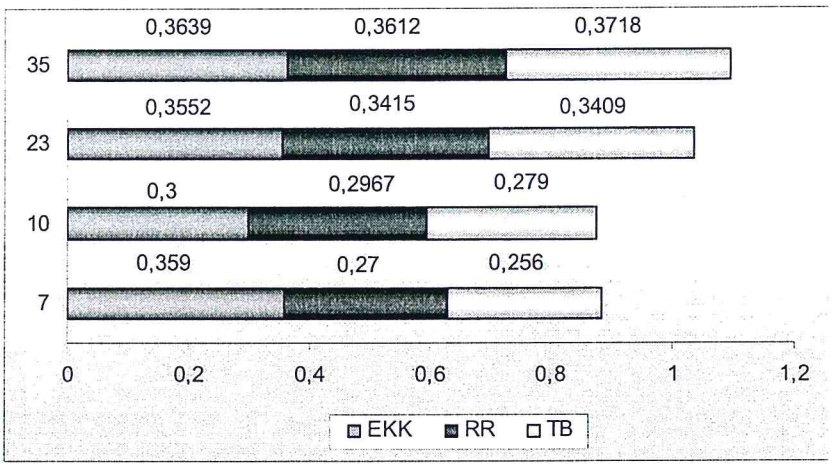

Şekil 3. n = 20, Güven Düzey i = 0.95 ve Sapan Değer Oranının %10 Olduğu Durumda İki Sapan Değer İçeren "Kirli $Y_i$" Veri Setine Ait Kutu Grafiği

## 4. Sonuç ve Tartışma

Çoklu doğrusal bağlantı probleminin varlığında kullanılan regresyon çözümlemesindeki yanlı kestirim yöntemlerinden RR ve TBR çözümlemlerinin bağımlı değişkenin sapan değer içermesi durumunda karşılaştırıldığı bu çalışmada elde edilen sonuçlar ve bu sonuçlara ilişkin yorumlar ilerleyen alt bölümlerde belirtilmiştir.

### 4.1 Simülasyon Çalışması ile Elde Edilen Sonuçlar

Simülasyon çalışması ile çözümlemelere ait MSE değerleri açısından incelenecek olursa; 20 birimlik 7, 10, 23 ve 35 adet çekilen örneklemler için yapılan regresyon modelleri çözümlemesinde en yüksek standart sapmalı MSE değerlerinin EKK'den elde edilirken en küçük standart sapmalı MSE değerleri de TBR'den elde edilmiştir. Bu verilerden de anlaşıldığı gibi TBR'nin sapan değer içeren ÇDB probleminin varlığında diğer çözümlemelere göre daha güçlü olduğu söylenebilir. Çözümlermelere ait standart sapmalı MSE değerleri Şekil 4'deki çubuk grafiğinde gösterilmiştir.




Şekil 4. EKK, RR ve TBR Çözümlemeleri İçin MSE Değerlerinin Standart Sapmalarının Karşılaştırılması

Yapılan simülasyon çalışmasıyla kirli $Y_i$'lerin değerleri kullanılarak EKK, RR ve TBR çözümlemelerinden elde edilmiş parametrelerin ve MSE değerlerinin ortalamaları ve standart sapmaları Tablo 3'de verilmiştir.

Tablo 3. EKK. RR ve TBR Çözümlemelerinden Elde Edilen Parametre Değerlerinin Ortalamaları

| | | $\hat{\beta}_0$ | | | $\hat{\beta}_1$ | | | $\hat{\beta}_2$ | | |
|---|---|---|---|---|---|---|---|---|---|---|
| #* | | **EKK** | **RR** | **TBR** | **EKK** | **RR** | **TBR** | **EKK** | **RR** | **TBR** |
| | **Ort** | 1.057 | 1.536 | -3.49 | 3.15 | 1.90 | 0.6602 | -3.49 | -1.12 | 1.259 |
| **7** | **Std** | **2.228** | 2.432 | 3.08 | 6.07 | 3.00 | **0.1973** | 12.01 | 6.16 | **0.450** |
| | **Ort** | 0.0350 | 0.0268 | 1.91 | 0.572 | 0.763 | 0.7902 | 1.91 | 1.535 | 1.441 |
| **10** | **Std** | 3.125 | 3.072 | **2.785** | 1.991 | 1.492 | **0.1876** | 3.96 | 2.929 | **0.332** |
| | **Ort** | 2.151 | 1.800 | -0.451 | 1.516 | 1.385 | 0.6582 | -0.451 | -0.0880 | 1.214 |
| **23** | **Std** | **4.009** | 4.023 | 4.144 | 2.863 | 2.219 | **0.2654** | 5.55 | 4.222 | **0.484** |
| | **Ort** | 2.351 | 2.193 | 2.00 | 0.271 | 0.631 | 0.6452 | 2.00 | 1.318 | 1.2436 |
| **35** | **Std** | 4.380 | **4.235** | 4.272 | 4.828 | 2.432 | **0.2657** | 9.77 | 5.090 | **0.5099** |

