

**T.C.
İNÖNÜ ÜNİVERSİTESİ
SOSYAL BİLİMLER ENSTİTÜSÜ**



**RİDGE REGRESYON PARAMETRE SEÇİMİ:
TÜRKİYE’NİN DOĞRUDAN YABANCI
YATIRIM ÖRNEĞİ**

YÜKSEK LİSANS TEZİ

**Danışman
Doç. Dr. Bahadır YÜZBAŞI**

**Hazırlayan
Mustafa PALA**

MALATYA 2022

T.C.
İNÖNÜ ÜNİVERSİTESİ
SOSYAL BİLİMLER ENSTİTÜSÜ

RİDGE REGRESYON PARAMETRE SEÇİMİ: TÜRKİYE’NİN DOĞRUDAN
YABANCI YATIRIM ÖRNEĞİ

YÜKSEK LİSANS TEZİ

Hazırlayan
Mustafa PALA

Danışman
Doç. Dr. Bahadır YÜZBAŞI

MALATYA – 2022

ONUR SÖZÜ

Doç. Dr. Bahadır YÜZBAŞI'nın danışmanlığında yüksek lisans tezi olarak hazırladığım **“RİDGE REGRESYON PARAMETRE SEÇİMİ: TÜRKİYE’NİN DOĞRUDAN YABANCI YATIRIM ÖRNEĞİ”** başlıklı bu çalışmanın, bilimsel ahlak ve geleneklere aykırı düşecek bir yardıma başvurmaksızın tarafımdan yazıldığını ve yararlandığım bütün yapıtların hem metin içinde hem de kaynakçada yöntemine uygun biçimde gösterilenlerden oluştuğunu belirtir, bunu onurumla doğrularım.

Mustafa PALA

İmza

ÖNSÖZ

Tez çalışmam süresince bilgi ve değerli fikirlerinden istifade ettiğim, hiçbir zaman desteğini ve kıymetli zamanlarını esirgemeyen ayrıca bu tezin oluşmasında büyük rol oynayan değerli tez danışmanım Doç. Dr. Bahadır YÜZBAŞI'ya, maddi ve manevi olarak desteklerini esirgemeyen dostlarıma ve kıymetli Pala ailesine, sonsuz teşekkür eder saygılarımı sunarım.

Bu tez çalışmamı yakın zamanda kaybettiğim rahmetli annem Meryem PALA'ya ithâf ediyorum.



ÖZET

Çoklu doğrusal regresyon modelindeki bir veya daha fazla bağımsız değişkenin aralarında doğrusal bir ilişki olması durumunda ortaya çıkan probleme çoklu doğrusal bağlantı problemi denir. Çoklu doğrusal bağlantı durumunda en küçük kareler tahmincilerinin varyansı yüksek ve parametre tahminleri yanlış hesaplanmaktadır. Söz konusu durumda en küçük kareler tahmincilerinin sonuçları yorumlamada yetersiz kalması, en küçük kareler yöntemini güvenilir hale getirir. Çoklu doğrusal bağlantı problemini çözmek ve model parametrelerini tahmin etmek için en küçük kareler tahmincisine alternatif olarak yanlı regresyon tahmincileri kullanılır. Bu çalışmada yanlı regresyon tahmincilerinden Ridge regresyon yöntemi ele alınmıştır. Ridge regresyon parametresinin çözümü ayar parametresi k 'ya bağlıdır. Yapılan bu çalışmada da Ridge regresyon ayar parametresi k 'nın seçimi üzerine analizler yapılmıştır.

Bu çalışmada, Türkiye'nin 1974-2019 yılları arasındaki doğrudan yabancı yatırımlarını etkileyen faktörler veri seti olarak alınmıştır. Doğrudan yabancı yatırımları etkileyen faktörlerle kurulan çoklu doğrusal regresyon modeliyle simülasyon çalışması yapılmıştır. Yapılan simülasyon çalışması ile ayar parametresi k 'nın seçimi için model seçim kriterleri arasından en iyi sonuç vereni seçmek amaçlanmıştır. Simülasyondan sonra gerçek verilerle analiz yapılmış ve sonuçlara göre çoklu doğrusal bağlantı tespit edilmiştir. Modele dahil edilen yanlı tahmincilerden Ridge regresyon ile tekrardan analiz yapıp Ridge regresyonun en küçük kareler yönteminden daha iyi sonuç vererek bağlantı sorunu çözdüğü tespit edilmiştir.

Anahtar kelimeler: Çoklu Doğrusal Bağlantı, Ridge Regresyon, Parametre Tahmincileri, LASSO.

ABSTRACT

The problem that arises when there is a linear relationship between one or more independent variables in the multiple linear regression model is called the multicollinearity problem. In the case of multicollinearity, the variance of the least squares estimators is high and the parameter estimates are calculated incorrectly. In this case, least squares estimators are incapable of interpreting the results, making the least squares method unreliable. Biased regression estimators are used as an alternative to the least squares estimator to solve the multicollinearity problem and estimate the model parameters. In this study, Ridge regression method, which is one of the biased regression estimators, is discussed. The solution of the Ridge regression parameter depends on the tuning parameter k . In this study, analyzes were made on the selection of the Ridge regression adjustment parameter k .

In this study, the factors affecting Turkey's foreign direct investments between 1974-2019 were taken as a data set. A simulation study was carried out with the multiple linear regression model established with the factors affecting foreign direct investments. With the simulation study, it was aimed to choose the one that gives the best result among the model selection criteria for the selection of the tuning parameter k . After the simulation, analysis was made with real data and multicollinearity was determined according to the results. By reanalyzing with Ridge regression, one of the biased estimators included in the model, it was determined that Ridge regression solved the connection problem by giving better results than the least squares method.

Keywords: Multiple Linear Connection, Ridge Regression, Parameter Estimators, LASSO.

İÇİNDEKİLER

ONUR SÖZÜ	i
ÖNSÖZ	ii
ÖZET	iii
ABSTRACT.....	iv
İÇİNDEKİLER	v
TABLolar LİSTESİ	vii
ŞEKİLLER LİSTESİ	viii
KISALTMALAR	ix
GİRİŞ	1
1. REGRESYON ANALİZİ.....	3
1.1. Basit Doğrusal Regresyon Analizi	3
1.2. Çoklu Doğrusal Regresyon Analizi	4
1.2.1. Otokorelasyon.....	5
1.2.2. Değişen Varyans.....	6
1.2.3. Normallikten Sapmalar.....	6
1.2.4. Çoklu Doğrusal Bağlantı	7
2. ÇOKLU DOĞRUSAL BAĞLANTI.....	8
2.1. Tam Çoklu Doğrusal Bağlantı	9
2.2. Tama Yakın Çoklu Doğrusal Bağlantı	9
2.3. Çoklu Doğrusal Bağlantının Tespiti.....	10
2.3.1. Basit Korelasyon Katsayılarının İncelenmesi	10
2.3.2. Varyans Artış Faktörü (VIF) Değerlerinin İncelenmesi.....	10
2.3.3. Koşul İndeksinin (CI) İncelenmesi.....	11
3. RİDGE REGRESYON	12
3.1. Ridge Kestiricisi ile EKK Kestiricisi Arasındaki ilişki.....	13
3.2. Z ve W matrislerinin özellikleri	14
3.3. Ridge Regresyon Parametre Seçim Kriterleri	14
3.3.1. Akaike Bilgi Kriteri (AIC)	15
3.3.2. Bayes Bilgi Kriteri (BIC)	15
3.3.3. Çapraz Geçerlilik (CV) Model Seçim Kriteri	16

3.3.4. Genelleştirilmiş Çapraz Geçerlilik Ölçütü (GCV)	16
3.3.5. Mallow'un Cp Kriteri	17
3.3.6. En Küçük Mutlak Küçülme ve Seçim Operatörü (LASSO)	17
4. SİMÜLASYON	19
5. UYGULAMA	28
5.1. Uygulamanın Amacı ve Modeli	28
5.2. Uygulama Analiz Sonuçları	29
SONUÇ	35
KAYNAKÇA	36
EKLER	38



TABLÖLAR LİSTESİ

Tablo 4.1. Örnek 1 için Simülasyon Çıktısı	23
Tablo 4.2. Örnek 2 için Simülasyon Çıktısı	24
Tablo 4.3. Örnek 3 için Simülasyon Çıktısı	25
Tablo 4.4. Örnek 4 için Simülasyon Çıktısı	26
Tablo 4.5. Örnek 5 için Simülasyon Çıktısı	27
Tablo 5.1. Değişkenler ve Açıklamaları	29
Tablo 5.2. VIF Değerleri	29
Tablo 5.3. Parametre Dereceleri	30



ŞEKİLLER LİSTESİ

Şekil 5.1. Seçim Kriterleri için Ayar Parametresi Fark Grafiği	32
Şekil 5.2. Değişkenlerin Korelasyon Değerleri ve Saçılım Grafikleri	33



KISALTMALAR

EKK	En Küçük Kareler
HKO	Hata Kareler Ortalaması
HKT	Hata Kareler Toplamı
AIC	Akaike Bilgi Kriteri
BIC	Bayes Bilgi Kriteri
CV	Çapraz Geçerlilik
GCV	Genelleştirilmiş çapraz geçerlilik
DYY	Doğrudan Yabancı Yatırım
ÇDB	Çoklu Doğrusal Bağlantı
MSE	Hata Kareler Ortalaması
RMSE	Kök Ortalama Kereler Hatası
LASSO	En Küçük Mutlak Küçülme ve Seçim Operatörü
PE	Tahmin Hatası
RPE	Göreceli Tahmin Hatası
$\hat{\beta}$	En Küçük Kareler Tahmin Edicisi
$\hat{\beta}_k$	Ridge Tahmin Edicisi
k	Ridge Tahmin Edicisini Yan Parametresi
Y	Bağımlı Değişken
X	Bağımsız Değişken
ε	Hata Terimi

GİRİŞ

Uluslararası ekonomide gitmiş oldukları bölgelerde yeni firma kurulması, mevcut yerli ve yabancı firmalara ilave yatırım yapmak ya da o bölgedeki firmaları satın almasıyla ortaya çıkan sermaye akımına doğrudan yabancı yatırım (DYY) adı verilmektedir. DYY'ler gitmiş olduğu bölgelerde tesis kurmak, şirket kurmak, mevcut firma ve şirketleri satın almak, ortak girişimlerde bulunmak şeklinde olabileceği gibi çeşitli anlaşmalarla da olabilmektedir (Çinko, 2009: 118).

Gelişmekte olan ülkeler kalkınma sürecinde sermaye problemi yaşarken gelişmiş ülkeler ise sermaye açısından zengin olması dünya ekonomisinde kaynak dağılımının olmaması sorununu ortaya çıkarmıştır. Bununla birlikte ticari serbestleşme ve devamında finansal serbestleşme ile 1980'li yıllarda sermaye hareketlilikleri büyük oranda hız kazanmıştır. DYY'ler başlarda gelişmiş ülkelerde yaygın olarak görülse de 1980'li yılların sonlarına doğru gelişmekte olan ülkelerde de yoğun bir şekilde görülmeye başlanmıştır. Gelişmekte olan ülkeler sınırları içindeki büyüme ve kalkınma için kullanılacak altyapıda tasarrufa gitmişlerdir. Bu tasarrufla birlikte meydana gelen yetersizlikleri de gidermek için dışarıdan sermayeye ihtiyaç duymuşlardır. DYY'yi çekme konusunda ki oranlar ülkelere değişiklik göstermektedir. Oranlardaki bu değişiklik yatırım yapan firmanın hedefleri ve yatırım yapılacak ülkenin sahip olduğu imkan ve yeterlilikle açıklanabilir. Bununla birlikte DYY'yi etkileyen faktörleri genel olarak tanımlamak mümkün değildir (Karagöz, 2007: 931).