* 20 Birimli Çekilen Örneklem Sayısı

EKK, RR, TBR çözümlemeleri parametere ortalamaları ve standart sapmaları açısından incelendiğinde; $\hat{\beta}_0$ için, çekilen örneklem sayısı 7 ve 23 olduğu durumda en iyi parametre değerini tahmin eden çözümleme EKK yöntemi iken, çekilen örneklem sayısı 10 olduğunda TBR ve 35 olduğunda ise RR çözümlemesi olduğu görülmüştür. Tahmin edilen $\hat{\beta}_1$ ve $\hat{\beta}_2$ için çekilen tüm örneklemlerde TBR çözümlemesinin diğer çözümlemelere göre en küçük standart sapmalı değerlere sahip olduğu görülmüştür. Bu durumlar değerlendirildiğinde TBR çözümlemesinin parametre tahminleri için belirli oranlarda sapan değerler içeren küçük örneklemlerde daha güçlü ve daha iyi tahmin değerleri verdiği görülmüştür.

Bu çalışmada kirli $Y_i$'lerin değerleri kullanılarak EKK, RR ve TBR çözümlemelerinden elde edilen modelin MSE değerlerinin ortalamaları Tablo 4'de verilmiştir.

Tablo 4. EKK. RR ve TBR Çözümlemelerinden Elde Edilen Parametre Değerlerinin MSE'lerin Ortalamaları

| #* | EKK | RR | TBR |
|---|---|---|---|
| **7** | 2.068 | 2.136 | 2.2036 |
| **10** | 1.7360 | 1.7440 | 1.7736 |
| **23** | 2.2539 | 2.2427 | 2.2922 |
| **35** | 2.2112 | 2.2413 | 2.2948 |
| **Genel Ortalama** | 2.0672 | 2.091 | 2.1410 |

* Çekilen Örneklem Sayısı

Tablo 4'ün sonuçlarına göre sapan değer içeren tüm veri setleri için regresyon modellerinin EKK çözümlemesine ait MSE değerlerinin diğer çözümleme yöntemlerine göre daha küçük olduğu görülmüştür. Ancak tüm çözümlemelerin MSE değerlerinin birbirlerine çok yakın olduğu da görülmektedir.

ÇDB veri toplama yönteminin yanlış olmasından, kitle veya modelde yapılan kısıtlamalardan, modelin tanımlanmasında ve model seçiminde yapılan hatalardan kaynaklanır. ÇDB probleminin olduğu veri setinde regresyon katsayıları, bu katsayıların standart hataları ve işaretleri doğru olarak tahmin edilmez. Bir başka deyişle, regresyon katsayılarının varyans ve kovaryansları artığından bu değişkenlerin regresyon katsayılarının kısmi t testleri istatistiksel açıdan anlamsız hale gelir. ÇDB probleminin çözümü için modeli yeniden tanımlamak, konu ile ilgili ek veriler toplamak ve EKK yöntemi yerine RR veya TBR yöntemlerini kullanmak gereklidir.

RR, EKK yönteminin düzeltilmiş şeklidir ve yanlı standartlaştırılmış regresyon katsayılarını tahmin ederken TBR ise birbirinden bağımsız bileşenler türetir.

Bu çalışmada ÇDB probleminin yanında veri seti üzerinde bazı değişiklikler yapıldığında bu çözümleme yöntemlerinin sonuçlarının nasıl değiştiği incelenmiştir. Yapılan bu değişiklikler veri setinin,

- Küçük olması (n=20),
- Belirli oranlarda sapan değerler içermesi (% 5, %10),
- Belirli güven aralıklarında incelenmesi (%95, %99) ve
- Belirli güven aralığı ve sapan değer yüzdesinin belirlenmesinden sonra çalışılacak olan örneklem sayılarının belirlenmesi (7, 10, 23, 35)

şeklinde olmasıdır.

Bu kısıtlar altıda yapılan çözümlemelerin karşılaştırılması Tablo 5'de verilmiştir.

Tablo 5. Tüm Kısıtlar İçin Çözümlemelerin Karşılaştırılması

| | | Güven Düzeyi (1-α) | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | %95 | | | | %99 | | | |
| Sapan Değer Yüzdesi | Çekilen Örneklem Sayısı | Parametre Tahmini | | | | Parametre Tahmini | | | |
| | | $\hat{\beta}_0$ | $\hat{\beta}_1$ | $\hat{\beta}_2$ | MSE | $\hat{\beta}_0$ | $\hat{\beta}_1$ | $\hat{\beta}_2$ | MSE |
| 5 | 7 | EKK | TBR | TBR | EKK-TBR-RR | - | - | - | |
| | 10 | - | - | - | | TBR | TBR | TBR | EKK-TBR-RR |
| 10 | 23 | EKK | TBR | TBR | EKK-TBR-RR | - | - | - | |
| | 35 | - | - | - | | RR | TBR | TBR | EKK-TBR-RR |

### 4.2. Tartışma

EKK, RR, TBR çözümlemeleri parametere ortalamaları ve standart sapmaları açısından incelendiğinde; $\hat{\beta}_0$ için güven düzeyi %95 olduğunda en iyi parametre değerini tahmin eden çözümleme EKK yöntemi iken, güven düzeyi artırıldığında TBR ve RR çözümlemesinin parametre değerinin daha iyi olduğu görülmüştür. Tahmin edilen $\hat{\beta}_1$ ve $\hat{\beta}_2$ için çekilen tüm örneklemlerde TBR çözümlemesinin diğer çözümlemelere göre en küçük standart sapmalı değerlere sahip olduğu sonucu elde edilmiştir. Bu durumlar değerlendirildiğinde TBR çözümlemesinin parametre tahminleri için belirli oranlarda sapan değerler içeren küçük örneklemlerde daha güçlü ve daha iyi tahmin değerleri verdiği söylenebilir.

Sapan değer içeren tüm veri setleri için regresyon modelleri MSE değerleri açısından incelendiğinde EKK çözümlemesinden elde edilen değerlerin çok az farkla da olsa diğer çözümleme yöntemlerine göre küçük olduğu görülmüştür. Ancak yapılan bu çalışmada ÇDB probleminin sapan değer içeren küçük veri setlerinde TBR çözümlemesinin diğer çözümlemelere göre daha güçlü parametre tahmin değerlerine sahip olması ve MSE değerlerinin yaklaşık olarak EKK çözümlemesi değerlerine benzemesinden dolayı, TBR çözümlemeisnin ÇDB probleminin sapan değer içeren küçük veri setlerinde kullanılması önerilmektedir.

## 5. Kaynaklar

1. Albayrak A.S. **Çoklu Doğrusal Bağlantı Halinde En Küçük Kareler Tekniğinin Alternatifi Yanlı Tahmin Teknikleri Ve Bir Uygulama**. Zonguldak Karaelmas Üniversitesi. Çaycuma İktisadi ve İdari Bilimler Fakültesi Sayısal Yöntemler Anabilim Dalı Öğretim Üyesi. İnternet Üzerinden Son Erişim Tarihi: 4 Temmuz 2007.
2. Barnett V. and Lewis T. **Outliers in Statistical Data.** 3th edt., John Wiley & Sons, Canada, (1994).
3. Birkes ve Dodge. **Alternative Methods of Regression**. 3th edt. John Wiley & Sons. Canada, (1993).
4. Dempster. A. P.. M. Schatzoff. and N. Wermuth. **A Simulation Study of Alternatives to Ordinary Least Square**. Journal of American Statistical Association. V: 72. 77-91, (1977).

5. Faden. V. B. **Shrinkage in Regression and Ordinary Least Squares Multiple Regression Estimators**. Yayınlanmamış Doktora Tezi. University of Maryland, (1978).
6. Foong N. et all. **An Overview of Biased Estimators.** Journal of Physical Science. V: 18(2), 89–106, (2007).
7. Fox J. **Applied Regression Analysis, Linear Models and Related Methods**. 3th edt., Sage Publication, USA, (1997).
8. Gujarati. D. N. **Basic Econometrics**. 3th Edition. McGraw-Hill. New York, (1995).
9. Hoerl. A. E. and Kennard R.W. **Ridge Regression: Biased Estimation for Nonorthogonal Problems**. Technometrics. V: 12. 69-82, (1970).
10. Hoaglin D.&Tukey J. **Understanding Robust and Exploratory Data Analysis.** John Wiley & Sons, Canada, (1983).
11. Karadavut U.. Genç A. ve ark.. **Nohut Bitkisinde Verime Etki Eden Bazı Karakterlerin Alternatif Regresyon Yöntemleriyle Karşılaştırılması**. Tarım Bilimleri Dergisi. 11 (3). 328-333, (2005).
12. Kleinbaum. Kupper. Muller. and Nizam. **Applied Regression Analysis and Other Multivariate Methods**. 3th edt. Duxbury. USA. (1998).
13. Krishnan T. **Regression Diagnostics**. Lecture Notes. (2008)
14. Kurt S. **Lecture Notes on Regression Analysis**. DEU-Department of Statistics. Turkey. (2000).
15. Marguard D.&Snee D. **Ridge Regression in Practice**. American Statisticion. V: 29 (1), (1975).
16. Neter. J.. W. Wasserman and M. Kunter. **Applied Linear Statistical Models**. 3th Edition. New Jersey. (1990).
17. Orhunbilge. N.. **Uygulamalı Regresyon ve Korelasyon Analizi**. İstanbul, Türkiye, (2000).
18. Roso V.N. et all. **Estimation of Genetic Effects in The Presence of Multicollinearity in Multibreed Beef Cattle Evaluation.** American Society of Animal Science. V:83:1788–1800. (2005).
19. Rousseeuw P. and Leroy A. **Robust Regression and Outlier Detection**, John Wiley & Sons, Canada, (1987).
20. Stromberg Arnold J. **Computation of High Breakdown Nonliner Regression Parameters**, Journal of the American Statistical association, V. 88, No:421, (1993).
21. Price. B. **Ridge Regression: Application to Nonexperimental Data**. Psychological Bulletin. V: 84. 759-766, (1979).