Günümüzde gelişmiş ve gelişmekte olan ülkeler DYY'lerin ülkelerine girişlerini sağlamak ve artırmak için büyük çaba sarf etmektedirler. Ülkelerin büyüme ve kalkınmaları için DYY'ler belirleyici unsur haline gelmiştir. Ülkelerdeki yabancı yatırımlar dış ticaret, istihdam, iş gücü ve ülke ekonomisine büyük oranda katkı sağlamaktadır. (Anbar ve Suleymanlı, 2006: 102).

DYY yatırım kararlarını birçok faktör etkileyebilmektedir. Özellikle sermayenin yetersiz olduğu az gelişmiş ve gelişmekte olan ülkeler için ekonomide büyüme ve kalkınmayı sağlayacak yatırımları belirlemek için borçlanmak yerine DYY yatırımlarını tercih etmektedirler.

Bu çalışmada doğrudan yabancı yatırımı etkileyen faktörlerle kurulan çoklu doğrusal regresyon modeliyle tahminler yapılacaktır. Yapılan tahminler sonucunda bağımsız değişkenler arasında ilişki olması sebebiyle doğru tahminler elde edilememiştir. Çoklu doğrusal regresyon modelindeki varsayımlardan biri de bağımsız değişkenlerin birbiri ile ilişkisiz olmasıdır. Eğer bağımsız değişkenler arasında doğrusal bir ilişki söz konusu ise çoklu doğrusal bağlantı (ÇDB) sorunu diye adlandırılan problem ortaya çıkmaktadır. Bu problem parametre tahminlerinin varyanslarının olması gerektiğinden daha büyük çıkmasına neden olmaktadır. Söz konusu bu durumda çoklu doğrusal regresyon modelinin tahmininde kullanılan en küçük kareler yöntemi (EKK) doğru sonuçlar verememektedir. Bağlantı sorunun çözümü için önerilen yöntemlerden yanlı regresyon tahmincisi olan Ridge regresyon ve modele dahil edilen yanlılık parametresi ile analiz tekrardan yapıp sorun giderilmeye çalışılmaktadır. Çalışmanın asıl amacında modele dahil edilen Ridge regresyon aya parametresi olan k 'nin seçimi için kullanılan kriterlerde analize dahil edilerek en iyi sonucu veren kriter seçimi de yapılmıştır.

Çalışmanın birinci bölümünde, regresyon analizi, basit ve çoklu regresyon analizi açıklanmıştır. Çoklu regresyon analizinin varsayımdan sapmaları açıklanmıştır. Çalışmanın ikinci bölümünde, çoklu doğrusal bağlantı durumu açıklanmıştır. Bağlantı varsayımlarını ve bu varsayımların çözüm yolları hakkında teorik açıklamalar yapılmıştır. Çalışmanın üçüncü bölümünde, çoklu doğrusal bağlantı durumunda çözüm için yanlı tahmincilerden Ridge regresyon anlatılmıştır. Ridge regresyonun ayar parametre seçim kriterleri bu bölümde teorik olarak işlenmiştir. Çalışmanın dördüncü bölümünde, Monte Carlo simülasyon modeli ile ayar parametre tahmincileri analiz edilmiştir. Çalışmanın beşinci bölümünde, gerçek veri seti kullanılarak yapılan analizler ve bu analizler sonunda en iyi sonucu veren ayar parametre tespiti yapılmıştır.

1. REGRESYON ANALİZİ

Regresyon analizi bir değişkenin (bağımlı) bir ya da daha fazla değişkenle (bağımsız) olan ilişkisini matematiksel olarak yazan bir fonksiyondur. Bu fonksiyona da regresyon denklemi adı verilir. Regresyon denkleminde amaç bağımsız değişkenlerin bağımlı değişkeni ne boyutta etkilediğini analiz etmektir.

Regresyon analizi ile ilgili çalışmaların ilkinin 19.yüzyılın sonlarında Galton tarafından yapılmıştır. Galto'nun çalışmasında anne-babaların boyları ile çocuklarının boyları arasındaki ilişkiyi incelemiştir. Galto'nun çalışmaları bugünkü değişkenler arasındaki ilişkiyi inceleyen regresyon analizinin başlangıcı olmuştur.

Regresyon analizinde bağımlı değişken Y , bağımsız değişken X ile gösterilir. Regresyon analizinde bağımsız değişkenlerin sayısına göre ikiye ayrılmaktadır.

1.1. Basit Doğrusal Regresyon Analizi

Basit doğrusal regresyon analizi, bağımlı değişken ile bağımsız değişkenler arasındaki ilişkinin ve bağımsız değişkenlerin bağımlı değişken üzerindeki etkilerinin belirlenmesinde kullanılır. Basit doğrusal regresyon analizi değişkenler arasındaki ilişkinin yapısı ve derecesi ile ilgilenmektedir. Basit doğrusal regresyon analizinde bir bağımsız değişken bulunur.

$$Y = \beta_0 + \beta_1 X + \varepsilon_i, \quad i = 1, 2, \dots, n$$

Modeline basit doğrusal regresyon modeli denir. Y bağımlı değişken, X bağımsız değişkendir. β_0 sabit olup doğrunun y eksenini kestiği noktadır. β_1 doğrunun eğimi ve regresyon katsayısıdır. ε hata terimini gösterir. i alt indisleri belirlenmiş mümkün gözlemi ifade eder.

Basit doğrusal regresyon modelinin varsayımları;

1. Bağımsız değişkenin değerleri sabit kabul edilir. Bağımlı değişkenin değerleri ise rastgeledir.
2. Değişkenler hatasız ölçülmüştür.
3. Her X_i değeri için Y_i değerleri birbirinden bağımsızdır. Y_i gözlemlerinin dağılımı normaldir.

4. Bağımlı ve bağımsız değişkenler arasındaki ilişki doğrusaldır.
5. Gözlem sayısı n , değişken sayısı p 'den çok olmalıdır.

$$n > p$$

Hata terimi (ε_i) varsayımları;

1. Hataların beklenen değeri sıfırdır.
2. ε 'ların olasılık dağılımının varyansı sabittir.
3. Hata değerleri birbirinden bağımsızdır.
4. Hatalar ile bağımlı değişken arasında ilişki yoktur.
5. Hatalar ile bağımsız değişkenler birbirinden bağımsızdır.

1.2. Çoklu Doğrusal Regresyon Analizi

Basit doğrusal regresyon analizinde bağımlı değişkeni etkileyen tek bağımsız değişken vardır. Ancak bazı durumlarda bağımlı değişkeni açıklamak için birden fazla bağımsız değişkene ihtiyaç duymaktadır. Bu durumlarda çoklu doğrusal regresyon kullanılır. Bu gibi durumlarda bağımlı değişkeni etkileyen hangi değişken olduğu analiz edilir.

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_p X_{ip} + \varepsilon_i, \quad i = 1, 2, \dots, n$$

Y_i 'ler bağımlı rassal değişkendir. $X_{i1}, X_{i2}, \dots, X_{ip}$ 'ler bağımsız değişkenlerdir. β_0 bağımsız değişken değerleri sıfır olduğu durumda bağımlı değişkenin alacağı değeri belirten sabit terimdir. $\beta_1, \beta_2, \dots, \beta_p$ kısmi regresyon katsayıları olarak tanımlanır. ε_i aynı dağılımlı özdeş rassal hata terimidir. i alt indisleri belirlenmiş olan mümkün gözlemi ifade eder.

Çoklu doğrusal regresyon modelinin varsayımları;

1. Hata teriminin beklenen değerinin sıfırdır.

$$E(\varepsilon_i) = 0$$

2. Hata terimin varyansı sabittir.

$$\text{Var}(\varepsilon_i) = \sigma^2$$

3. Hata terimlerinin birbirlerinden bağımsız olmalıdır.

$$\text{cov}(\varepsilon_i, \varepsilon_j) = 0$$

4. Hataların varyans ve kovaryansları sınırlı değere sahip olmalıdır.

5. Bağımsız değişkenler arasında bir ilişki olmamasıdır.

$$\hat{\beta} = (X'X)^{-1}X'Y$$

$|X'X| = 0$ regresyon analizindeki parametre kestirimleri sürekliliğini sağlamak ve varyansı küçültebilmek için bu varsayım kabul edilir. $Y = (Y_1, Y_2, \dots, Y_n)'$, $X = (X_1, X_2, \dots, X_n)'$, olarak tanımlanır. (') kullanılan üst simge bir vektörün veya matrisin devriğini almayı gösterir. $\hat{\beta}$ ise $\beta = (\beta_1, \beta_2, \dots, \beta_p)'$ nın EKK tahmincisidir.

Değişkenler arasındaki ilişkileri belirleyebilmek için bu varsayımların gerçekleşmesi lazımdır. Ancak çoklu regresyon varsayımları her olayda geçerli olmayıp sapmalar meydana gelmektedir. Bunlar otokorelasyon, değişen varyans, çoklu doğrusal bağlantı ve normallikten sapmalardır.

1.2.1. Otokorelasyon

Çoklu doğrusal regresyon modeli, ε_i hata terimleri arasında otokorelasyon olmadığını varsayar. Bu durum $E(\varepsilon_i) = 0$ varsayımıyla;

$$E(\varepsilon_i, \varepsilon_j) = E(\varepsilon_i).E(\varepsilon_j) = 0 \quad i \neq j$$

değerini alır. Bu değer sıfırdan farklı olması durumunda hata terimleri arasında otokorelasyon olduğunu gösterir.

$$E(\varepsilon_i, \varepsilon_j) \neq 0 \quad i \neq j$$

Bu durumda En Küçük Kareler (EKK) tahmincileri sapmasız ve tutarlı olup etkin değildirler. Hata teriminin varyansı sapmalıdır. Bunun için parametrelerin varyansları da sapmalı olur. Bunun sonucunda ise t test istatistiği değeri büyük çıkar. Böylece anlamsız bir katsayının anlamlı olma olasılığı da artar. R^2 'de yükselir. R^2 çoklu belirlilik katsayısı olarak adlandırılır ve değişkenler arasındaki ilişkinin kuvvetini ölçer. Bu sebeple F değeri olduğundan büyük bulunur. Sonuç olarak da t ve F testleri güvenilirliğini yitirip yanıltıcı sonuçlar vermektedir. Otokorelasyonun ana nedenleri modele bazı açıklayıcı

değişkenlerin alınmaması, bağımlı değişkenin ölçme hatalı olması, verilerin sistematik incelenmesi gibi nedenler neden olabilir.

Otokorelasyon sorunun çözüm yöntemleri, genelleştirilmiş farklar yöntemi, ilk farklar yöntemi, iki aşamalı Durbin yöntemi, Durbin Watson d istatistiğine dayanan tahmin yöntemi, Cochrane-Orcutt yöntemidir.

1.2.2. Değişen Varyans

Değişen varyans hata teriminin varyansının tüm gözlemler için aynı olmaması durumudur.

$$Var(\varepsilon_i^2) = \sigma^2 \quad \text{ise sabit varyans,}$$

$$Var(\varepsilon_i^2) = \sigma_i^2 \quad \text{ise değişen varyans}$$

Değişen varyansın olmasının genel olarak sebepleri, verilerin toplanmasında yapılan ölçme hataları, dahil edilmesi gerekli değişkenlerin modele eklenmemiş olması, verilerde aşırı uç değerlerin bulunması gibi sebepler değişen varyans sorununun ortaya çıkmasına neden olur.

Değişen varyans sorunun çözümü için önerilen yöntemler, önemli bir değişken varsa modele dahil edilmeli, değişkenlerde dönüşümler yapılması ve ağırlıklı en küçük kareler yönteminin uygulanması gibi değişen varyans sorununu ortadan kalkması için önerilen yöntemlerdir.

1.2.3. Normallikten Sapmalar

Regresyon analizinin temel varsayımlarından birde hata terimlerinin dağılımının normal dağılıma uyduğu varsayımdır.

$$E(\varepsilon_i) = 0$$

$$E(\varepsilon_i^2) = \sigma^2$$

$$E(\varepsilon_i, \varepsilon_j) = 0 \quad i \neq j$$

$$\varepsilon_i = N(0, \sigma^2)$$

olarak ifade edilir. Bu varsayımlarının çoğu durumlarda sağlanması beklenir.

Normal dağılımdan sapmalar meydana geldiğinde çözüm olarak, modelde unutulmuş bağımsız değişken veya değişkenlerin modele dahil edilmesi, güçlü regresyon, örneklem büyüklüğünü artırmak, ağırlıklı en küçük kareler yönteminin uygulanması gibi önerilen yöntemlerdir.

1.2.4. Çoklu Doğrusal Bağlantı

Doğrusal regresyon modelinin varsayımlarından birisi bağımsız değişkenler arasında ilişki olmamasıdır. Çoklu doğrusal regresyon analizinde her bir gözlem için bağımsız değişkenler arasında tam veya tama yakın doğrusal bir ilişki olursa ortaya çıkan probleme çoklu doğrusal bağlantı adı verilir. Bağlantı durumunda doğrusal ve yansız tahminicileri içinde küçük varyansa sahip olan EKK tahmincisi önemli oranda olumsuz etkilenmektedir (Kennedy, 2003: 206).

Modeldeki bağımsız değişkenlerin aynı eğilime sahip olmaları, modelin değişken sayısının gözlem sayısından çok olması, kullanılan veri toplama yöntemlerinin uygun olmaması çoklu doğrusal bağlantının nedenleri arasında gösterilebilir.

Bağlantı sorunun çözümü için panel veri, soruna neden olan değişkenleri modelden çıkarmak ve yanlış tahminicileri kullanarak bağlantı sorununu çözmek için başvurulan yollardandır. Bu yanlış tahmin edicilerden Ridge regresyonu çalışmanın üçüncü bölümde ayrıntıları ile birlikte incelenecektir.

2. ÇOKLU DOĞRUSAL BAĞLANTI

Doğrusal regresyon modelinin varsayımlarından biri de bağımsız değişkenler arasında çoklu doğrusal bağlantı (olmamasıdır). ÇDB problemi, çoklu regresyon analizinde, her bir gözlem için, bağımsız değişkenlerden biri ya da birkaçı arasında, tam veya tama yakın doğrusal bir ilişki olması durumudur. Bu durum, doğrusal ve yansız tahminciler arasında en küçük varyansa sahip olan EKK tahmincisini önemli oranda olumsuz etkilemektedir (Kennedy, 2003: 206).

Bağlantı kavramı, ilk olarak 1934 yılında Ragnar Frisch tarafından bahsedilmiştir. Frisch, bağımlı ve bağımsız değişken ayırımına gitmeden tüm değişkenlerin hata içerdiği ve asıl problemin gerçek değişkenler arasındaki farklı doğrusal ilişkileri tahmin etmek olduğunu ileri sürmüştür.

Literatüründe, ÇDB probleminin ana kütleden kaynaklanan bir problem olmadığı, genel olarak örnekleme ile ilgili ve veri yetersizliğinden kaynaklanan bir problem olduğu konusunda yaygın bir görüş bulunmaktadır. ÇDB problemi bu özelliği ile diğer doğrusal regresyon varsayımlarından sapmalar sonucu ortaya çıkan problemlerden farklıdır (Karakaş, 2008: 3).

ÇDB probleminin olduğu durumunlar da sıklıkla başvurulanan yöntemlerden biri de değişken seçimidir. Çoklu doğrusal regresyon modelinde, bağımlı değişkendeki değişimi açıklamada, uygun bağımsız değişkenlerin alt kümesinin belirlenme işlemi değişken seçimi olarak adlandırılmaktadır (Draper ve Smith, 1998: 327-328).

Bu gibi durumlarda güvenilir yorumlar elde etmek için çoklu doğrusal bağıntıyı ortadan kaldırmak veya etkisini azaltmak gerekir. Bazı araştırmacılar bu sorunu ortadan kaldırmak için soruna sebep olan bağımsız değişken ya da değişkenleri modelde çıkarmayı uygulamışlardır. Bu da analizin doğru ve güvenilir sonuçlar vermemesine neden olmuştur. Bunun yerine modelin katsayılarını yanlış olarak tahmin eden Ridge regresyon modelini uygulayarak sorunu çözmeye çalışmışlardır.

2.1. Tam Çoklu Doğrusal Bağlantı

Tam çoklu doğrusal bağlantı olması durumunda bağımsız değişkenlerin bir veya daha fazlasının diğer değişkenlerle doğrusal bileşimde olması durumudur. Bağımsız değişken vektörleri $X_1, X_2, \dots, X_p$ için, $\lambda_1, \lambda_2, \dots, \lambda_p$ sabitleri ile;

$$\lambda_1 X_1 + \lambda_2 X_2 + \dots + \lambda_p X_p = 0$$

eşitliğinin sağlanabilmesi için sıfır olmayan $k_1, k_2, \dots, k_p$ sabitleri elde edilebiliyorsa, bağımsız değişken vektörlerinde tam doğrusal bağlantı vardır. Genel olarak;

$$\sum_{i=1}^p \lambda_i X_i = 0$$

eşitliği sağlandığı durumda, $i \neq j$ olması koşuluyla;

$$X_j = -\lambda_j^{-1} \sum_{i=1}^p \lambda_i X_i$$

şeklinde değişkenlerin doğrusal bileşimi türünde yazılır. $p \leq n$, p bağımsız değişkene sahip n gözlemlili X veri matrisinin rankı p 'den küçük ve $X'X$ matrisinin özdeğerlerinden en az biri sıfırdır. Tam çoklu doğrusal bağlantı $X'X$ matrisinin tekil olması durumudur (Karakaş, 2008: 9).

2.2. Tama Yakın Çoklu Doğrusal Bağlantı

Tama yakın çoklu doğrusal bağlantı olması durumunda bağımsız değişkenler arasındaki ilişkinin tam değil, yüksek derecede olması durumudur. $p \leq n$ olmak üzere;

$$\sum_{i=1}^p \lambda_i X_i = 0$$

eşitliğinin sağlanması durumudur.

$X'X$ matrisi tekil olmayan bir matristir. $X'X$ matrisinin determinantı sıfır değil sıfıra yakın değer almaktadır. Matrisin tersi hesaplanabilir lakin elemanları büyümektedir. Bu durum söz konusu olduğunda EKK tahmincilerine ait standart hatalar gerçek değerlerinden büyük tahmin edilecektir (Karakaş, 2008: 10).

Son yıllarda ÇDB durumunda, problemi çözmek için önerilen yöntemlerin yanında ÇDB'nin ortaya çıkardığı olumsuzluklardan etkilenmeyen başarılı tahminler yapmaya olanak sağlayan yanlı regresyon tahmincileri kullanılmaktadır. Bir sonraki bölüm olan üçüncü bölümde bu yanlı tahmincilerden Ridge regresyonu ayrıntılı bir şekilde aktarılacaktır.

2.3. Çoklu Doğrusal Bağlantının Tespiti

Çoklu doğrusal bağlantıyı belirleyebilmek için birçok yöntem vardır. Bunlardan bazıları korelasyon matrisinin incelenmesi, varyans artış faktörleri (VIF), öz değerler ve öz vektör analizidir. Bu yöntemler sayesinde sorunun boyutunu ve hangi değişkenden kaynaklandığını görülmektedir. İki değişken arasındaki korelasyon katsayısının 1'e yakın olması çoklu doğrusal bağlantının olduğuna güçlü bir işarettir.

Çoklu doğrusal bağlantıyı tespit etmede yaygın olarak kullanılan bazı yöntemleri aşağıda açıklanacaktır.

2.3.1. Basit Korelasyon Katsayılarının İncelenmesi

p tane bağımsız değişkenlerinin hepsi aşağıdaki gibi standartlaştırma yapılmaktadır.

$$\tilde{X}_{ij} = \frac{X_{ij} - \bar{X}_j}{\sqrt{\sum_{i=1}^n (X_{ij} - \bar{X}_j)^2}} = \tilde{X} \quad \text{ve,}$$
$$\tilde{Y}_{ij} = \frac{Y_i - \bar{Y}}{\sqrt{\sum_{i=1}^n Y_i^2 - \left(\frac{\sum_{i=1}^n Y_i}{n}\right)^2}}, \quad i = 1, 2, \dots, n \text{ ve } j = 1, 2, \dots, k$$

Yukarıdaki standartlaştırma ile elde edilen $\tilde{X}$ matrisi yardımıyla korelasyon matrisi elde edilir.

$$R = (\tilde{X}'\tilde{X})_{p \times p} = \begin{pmatrix} 1 & \cdots & r_{1p} \\ \vdots & \ddots & \vdots \\ r_{p1} & \cdots & 1 \end{pmatrix}_{p \times p}$$

Korelasyon matrisinin köşegen dışı elamanları kontrol edilir. Kontrol edilen elamanların 1'e yaklaşmaları durumunda çoklu doğrusal bağlantı tespit edilir.

2.3.2. Varyans Artış Faktörü (VIF) Değerlerinin İncelenmesi

Varyans arttırıcı faktör (VIF) çoklu doğrusal bağlantıyı belirlenmesinde kullanılan bir yöntemdir. VIF değerlerini bulmanın bir yolu korelasyon matrisinin tersini $(\tilde{X}'\tilde{X})^{-1}$ almaktır. Tersini aldığımız korelasyon matrisini köşegenleri bize VIF değerlerini verir. Diğer bir yolu ise;

$$VIF = \frac{1}{1-R^2}$$

VIF değeri 10 veya daha büyük ($VIF \geq 10$) ise anlamlı çoklu doğrusal bağlantı söz konusudur.

2.3.3. Koşul İndeksinin (CI) İncelenmesi

Çoklu bağlantının olmadığı durumlarda özdeğerler 1'e eşittir. En az bir özdeğerin 1'den farklı olması ya da en az birinin 0'a yakın olması çoklu doğrusal bağlantının varlığını tespit edildiğini gösterir. Ancak özdeğerlerin tek tek incelenmesi bir anlam ifade etmemektedir. Bunun üzerine Vinod ve Ullah 1981 yılındaki çalışmasında en büyük özdeğer ile en küçük özdeğer üzerine dayalı koşul indeksini önermişlerdir. Koşul indeksi için EKK tahmincilerinin hesaplanmasında kullanılan $\tilde{X}'\tilde{X}$ korelasyon matrisinin, özdeğerleri (λ özdeğerleri gösterir);

$$CI = \sqrt{\frac{\lambda_{max}}{\lambda_{min}}} = \sqrt{\frac{en\ büyük\ özdeğer}{en\ küçük\ özdeğer}}$$

$$(\lambda_{max} = \lambda_1 > \lambda_2 > \dots \lambda_k = \lambda_{min})$$

Buradan da $CI > 10$ olduğu durumlarda çoklu doğrusal bağlantı olduğu görülmektedir.

3. RİDGE REGRESYON

Çoklu doğrusal bağlantı problemi durumunda EKK yönteminin kullanılması tahminlerin büyük varyansa sahip olmasına neden olmaktadır. 1970 yılında Hoerl ve Kennard çoklu doğrusal bağlantı problemi olduğunda bu problemi gidermek için Ridge tahmin ediciyi önermişlerdir. Böylelikle daha küçük varyanslı tahmin ediciler elde edilmektedir. Çoklu doğrusal bağlantı problemi olduğunda $X'X$ matrisi tekil değildir. Hoerl ve Kennard ilk kez 1962 yılında $X'X$ matrisine k negatif olmayan çok küçük bir sayı olmak üzere, kI sabitini modele ekleyerek, Ridge tahmin edicisini elde etmişlerdir (Hoerl ve Kennard, 1970: 57).

$X'X$ matrisinde çoklu doğrusal bağlantıdan dolayı bir veya daha fazla özdeğerin küçük olacağını ve bu nedenle $\hat{\beta}$ ile β değerleri arasındaki uzaklığın yüksek olacağını açıklamışlardır. Bu sorunun çözümü içinde en uygun tahmin edicinin ridge tahmin edici olduğunu söylemişlerdir (Hoerl ve Kennard, 1970: 58).

Açıklayıcı değişkenler arasında çoklu bağlantı olması durumunda Ridge regresyon yöntemi ile tahmin edilen β regresyon katsayılarının EKK yöntemiyle yapılan tahminlerden daha küçük HKO'ya sahip olduğu Hoerl ve Kennard 1970 yılında tarafından gösterilmiştir. Hoerl ve Kennard tarafından 1970 yılında önerilen Ridge regresyon tahmin edicisi aşağıdaki gibidir;

Kayıp fonksiyon;

$$M(\beta) = \sum_{i=1}^n (Y_i - \beta_0 - X_i' \beta)^2 + k \sum_{j=1}^p \beta_j^2$$

türevini alıp yazılırsa;

$$\begin{aligned} M(\beta) &= (Y - X\beta)'(Y - X\beta) + k\beta'\beta \\ &= \beta'(X'X + kI)\beta - \beta'X'Y - Y'X'\beta + Y'Y \end{aligned}$$

eşitliği sıfıra eşitlersek;

$$\hat{\beta}_k = (X'X + kI_p)^{-1}X'Y \quad k \geq 0$$

$k = 0$ ise $\hat{\beta}_k = \hat{\beta}$, $k = \infty$ ise $\hat{\beta}_k = 0$ 'dır.

şeklinde ifade edilir. Burada $k \geq 0$ Ridge regresyon ayar parametresi ve I_p ise $p \times p$ boyutunda birim matrisidir. $X'X$ matrisin köşegenine küçük ve pozitif bir k değeri eklenerek yukarıdaki çözüm modeli kurulmuştur.

3.1.Ridge Kestiricisi ile EKK Kestiricisi Arasındaki ilişki

EKK kestiricisi;

$$\hat{\beta} = (X'X)^{-1}X'Y$$

her iki tarafı da $X'X$ ile çarpılırsa;

$$X'X\hat{\beta} = X'Y$$

elde edilir.

Ridge kestiricisi;

$$\hat{\beta}_k = (X'X + kI)^{-1}X'Y$$

bu denklemde de $X'Y$ yerine eşiti yazılırsa;

$$\hat{\beta}_k = (X'X + kI)^{-1}X'X\hat{\beta}$$

elde edilir.

$X'X$ matrisinin tersinin tersi kendisi olduğu için;

$$\hat{\beta}_k = (X'X + kI)^{-1} [(X'X)^{-1}]^{-1}\hat{\beta}$$

şeklinde yazılabilir. Her iki matriste tekil olmadıkları için;

$$\hat{\beta}_k = [(X'X)^{-1} (X'X + kI)]^{-1} \hat{\beta}$$

şeklinde yazılabilir. Buradan da,

$$\hat{\beta}_k = [(X'X)^{-1} X'X + k(X'X)^{-1}]^{-1} \hat{\beta}$$

şeklinde ifade edilir. Yapılan bu işlemlerden sonra

$$\hat{\beta}_k = [I + k(X'X)^{-1}]^{-1} \hat{\beta}$$

olur. $Z = [I + k(X'X)^{-1}]^{-1}$ olarak tanımlanırsa;

$$\hat{\beta}_k = Z\hat{\beta}$$

şeklinde ifade edilir. Bu eşitlik Ridge kestiricisinin EKK kestiricinin bir dönüşümü olduğunu göstermektedir.

3.2. Z ve W matrislerinin özellikleri

$$\hat{\beta}_k = [I + k(X'X)^{-1}]^{-1}\hat{\beta}$$

EKK tahmin edicinin bir fonksiyonu olarak ifade edilir (Hoerl ve Kennard, 1970; 57).

$$W = (X'X + kI)^{-1}$$

olmak üzere, Ridge regresyon tahmin edicisi,

$$\hat{\beta}_k = WX'Y$$

şeklinde ifade edilir.

$X'X$ kolerasyon matrisinin ξ_i (W), ξ_i (Z) ise W ve Z matrisinin özdeğerlerini gösterir.

$$\xi_i (W) = \frac{1}{\lambda_i + k}$$

$$\xi_i (Z) = \frac{\lambda_i}{\lambda_i + k}$$

şeklinde gösterilir. Bu sonuçlar $|W - \xi I|$ ve $|Z - \xi I|$ karakteristik denklemin çözümü ile bulunmuştur.

$$Z = (X'X + kI)^{-1} (X'X)$$

eşitliğin her iki tarafı soldan W^{-1} ile çarpıldığında;

$$Z = I - k(X'X + kI)^{-1} = I - kW$$

W fonksiyonu elde edilir.

3.3. Ridge Regresyon Parametre Seçim Kriterleri

Parametre seçimi yanlı regresyon tahmincilerinin performanslarını etkileyen en önemli konudur. Seçmek istediğimiz parametre, ÇDB'nin çözümünde en önemli

faktördür. Parametre tahmini için birçok kriter bulunmaktadır. Bu kriterlerden AIC, BIC, CV, GCV, CP ve LASSO alt başlıklar halinde verilecektir.

$$Y = X\beta + \varepsilon$$

$$\hat{\beta}_k = (X'X + kI)^{-1}X'Y$$

$$\hat{\varepsilon} = Y - X\hat{\beta}$$

$$H = X(X'X + kI)^{-1}X'$$

$$df = tr(H)$$

$$HKT(Hata Kareler Toplamı) = \hat{\varepsilon}'\varepsilon$$

Belirtilen bilgi kriterleri için gerekli denklemler yukarıda ifade edildiği gibidir.

3.3.1. Akaike Bilgi Kriteri (AIC)

Akaike'nin bilgi kriterleri 1973'te Hirotugu Akaike tarafından geliştirilmiştir. AIC, tahmin edilen herhangi bir istatistiksel modelin uyum iyiliğinin bir göstergesi olarak adlandırılabilir. AIC genellikle yüksek boyutlu gerçekliğe sahip bilinmeyen bir model bulmaya çalışır. Bu modellerin AIC'deki gerçek modeller olmadığı anlamına gelir. AIC kriterleri asimptotik olarak çapraz geçerliliğe eşdeğerdir.

$$AIC = n\log(HKT) + 2df$$

$$AIC = n\log(\hat{\varepsilon}'\varepsilon) + 2tr(H)$$

3.3.2. Bayes Bilgi Kriteri (BIC)

Bayesian bilgi kriterleri 1978'de Gideon E. Schwarz Bayesçi bilgi ölçütünü geliştirmiştir. BIC, farklı sayıda parametreye sahip bir grup parametrik model arasında bir model seçimidir. BIC serbest parametreleri daha güçlü cezalandırmaktadır. Bayes Bilgi kriterleri sadece gerçek modellerle karşılaşmaktadır. Bayes bilgi kriterlerinin tutarlı olduğu söylenebilir.

$$BIC = n\log(HKT) + df\log(n)$$

$$BIC = n\log(\hat{\varepsilon}'\varepsilon) + tr(H)\log(n)$$

3.3.3. Çapraz Geçerlilik (CV) Model Seçim Kriteri

Eğer kare hatalar kullanılacaksa iyi bir tahmin edici olarak birini dışarıda bırakarak çapraz geçerlilik tahmincisi önerilebilir. Bu metot modellerin öngörü yeteneğine göre model seçimi yapılmasını benimsemektedir.

EKK tahmincisi;

$$\hat{\beta} = (X'X)^{-1}X'Y$$

ve tahmin vektörü;

$$\hat{Y} = X\hat{\beta}$$

şeklindedir. $\hat{\beta}$ değeri de yerine konulursa;

$$\hat{Y} = X(X'X)^{-1}X'Y$$

$$H = X(X'X)^{-1}X'$$

$$\hat{Y} = HY$$

şeklinde olur. H matrisi Y ile çarpıldığında tahmin vektörünü vermektedir. H matrisine şapka matrisi denilmektedir. Çapraz geçerlilik modeli;

$$CV = \sum_{i=1}^n (Y_i - \hat{Y}_i^{-1})^2 = \sum_{i=1}^n \left(\frac{Y_i - \hat{Y}_i}{1 - h_{ii}} \right)^2$$

olarak elde edilir. Burada h_{ii} , H şapka matrisinin i . köşegen elemanıdır. Ridge ve Lasso regresyon parametre seçiminde k kat çapraz geçerlilik kullanılmaktadır. n sayıda gözlemi olan veri seti k sayıda eşit parçaya bölünür. k parçaya bölünen veri için $k - 1$ parça eğitim ve geriye kalan ise test seti olarak seçilir. İşlem k adım tekrarlanarak her adımda farklı test seti seçilir. Söz konusu her adım için ortalama kare hatası elde edilir ve tüm tahminlerin ortalaması ele alınarak çapraz doğrulama hata eğrisi elde edilir. Yapılan bu tez çalışmasında $k = 10$ alınmıştır.

3.3.4. Genelleştirilmiş Çapraz Geçerlilik Ölçütü (GCV)

$$CV = \sum_{i=1}^n (Y_i - \hat{Y}_i^{-1})^2 = \sum_{i=1}^n \left(\frac{Y_i - \hat{Y}_i}{1 - h_{ii}} \right)^2$$

Yukarıdaki (CV) modeli üzerinde yapılan değişikliklerle elde edilen genelleştirilmiş çapraz geçerlilik (GCV) ölçütü CV modelindeki h_{ii} yerine $\hat{H}$ matrisinin ortalanması alınarak elde edilen bir ölçüttür. GCV modeli ise;

$$GCV = \frac{n(HKT)}{(n-df)^2}$$

$$GCV = \frac{n(\hat{\varepsilon}'\varepsilon)}{(n-tr(H))^2}$$

şeklindedir.

3.3.5. Mallow'un Cp Kriteri

Mallow'un model seçim kriteri aşağıdaki gibidir.

$$C_p = HKT + 2df\hat{\sigma}^2$$

$$C_p = \hat{\varepsilon}'\varepsilon + 2tr(H)\hat{\sigma}^2$$

Burada; $\hat{\sigma}^2$, σ^2 'nin tahmin edicisidir. Bulunan alt küme modelleri arasında C_p değerini minimum olan model uygun model olarak seçilir. Bu durumda, $\hat{\sigma}^2$, σ^2 'nin iyi bir tahmincisi olmalıdır. Bu durumu sağlayabilmek için modelin Hata kareler Toplamı (HKT) kullanılabilir.

3.3.6. En Küçük Mutlak Küçülme ve Seçim Operatörü (LASSO)

Ridge tahmin edicisinin tahmin gücü oldukça yüksektir. Ancak tahmin sürecinde bir takım problemler ortaya çıkmaktadır. Ridge regresyon katsayıları daraltan bir tahmin edici olması sebebiyle daha kararlı bir yapıya sahiptir. Fakat regresyon katsayılarını sıfıra daraltmadığı için modelin yorumlanması zorlaşır. Bu sebeple modeli yorumlamak için ve tahmin doğruluğunu elde edebilmek için bazı katsayıları sıfıra daraltmak söz konusu olabilir. Bu işlemten sonra varyans küçülmekte fakat yanlı açıdan sapma meydana gelmektedir. Çıkan sonuçta tahmin doğruluğu daha güvenilir olduğundan bu tahmin edicisinin yansızlık özelliği göz ardı edilebilir(Küçük, 2009: 24). 1996 yılında Tibshirani, LASSO yöntemi ile bazı katsayıları sıfıra daraltarak parametre tahminini ve model seçimini eş zamanlı yapmıştır (Tibshirani, 1996; 268). LASSO yöntemi genel olarak aşağıdaki gibidir.

$$\hat{\beta}_{LASSO} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^n (Y_i - X_i' \beta)^2 + \lambda \sum_{j=1}^p |\beta_j| \right\}$$

Burada λ ayar parametresidir.



4. SİMÜLASYON

Bu bölümde, ridge parametresinin seçilmesi için kullanılacak altı yöntemin (AIC, BIC, CV, GCV, CP ve LASSO) karşılaştırılması için Monte Carlo simülasyon çalışması ile yapılmıştır.

1975 yılında McDonald ve Galarneau çalışması takip edilerek farklı derecelerde çoklu bağlantıya sahip veri matrisi üretebilmek için;

$$X_{ij} = (1 - \rho^2)^{1/2} z_{ij} + \rho z_{i(p+1)} \quad i = 1, 2, \dots, n \quad j = 1, 2, \dots, p$$

Burada p , bağımsız değişken sayısı, n , örneklem hacmi olmak üzere ρ^2 bağımsız değişkenler arasındaki korelasyon katsayısı, z_{ij} ise $N(0,1)$ dağılımından olan rassal sayılardır.

Yapılan bu çalışmada, değişkenler arasındaki yüksek ve düşük korelasyonların olduğu örnek çalışmalarında, Ridge ve LASSO regresyonların performanslarını karşılaştırmak ve çoklu doğrusal bağlantı durumunda en iyi tahmini veren yöntemi seçmek amaçlanır. Bağımsız değişkenler arasındaki farklı korelasyon değerleri ile altı örnek için simülasyon çalışması yapıp sonuç elde etmek amaçlanmıştır (Zou ve Hastie, 2005; 312).

Simülasyonda kullanılacak doğrusal regresyon modeli aşağıdaki gibidir;

$$Y = X\beta + \sigma\varepsilon \quad \varepsilon \sim N(0, I)$$

Modellerin değerlendirilmesi için simülasyonda eğitim verisi ve test verisi için n birimlik örneklemeler kullanılmıştır. Simülasyon için R Studio’da *glmnet* paketi kullanılmıştır. Eğitim setinde en iyi ceza parametresinin değeri model seçim kriterleri ile belirlenip en iyi model oluşturmak amaçlanmaktadır. Ardından test seti ile modelin iyiliği test edilmektedir. Simülasyonlarda tüm değişkenleri eğitim setine dayalı olarak ortalarız. $\bar{X}_{train} = (\bar{X}_{1,train}, \dots, \bar{X}_{p,train})$ eğitim verilerini vektörlerini belirtir. Test veri setindeki gözlem sayısı n_{test} ve $\bar{Y}_{train} = (Y_{train} - \bar{Y}_{train})$ eğitim verilerinde ortalama yanıtları gösterir. Son olarak iki performans ölçüsü hesaplanmıştır. Bunlar, test hatası (mean-

squared error: hata kareler ortalaması) $MSE_Y = \frac{1}{n_{test}} r_{sim}^T r_{sim}$ ise $r_{sim} = X_i \beta - (\bar{Y}_{train} + (X_i + \bar{X}_{train})^T \hat{\beta})$ ve tahminin ortalama kare hatası β ise $MSE_{\beta} = |\hat{\beta} - \beta|^2$ olarak gösterilir (Tutz and Ulbricht, 2009; 242).

Sırasıyla eğitim ve test setini göstermek için ./ gösterimi kullanılacaktır. Simülasyon çalışması için incelenen örnekler aşağıdaki gibidir.

1. Her veri seti 20/200 gözleminden oluşmaktadır. $\beta = (3, 1.5, 0, 0, 2, 0, 0, 0)'$ olarak $\sigma = 1, 3$ ve $X \sim N(0, \Sigma)$ ise $\Sigma_{ij} = \rho^{|i-j|}$, $\rho = 0.3, 0.6, 0.9$
2. Her veri seti 20/200 gözleminden oluşmaktadır. $\beta = \left(\underbrace{0.85, 0.85, \dots, 0.85}_8 \right)'$ olarak $\sigma = 1, 3$ ve $X \sim N(0, \Sigma)$ ise $\Sigma_{ij} = \rho^{|i-j|}$, $\rho = 0.3, 0.6, 0.9$
3. Her veri seti 20/200 gözleminden oluşmaktadır. $\beta = (3, 1.5, 0, 0, 0, 0, -1, -1)'$ olarak $\sigma = 1, 3$ ve $X \sim N(0, \Sigma)$ ise $\Sigma_{ij} = \rho^{|i-j|}$, $\rho = 0.3, 0.6, 0.9$
4. Her veri seti 100/200 gözleminden, 30 tahminciden oluşmaktadır. $\beta = \left(\underbrace{2, \dots, 2}_8, \underbrace{0, \dots, 0}_{22} \right)'$ olarak $\sigma = 1, 3$ ve $X \sim N(0, \Sigma)$ ise $\Sigma_{ij} = \rho^{|i-j|}$, $\rho = 0.3, 0.6, 0.9$
5. Her veri seti 100/200 gözleminden, 40 tahminciden oluşmaktadır. $\beta = \left(\underbrace{0, \dots, 0}_{10}, \underbrace{2, \dots, 2}_{10}, \underbrace{0, \dots, 0}_{10}, \underbrace{2, \dots, 2}_{10} \right)'$ olarak $\sigma = 1, 3$ ve $X \sim N(0, \Sigma)$ ise $\Sigma_{ij} = \rho^{|i-j|}$, $\rho = 0.3, 0.6, 0.9$

Yukarıda verilen örnek modellerini 1000 veri seti ile simülasyon yaparak araştırıyoruz. Bu örnek simülasyon çalışmaları aşağıdaki tablolarda verilmiştir. Bu örneklerle yapılan simülasyon çalışmasına Koşul İndisi (CI) eklenerek çoklu doğrusal bağlantının varlığı durumunda seçim parametrelerinin sonuçları karşılaştırılmıştır (Arashi vd., 2021: 7-8).

$$RMSE(\hat{\beta}^*) = MSE(\hat{\beta}) / MSE(\hat{\beta}^*)$$

burada $\hat{\beta}^*$ yukarıda belirttiğimiz tahmincilerden bir tanesidir. Eğer, $RMSE(\hat{\beta}^*)$ değerinin 1'den büyük ise, EKK tahmincisinden daha üstün performans gösterdiğini belirtir.

Simülasyon sonuçlarına bakıldığında tüm örneklerde EKK, $RMSE_{\beta}$ ve $RMSE_Y$ göre de en kötü sonuçları vermiştir. Simülasyon çıktılarının yorumları aşağıda verilmiştir.

Örnek 1 için simülasyon sonuçlarına Tablo 4.1'den bakıldığında, çoklu doğrusallığın düşük, orta ve yüksek olduğu aralıklarda $\sigma = 1$ 'de $RMSE_{\beta}$ ve $RMSE_Y$ değerlerinde EKK'ya göre en iyi sonucu veren kriter LASSO(CV)'dir. $\sigma = 3$ ise en iyi sonucu veren kriter RİDGE (CV)'dir. Çoklu doğrusallığın yüksek ve çoklu doğrusal bağlantının arttığı (CI değeri yüksek olduğu değerler) durumda ise EKK'ya göre en kötü sonucu veren kriter RİDGE(AIC) en iyi sonucu veren kriter ise RİDGE(CV) olduğu görülmüştür.

Örnek 2 için simülasyon sonuçlarına Tablo 4.2'den bakıldığında, çoklu doğrusallığın düşük, orta ve yüksek olduğu aralıklarda $\sigma = 1$ 'de $RMSE_{\beta}$ ve $RMSE_Y$ değerlerinde EKK'ya göre en iyi sonucu veren kriter RİDGE(CV)'dir. $\sigma = 3$ ise en iyi sonucu veren kriter LASSO(CV)'dir. Fakat çoklu doğrusallığın yüksek ve çoklu doğrusal bağlantının arttığı (CI değeri yüksek olduğu değerler) durumda ise EKK'ya göre en kötü sonucu veren kriter RİDGE(AIC) en iyi sonucu veren kriter ise RİDGE(CV) olduğu görülmektedir.

Örnek 3 için simülasyon sonuçlarına Tablo 4.3'den bakıldığında, çoklu doğrusallığın düşük olduğu aralıkta $\sigma = 1$ de $RMSE_{\beta}$ ve $RMSE_Y$ değerlerinde EKK'ya göre en iyi sonucu veren kriter LASSO(CV), $\sigma = 3$ ise en iyi sonucu RİDGE(CV) vermiştir. Çoklu doğrusallığın orta olduğu aralıkta $\sigma = 1$ de $RMSE_{\beta}$ ve $RMSE_Y$ değerlerinde EKK'ya göre en iyi sonucu veren kriter LASSO(CV), $\sigma = 3$ ise en iyi sonucu RİDGE(CV) vermiştir. Çoklu doğrusallığın yüksek ve çoklu doğrusal bağlantının arttığı (CI değeri yüksek olduğu değerler) durumda ise EKK'ya göre en kötü sonucu veren kriter RİDGE(AIC) en iyi sonucu veren kriter ise RİDGE(CV) olduğu görülmektedir.

Örnek 4 için simülasyon sonuçlarına Tablo 4.4'den bakıldığında, çoklu doğrusallığın düşük ve orta olduğu aralıkta $\sigma = 1$ ve $\sigma = 3$ de $RMSE_{\beta}$ ve $RMSE_Y$ değerlerinde EKK'ya göre en iyi sonucu veren kriter LASSO(CV) olduğu görülmüştür. Çoklu doğrusallığın yüksek ve çoklu doğrusal bağlantının yüksek oranda arttığı (CI değeri yüksek olduğu değerler) durumda ise EKK'ya göre en kötü sonucu veren kriter $\sigma = 1$ 'de $RMSE_{\beta}$ 'ya göre RİDGE(AIC), $RMSE_Y$ 'ye göre RİDGE(BIC) en iyi sonucu ise LASSO(CV) vermiştir. $\sigma = 3$ ise en kötü sonucu $RMSE_{\beta}$ göre RİDGE (AIC), $RMSE_Y$

göre RİDGE(BIC)'dir, en iyi sonucu da $RMSE_{\beta}$ göre RİDGE(BIC), $RMSE_Y$ göre LASSO(CV) olduğu görülmüştür.

Örnek 5 için simülasyon sonuçlara Tablo 4.5'den bakıldığında, çoklu doğrusallığın düşük olduğu aralıkta $\sigma = 1$ ve $\sigma = 3$ de $RMSE_{\beta}$ ve $RMSE_Y$ değerlerinde EKK'ya göre en iyi sonucu veren kriter LASSO(CV) olduğu görülmüştür. Çoklu doğrusallığın orta olduğu aralıkta $\sigma = 1$ en iyi sonucu veren kriter LASSO(CV), $\sigma = 3$ de $RMSE_{\beta}$ göre en iyi sonuç RİDGE(CV) ve $RMSE_Y$ değerlerinde ise LASSO(CV) en iyi sonucu vermiştir. Çoklu doğrusallığın ve çoklu doğrusal bağlantını yüksek olduğu değerde $\sigma = 1$ için en iyi sonuç $RMSE_{\beta}$ göre RİDGE(BIC), $RMSE_Y$ göre ise RİDGE(GCV)'dir. $\sigma = 3$ için en kötü sonucu $RMSE_{\beta}$ ve $RMSE_Y$ değerlerinde LASSO(CV) vermiştir. En iyi sonucu ise $RMSE_{\beta}$ göre RİDGE(BIC), $RMSE_Y$ göre de RİDGE(CV) vermiştir.

Simülasyon sonuçlarına bakıldığında çoklu doğrusal bağlantını yüksek olduğu (CI yüksek olduğu değer) durumda Ridge regresyon ayar parametresi seçim kriterleri EKK tahmininden daha iyi sonuçlar verdiği görülmüştür. Bağlantının en yüksek olduğu durumda EKK'dan sonra en kötü sonucu LASSO(CV) vermiştir. En iyi sonuçları ise RİDGE(BIC) ve RİDGE(CV) vermiştir.

Tablo 4.1. Örnek 1 için Simülasyon Çıktısı

ρ	CI	σ	YÖNTEM	MSE_{β}	$RMSE_{\beta}$	MSE_Y	$RMSE_Y$
0.3	1.886	1	EKK	0.894	1.000	0.887	1.000
			RIDGE(GCV)	0.827	1.081	0.839	1.057
			RIDGE(AIC)	0.837	1.068	0.842	1.053
			RIDGE(BIC)	0.830	1.078	0.840	1.056
			RIDGE(CP)	0.827	1.081	0.840	1.056
			RIDGE(CV)	0.855	1.046	0.892	0.995
			LASSO(CV)	0.591	1.514	0.625	1.419
		3	EKK	8.049	1.000	7.985	1.000
			RIDGE(GCV)	5.104	1.577	5.595	1.427
			RIDGE(AIC)	5.546	1.451	5.882	1.357
			RIDGE(BIC)	5.448	1.478	6.014	1.328
			RIDGE(CP)	5.040	1.597	5.559	1.436
			RIDGE(CV)	5.013	1.606	5.529	1.444
			LASSO(CV)	5.116	1.573	5.587	1.429
0.6	3.605	1	EKK	1.494	1.000	0.887	1.000
			RIDGE(GCV)	1.249	1.195	0.802	1.106
			RIDGE(AIC)	1.288	1.160	0.807	1.100
			RIDGE(BIC)	1.257	1.189	0.805	1.102
			RIDGE(CP)	1.246	1.198	0.805	1.102
			RIDGE(CV)	1.250	1.195	0.872	1.017
			LASSO(CV)	0.863	1.730	0.606	1.464
		3	EKK	13.443	1.000	7.958	1.000
			RIDGE(GCV)	6.397	2.102	4.954	1.612
			RIDGE(AIC)	7.596	1.770	5.405	1.477
			RIDGE(BIC)	6.466	2.079	5.143	1.553
			RIDGE(CP)	6.246	2.152	4.961	1.610
			RIDGE(CV)	5.705	2.356	4.621	1.728
			LASSO(CV)	6.609	2.034	4.971	1.606
0.9	10.880	1	EKK	6.322	1.000	0.887	1.000
			RIDGE(GCV)	3.541	1.785	0.633	1.401
			RIDGE(AIC)	4.052	1.560	0.673	1.318
			RIDGE(BIC)	3.613	1.750	0.651	1.363
			RIDGE(CP)	3.486	1.814	0.633	1.401
			RIDGE(CV)	3.046	2.076	0.639	1.390
			LASSO(CV)	3.087	2.048	0.561	1.581
		3	EKK	56.895	1.000	7.985	1.000
			RIDGE(GCV)	14.562	3.907	3.558	2.244
			RIDGE(AIC)	21.676	2.625	4.341	1.840
			RIDGE(BIC)	12.978	4.384	3.528	2.263
			RIDGE(CP)	13.739	4.141	3.518	2.270
			RIDGE(CV)	8.250	6.896	2.860	2.792
			LASSO(CV)	18.000	3.161	3.884	2.056

Tablo 4.2. Örnek 2 için Simülasyon Çıktısı

ρ	CI	σ	YÖNTEM	MSE_{β}	$RMSE_{\beta}$	MSE_Y	$RMSE_Y$
0.3 1.886		1	EKK	0.894	1.000	0.887	1.000
			RIDGE(GCV)	0.649	1.358	0.740	1.199
			RIDGE(AIC)	0.707	1.266	0.756	1.173
			RIDGE(BIC)	0.665	1.346	0.747	1.187
			RIDGE(CP)	0.652	1.373	0.742	1.195
			RIDGE(CV)	0.618	1.447	0.708	1.253
			LASSO(CV)	0.940	0.952	0.964	0.920
		3	EKK	8.049	1.000	7.985	1.000
			RIDGE(GCV)	3.193	2.521	4.227	1.889
			RIDGE(AIC)	4.003	2.011	4.780	1.671
			RIDGE(BIC)	3.243	2.482	4.603	1.735
			RIDGE(CP)	3.103	2.594	4.246	1.881
			RIDGE(CV)	3.198	2.517	4.286	1.863
			LASSO(CV)	5.098	1.579	6.174	1.293
0.6 3.605		1	EKK	1.494	1.000	0.887	1.000
			RIDGE(GCV)	0.686	2.176	0.587	1.511
			RIDGE(AIC)	0.843	1.771	0.632	1.405
			RIDGE(BIC)	0.683	2.188	0.595	1.492
			RIDGE(CP)	0.657	2.273	0.587	1.511
			RIDGE(CV)	0.602	2.483	0.543	1.633
			LASSO(CV)	1.429	1.045	0.907	0.978
		3	EKK	13.443	1.000	7.958	1.000
			RIDGE(GCV)	3.409	3.943	3.533	2.260
			RIDGE(AIC)	4.996	2.691	4.190	1.906
			RIDGE(BIC)	3.041	4.420	3.597	2.220
			RIDGE(CP)	3.171	4.240	3.537	2.258
			RIDGE(CV)	3.094	4.344	3.399	2.349
			LASSO(CV)	6.427	2.092	5.367	1.488
0.9 10.880		1	EKK	6.322	1.000	0.887	1.000
			RIDGE(GCV)	1.297	4.876	0.369	2.406
			RIDGE(AIC)	2.090	3.024	0.450	1.971
			RIDGE(BIC)	1.140	5.544	0.364	2.436
			RIDGE(CP)	1.197	5.280	0.367	2.419
			RIDGE(CV)	0.574	11.011	0.287	3.092
			LASSO(CV)	3.674	1.720	0.654	1.356
		3	EKK	56.895	1.000	7.985	1.000
			RIDGE(GCV)	8.990	6.328	2.681	2.979
			RIDGE(AIC)	16.477	3.453	3.549	2.250
			RIDGE(BIC)	6.601	8.619	2.531	3.155
			RIDGE(CP)	8.577	6.633	2.669	2.992
			RIDGE(CV)	3.449	16.498	2.049	3.897
			LASSO(CV)	17.448	3.261	3.948	2.023

Tablo 4.3. Örnek 3 için Simülasyon Çıktısı

ρ	CI	σ	YÖNTEM	MSE_{β}	$RMSE_{\beta}$	MSE_Y	$RMSE_Y$
0.3	1.886	1	EKK	0.894	1.000	0.887	1.000
			RIDGE(GCV)	0.805	1.111	0.819	1.083
			RIDGE(AIC)	0.821	1.089	0.827	1.073
			RIDGE(BIC)	0.808	1.107	0.821	1.081
			RIDGE(CP)	0.804	1.113	0.820	1.083
			RIDGE(CV)	0.775	1.154	0.816	1.087
			LASSO(CV)	0.660	1.355	0.691	1.285
		3	EKK	8.049	1.000	7.985	1.000
			RIDGE(GCV)	4.773	1.686	5.321	1.501
			RIDGE(AIC)	5.321	1.513	5.704	1.400
			RIDGE(BIC)	5.161	1.560	5.860	1.363
			RIDGE(CP)	4.695	1.714	5.288	1.510
			RIDGE(CV)	4.686	1.718	5.265	1.517
			LASSO(CV)	4.989	1.613	5.541	1.441
0.6	3.605	1	EKK	1.494	1.000	0.887	1.000
			RIDGE(GCV)	1.130	1.322	0.770	1.152
			RIDGE(AIC)	1.213	1.231	0.786	1.129
			RIDGE(BIC)	1.142	1.308	0.775	1.145
			RIDGE(CP)	1.116	1.338	0.771	1.150
			RIDGE(CV)	0.991	1.507	0.781	1.135
			LASSO(CV)	0.975	1.532	0.692	1.282
		3	EKK	13.443	1.000	7.958	1.000
			RIDGE(GCV)	5.783	2.324	4.831	1.653
			RIDGE(AIC)	7.077	1.900	5.274	1.514
			RIDGE(BIC)	5.865	2.292	5.097	1.567
			RIDGE(CP)	5.603	2.399	4.840	1.650
			RIDGE(CV)	5.222	2.574	4.548	1.756
			LASSO(CV)	6.358	2.114	5.129	1.557
0.9	10.880	1	EKK	6.322	1.000	0.887	1.000
			RIDGE(GCV)	2.838	2.227	0.578	1.534
			RIDGE(AIC)	3.491	1.811	0.629	1.411
			RIDGE(BIC)	2.843	2.224	0.589	1.507
			RIDGE(CP)	2.730	2.315	0.577	1.537
			RIDGE(CV)	1.815	3.484	0.540	1.644
			LASSO(CV)	2.938	2.152	0.561	1.581
		3	EKK	56.895	1.000	7.985	1.000
			RIDGE(GCV)	13.620	4.177	3.581	2.230
			RIDGE(AIC)	20.929	2.719	4.345	1.838
			RIDGE(BIC)	12.436	4.575	3.734	2.138
			RIDGE(CP)	12.846	4.429	3.586	2.227
			RIDGE(CV)	8.227	6.916	3.049	2.619
			LASSO(CV)	16.571	3.433	4.022	1.986

Tablo 4.4. Örnek 4 için Simülasyon Çıktısı

ρ	CI	σ	YÖNTEM	MSE_{β}	$RMSE_{\beta}$	MSE_Y	$RMSE_Y$
0.3	2.551	1	EKK	0.530	1.000	0.457	1.000
			RIDGE(GCV)	0.498	1.065	0.439	1.042
			RIDGE(AIC)	0.504	1.052	0.440	1.038
			RIDGE(BIC)	0.493	1.075	0.444	1.029
			RIDGE(CP)	0.498	1.065	0.439	1.042
			RIDGE(CV)	0.536	0.989	0.532	0.860
			LASSO(CV)	0.220	2.409	0.224	2.037
		3	EKK	4.772	1.000	4.115	1.000
			RIDGE(GCV)	3.335	1.431	3.222	1.277
			RIDGE(AIC)	3.532	1.351	3.283	1.254
			RIDGE(BIC)	3.247	1.470	3.598	1.144
			RIDGE(CP)	3.328	1.434	3.224	1.276
			RIDGE(CV)	3.318	1.438	3.235	1.272
			LASSO(CV)	1.985	2.403	2.023	2.035
0.6	4.838	1	EKK	0.930	1.000	0.457	1.000
			RIDGE(GCV)	0.729	1.275	0.402	1.139
			RIDGE(AIC)	0.766	1.215	0.406	1.127
			RIDGE(BIC)	0.680	1.368	0.418	1.094
			RIDGE(CP)	0.728	1.277	0.402	1.138
			RIDGE(CV)	0.657	1.416	0.438	1.044
			LASSO(CV)	0.283	3.285	0.190	2.410
		3	EKK	8.369	1.000	4.115	1.000
			RIDGE(GCV)	3.556	2.353	2.538	1.621
			RIDGE(AIC)	4.051	2.066	2.622	1.569
			RIDGE(BIC)	2.838	2.949	2.882	1.428
			RIDGE(CP)	3.518	2.379	2.544	1.618
			RIDGE(CV)	3.531	2.370	2.547	1.616
			LASSO(CV)	2.546	3.287	1.708	2.409
0.9	18.898	1	EKK	4.119	1.000	0.457	1.000
			RIDGE(GCV)	1.430	2.880	0.260	1.756
			RIDGE(AIC)	1.670	2.467	0.270	1.694
			RIDGE(BIC)	1.110	3.712	0.296	1.547
			RIDGE(CP)	1.409	2.923	0.261	1.753
			RIDGE(CV)	1.072	3.842	0.294	1.556
			LASSO(CV)	0.933	4.415	0.155	2.952
		3	EKK	37.067	1.000	4.115	1.000
			RIDGE(GCV)	4.594	8.068	1.416	2.906
			RIDGE(AIC)	6.254	5.927	1.526	2.697
			RIDGE(BIC)	2.614	14.183	1.687	2.439
			RIDGE(CP)	4.511	8.217	1.427	2.883
			RIDGE(CV)	4.155	8.922	1.382	2.978
			LASSO(CV)	8.117	4.567	1.376	2.992

Tablo 4.5. Örnek 5 için Simülasyon Çıktısı

ρ	CI	σ	YÖNTEM	MSE_{β}	$RMSE_{\beta}$	MSE_Y	$RMSE_Y$
0.3	2.724	1	EKK	0.825	1.000	0.710	1.000
			RIDGE(GCV)	0.787	1.048	0.688	1.033
			RIDGE(AIC)	0.797	1.035	0.691	1.028
			RIDGE(BIC)	0.783	1.053	0.691	1.027
			RIDGE(CP)	0.787	1.048	0.688	1.033
			RIDGE(CV)	0.931	0.886	0.938	0.757
			LASSO(CV)	0.514	1.604	0.488	1.454
		3	EKK	7.424	1.000	6.392	1.000
			RIDGE(GCV)	5.510	1.347	5.185	1.233
			RIDGE(AIC)	5.911	1.256	5.328	1.200
			RIDGE(BIC)	5.351	1.387	5.476	1.167
			RIDGE(CP)	5.496	1.351	5.185	1.233
			RIDGE(CV)	5.464	1.359	5.193	1.231
			LASSO(CV)	4.633	1.602	4.402	1.452
0.6	5.124	1	EKK	1.443	1.000	0.710	1.000
			RIDGE(GCV)	1.159	1.246	0.630	1.127
			RIDGE(AIC)	1.231	1.173	0.640	1.110
			RIDGE(BIC)	1.107	1.304	0.644	1.102
			RIDGE(CP)	1.156	1.249	0.630	1.127
			RIDGE(CV)	1.078	1.339	0.742	0.958
			LASSO(CV)	0.724	1.995	0.423	1.679
		3	EKK	12.991	1.000	6.392	1.000
			RIDGE(GCV)	5.736	2.265	3.980	1.606
			RIDGE(AIC)	6.879	1.889	4.180	1.529
			RIDGE(BIC)	4.804	2.704	4.390	1.456
			RIDGE(CP)	5.670	2.291	3.988	1.603
			RIDGE(CV)	5.698	2.280	3.988	1.603
			LASSO(CV)	6.543	1.985	3.825	1.671
0.9	20.497	1	EKK	6.400	1.000	0.710	1.000
			RIDGE(GCV)	2.567	2.493	0.417	1.701
			RIDGE(AIC)	3.120	2.051	0.444	1.599
			RIDGE(BIC)	2.210	2.896	0.464	1.530
			RIDGE(CP)	2.540	2.519	0.418	1.698
			RIDGE(CV)	2.222	2.880	0.567	1.252
			LASSO(CV)	2.927	2.187	0.435	1.633
		3	EKK	57.596	1.000	6.392	1.000
			RIDGE(GCV)	7.998	7.202	2.139	2.988
			RIDGE(AIC)	11.856	4.858	2.394	2.670
			RIDGE(BIC)	5.423	10.621	2.565	2.492
			RIDGE(CP)	7.805	7.380	2.152	2.970
			RIDGE(CV)	7.341	7.845	2.080	3.073
			LASSO(CV)	22.891	2.516	3.135	2.039

5. UYGULAMA

Bu bölümde ÇDB probleminin EKK tahmincisi üzerindeki olumsuz etkisini ve yanlı regresyon tahmincilerinden Ridge regresyon ile çözmeyi ve hangi ayar parametre seçim kriterinin daha iyi sonuç vereceğine bakacağız. İlk bölümlerde bağlantı problemi, Ridge regresyonu ve parametre seçim kriterlerini teorik olarak anlattık. Bu bölümde iktisadi bir veri seti kullanarak parametre seçim kriterlerinden hangisinin daha iyi ÇDB'yi çözümlediğine bakılacaktır.

5.1. Uygulamanın Amacı ve Modeli

Yapılacak olan bu çalışmanın amacı Doğrudan Yabancı Yatırımı (DYY) etkilen faktörlerle kurulan modelle yapılan analiz sonucunda meydana gelen ÇDB problemini çözmek için yanlı tahmincilerden Ridge regresyon analizi yapmaktır. Yapılacak olan ridge regresyon analizi ile birlikte Ridge parametre tahmincilerinden hangisinin daha iyi sonuç vereceği incelenecektir.

Yapılan bu çalışmada kullanılacak olan veriler Dünya Bankasından(World Bank) <https://databank.worldbank.org/>(18.10.2021) elde edilmiştir. Elde edilen veri seti Türkiye'nin 1974-2019 yıllarına ait doğrudan yabancı yatırımlarını etkileyen faktörlerin verileridir. 46 gözlem ve 8 değişkenle kurulan çoklu doğrusal regresyon modeli aşağıdaki gibidir.

$$DYY = \beta_0 + \beta_1 KBGB + \beta_2 EGD + \beta_3 MHİH + \beta_4 MHİT + \beta_5 GHNT + \beta_6 AKH + \beta_7 CHD + \varepsilon_{it}$$

Kurulan regresyon modelindeki bağımlı ve bağımsız değişkenlerin açıklamaları Tablo 5.1'de detaylı bir şekilde gösterilmiştir (Yüzbaşı ve Ahmed, 2020: 10).

Tablo 5.1. Değişkenler ve Açıklamaları

DEĞİŞKENLER	AÇIKLAMALAR
Bağımlı Değişkenler	
DYY	Doğrudan yabancı yatırımlar, net girişler (GSYİH' nin yüzdesi)
Bağımsız Değişkenler	
KBGB	Kişi başına GSYİH büyümesi(Yıllık %)
EGD	Enflasyon, GSYİH deflatörü (Yıllık %)
MHİH	Mal ve hizmet ihracatı (GSYİH' nin yüzdesi)
MHİT	Mal ve hizmet ithalatı (GSYİH' nin yüzdesi)
GHNTH	Genel hükümet nihai tüketim harcaması (GSYİH' nin yüzdesi)
AKH	Alınan kişisel havaleler (GSYİH' nin yüzdesi)
CHD	Cari hesap dengesi (GSYİH'nin yüzdesi)

5.2. Uygulama Analiz Sonuçları

Tablo 5.2. VIF Değerleri

	EDG	MHİT	GHNTH	KBGB	MHİH	CHD
7.88	2.20	45.18	3.57	2.32	54.78	7.05

MHİT ve MHİH bağımsız değişkenlerinin VIF değerlerini 10'dan büyük olduğu analiz sonucunda tespit edilmiş ve çoklu doğrusal bağlantının varlığı ispat edilmiştir. Ayrıca $CI = 19.86$ değerine bakıldığında 10'dan büyük olduğu görülmektedir. İhracat ve ithalat verileri arasındaki kuvvetli ilişki çoklu doğrusal bağlantı problemine neden olduğu görülmektedir.

Tablo 5.3. Parametre Dereceleri

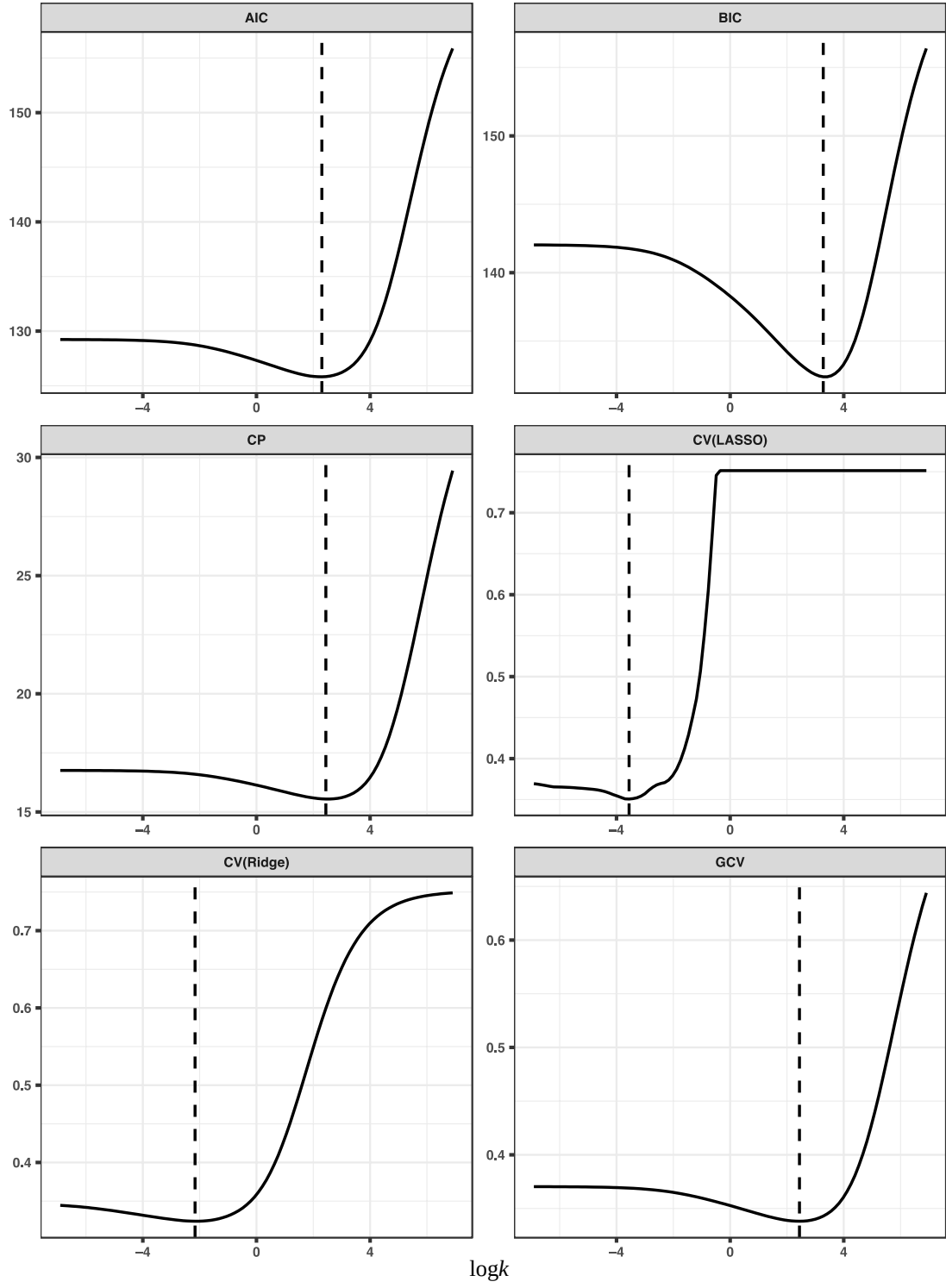
	EKK	RİDGE(GCV)	RİDGE(AIC)	RİDGE(BIC)	RİDGE(CP)	RİDGE(CV)	LASSO(CV)
<i>AKH</i>	-0.288	-0.186	-0.190	-0.161	-0.186	-0.204	-0.235
<i>EDG</i>	-0.151	-0.139	-0.140	-0.124	-0.139	-0.144	-0.135
<i>MHİT</i>	0.327	0.172	0.176	0.148	0.172	0.188	0.267
<i>GHNTH</i>	-0.109	0.041	0.034	0.076	0.041	0.005	0.000
<i>KBG</i>	-0.250	-0.112	-0.121	-0.061	-0.112	-0.150	-0.159
<i>MHİH</i>	0.062	0.155	0.158	0.136	0.155	0.165	0.074
<i>CHD</i>	-0.260	-0.178	-0.186	-0.132	-0.178	-0.213	-0.198
<i>k</i>	-	11.498	10.000	25.561	11.498	0.115	0.022
<i>PE</i>	0.533	0.476	0.479	0.481	0.476	0.459	0.526
<i>RPE</i>	1.000	1.120	1.113	1.110	1.121	1.161	1.014

Veri setini rastgele olacak şekilde iki parçaya bölerek eğitim ve test seti olarak analiz yapılmıştır. Tablo 5.3’de yapılan analiz sonuçlarının tahmin hatası $PE(\hat{\beta}^*) = \frac{1}{n_{test}} \sum_{i=1}^{n_{test}} (Y_{test} - \hat{Y}_{test})^2$ ve göreceli tahmin hatası $RPE(\hat{\beta}^*) = \frac{PE(\hat{\beta})}{PE(\hat{\beta}^*)}$ değerleri görülmektedir. Eğer *RPE* değeri 1’den büyük ise, bu EKK tahmincisinden daha iyi bir performans gösterdiğini gösterir. Tüm değerler için *RPE* değerleri üzerinden yorum yapılacak olursa en iyi tahmini RİDGE(CV) 1.161 değeriyle verdiği görülmektedir. Gerçek regresyon katsayılarına en yakın değeri RİDGE(CV)’nin verdiği görülmüştür. Diğer kriterler EKK’ya göre iyi sonuçlar verseler de RİDGE(CV)’ye göre kötü sonuçlar vermişlerdir. Kurulan regresyon modelindeki bağımsız değişkenlerin bağımlı değişkene etkilerini analiz sonuçlarına göre yorumlamakta mümkündür. En iyi tahmini veren RİDGE(CV) kriterine bakıldığında, değişkenlerdeki %1 oranındaki artışta bağımlı değişken olan *DYY* etkileri açık olarak görülmektedir. EKK’ye göre en iyi sonucu veren bilgi kriteri RİDGE(CV)’ye göre analizi yorumlanırsa, bağımsız değişkenlerin %1’lik artışta bağımlı değişken olan *DYY*’yi etkilerini görmek mümkündür. %1’lik artışta bağımsız değişkenler *AKH* % 0,20 oranında, *EDG* %0.13 oranında, *KBGB* %0.15 oranında ve *CHD* % 0.21 oranında *DYY*’yi azaltmaktadır. *MHİT* %0.18 oranında, *GHNTH* %0.005 oaranında ve *MHİH* %0.16 oranında *DYY*’yi arttırdığı görülmektedir. Analiz yorumlarına bakıldığında pozitif yönde *DYY*’yi etkileyen en iyi bağımsız değişkenler *MHİT* ve *MHİH* olduğu görülmüştür. Ekonomik olarak bakıldığında

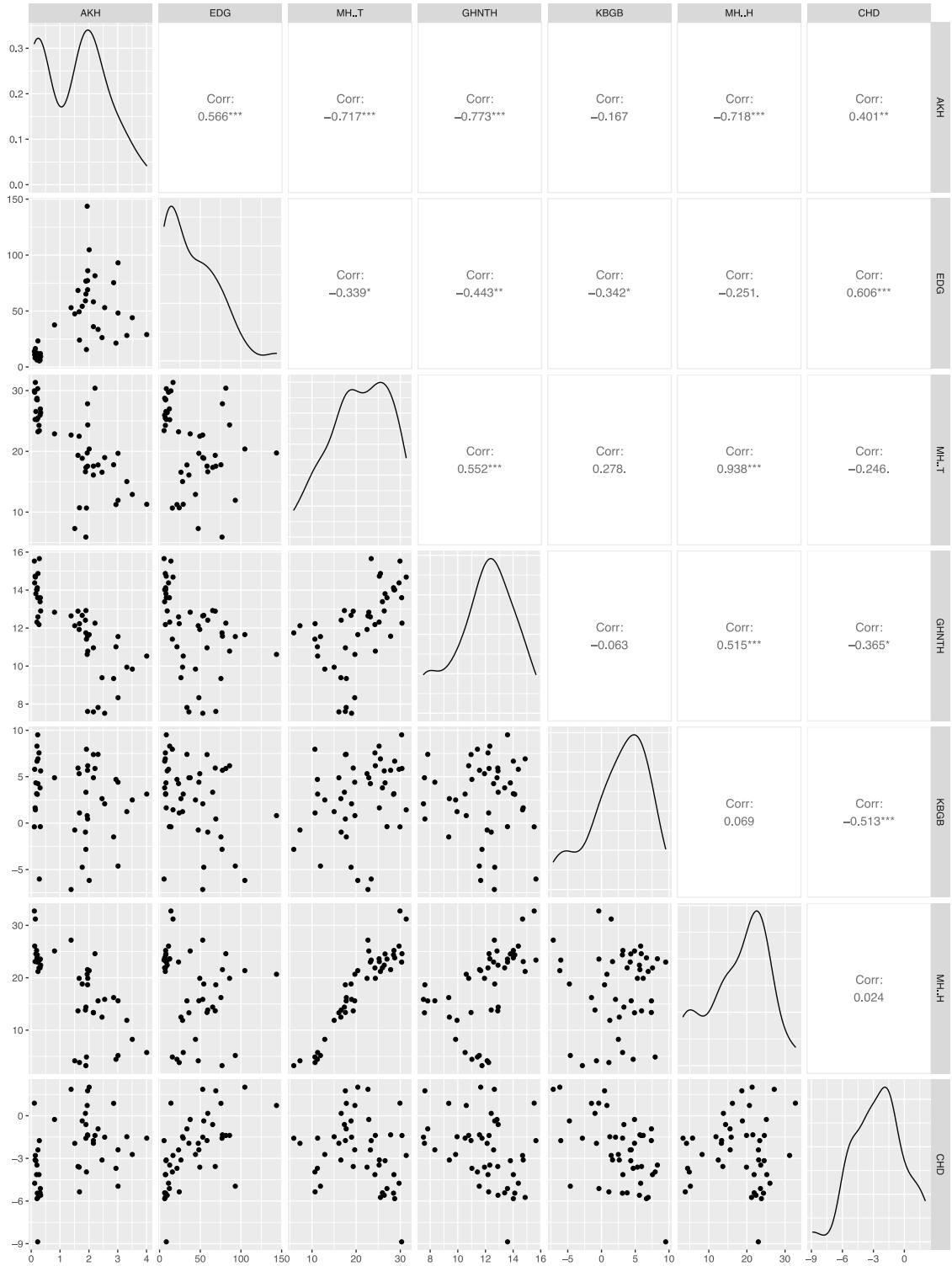
doğrudan yabancı yatırımı etkileyen en iyi değişkenler ithalat ve ihracat olduğu söylenilebilir.

Grafik 1. k ayar parametresi seçimi için kullanılan kriterlerin(AIC, BIC, CP, CV(LASSO), CV(RIDGE), GCV) optimum noktaları grafikte verilmiştir. Aralarındaki farka bakıldığında CV(RIDGE) kriterinin diğer kriterlere göre doğrusal formda en iyi sonuç verdiği görülmektedir.





Şekil 5.1. Seçim Kriterleri için Ayar Parametresi Fark Grafiği



Şekil 5.2. Değişkenlerin Korelasyon Değerleri ve Saçılım Grafikleri

Değişkenlerin korelasyon değerleri ve saçılım grafiklerine Grafik 2'ye bakılarak bağımsız değişkenlerin arasındaki ilişkiden söz etmek mümkündür. Korelasyon değerleri -1 ile $+1$ arasında değerler alır. Korelasyon değerleri 0 ile 1 arasında ise pozitif yönde,

–1 ile 0 arasında ise negatif yönde bir ilişki söz konusudur. Genel olarak korelasyon katsayısının 0.6 ile 0.8 arasında olması yüksek korelasyon, 0.8’in üzerinde olması ise çok yüksek korelasyon olduğunu göstermektedir. Grafik incelendiğinde bağımsız değişkenlerden MHİT ve MHİH arasında 0.93 oranında bir ilişki olduğu görülmektedir. Buna göre ithalat ve ihracat arasındaki ilişkinin pozitif yönde oldukça yüksek olduğu görülmektedir. Diğer değişkenler arasında korelasyon değerleri ve saçılım grafiklerine bakıldığında düşük ve orta düzeyde zayıf bir ilişki olduğu gözlemlenmiştir. Değişkenlere bakıldığında, MHİT ile AKH arasında –0.71, GHNTH ile AKH arasında –0.77 ve MHİH ile AKH arasında –0.71 oranında çıkan sonuçlara göre yüksek oranda ters yönde negatif bir ilişki olduğu görülmektedir. EDH ile AKH arasında 0.56, CHD ile EDH arasında 0.60, GHNTH ile MHİT arasında 0.55 ve MHİH ile GHNTH arasında 0.51 oranında çıkan sonuçlara göre pozitif yönde orta düzeyde bir ilişki olduğu görülmektedir.

SONUÇ

Çoklu doğrusal regresyon analizinde parametrelerin tahmini için genellikle kullanılan EKK yöntemi ile ancak belirli varsayımların sağlanması durumunda güvenilir sonuçlar elde edilir. Bu temel varsayımlardan biri olan bağımsız değişkenler arasında çoklu doğrusal bağlantının olmamasıdır. Değişkenler arasında çoklu doğrusal bağlantının varlığı durumunda EKK yöntemi ile elde edilen sonuçlar güvenilir olamamaktadır. Çoklu doğrusal bağlantının EKK tahminçileri üzerindeki birçok olumsuz etkisi bulunduğundan, bu problemin saptanmasına yarayan yöntemler ve bu problemin düzeltilmesi veya en aza indirgenmesi amacıyla önerilen Ridge regresyon ayrıntılarıyla incelenmiştir.

Bu tezde iki adet uygulama yapılmış ve sonuçlar R Studio programı yardımı ile elde edilmiştir. Birinci uygulama, Ridge ayar parametresinin seçilmesi için kullanılacak altı kriterin (AIC, BIC, CV, GCV, CP ve LASSO) karşılaştırılması bir Monte Carlo simülasyon çalışması ile yapılmıştır. Yapılan simülasyon çalışmasının sonucunda en iyi değeri veren RIDGE(CV) kriteri olduğu gözlemlenmiştir.

İkinci uygulamada ise gerçek bir veri seti ile çalışılmış ve 46 gözlem ve 8 değişkenle hesaplamalar yapılmıştır. Açıklayıcı değişkenler arasındaki çok yüksek korelasyon değerlerinin olmasından dolayı EKK ile uygun bir çözüm elde edilememiştir. Ayrıca açıklayıcı değişkenler için koşul indeksi, VIF değerleri çoklu doğrusal bağlantı sorununu göstermiştir. Dolayısıyla çoklu doğrusal regresyon modeli Ridge regresyon çözümü ile elde edilmiştir. Ridge regresyon EKK yöntemine göre daha iyi sonuçlar vermiştir. Gerçek veri seti ile yapılan çalışmada tüm kriterler için *RPE* değerleri üzerinden yapılan yorumlar sonucunda RIDGE(CV)'nin en iyi sonucu verdiği gözlemlenmiştir.

Her iki uygulamaya bakıldığında tüm kriterler için *RMSE* ve *RPE* değerlerine bakıldığında RIDGE(CV)'nin en iyi sonuçlar verdiği görülmüştür. Tezde kullanılan veri seti ile analize dahi edilen parametreler dışında diğer parametrelerle de analizler yapıp literatüre katkı sağlanabilir.

KAYNAKÇA

- Akaike, H. (1973). Information Theory and an Extension of the Maximum Likelihood Principle. In B. N. Petro and F. Csaki ed. 2nd International Symposium on Information Theory.
- Anbar, A. ve Süleymanlı, J. (2016). Azerbaycan’da Doğrudan Yabancı Yatırımları Etkileyen Faktörlerin Belirlenmesine Yönelik Bir Uygulama, *İktisat ve Girişimcilik Üniversitesi Akademik Bakış Dergisi*, 57(7).
- Arashi, M., Asar, Y., Yüzbaşı, B. (2021). SLASSO: a Scaled LASSO for Multicollinear Situations, *Journal of Statistical Computation and Simulation*, 91(15): 3170-3183
- Çinko, L. (2009). Doğrudan Yabancı Sermaye Hareketlerinin Makroekonomik Etkileri. *Marmara Üniversitesi İ.İ.B.F. Dergisi*, 16(1).
- Draper, R. N. ve Smith, H. (1998). Applied Regression Analysis, *Wiley Series in Probability & Statistics*, US.
- Hoerl, A. E. ve Kennard, R. W. (1970). Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, 12(1): 55-67.
- Karagöz, K. (2007). Türkiye’de Doğrudan Yabancı Yatırım Girişlerini Belirleyen Faktörler: 1970-2005. *Journal of Yasar University*, 2(8): 927-948.
- Karakaş, S. (2008). Çoklu Doğrusal Bağlantı Problemi ve Yanlı Regresyon Tahmincileri, *Fen Bilimleri Enstitüsü Yüksek Lisans Tezi*, s. 1-85.
- Kennedy, P. (2003). *A Guide to Econometrics*, The MIT Press, Fifth Edition, Cambridge.
- Schwarz, G. (1978). Estimating the Dimensions of a Model, *The Annals of Statistical*, 6: 461-464.
- Tibshirani R. (1996). Regression Shrinkage and Selection via the Lasso. *J R Stat Soc Ser B (Methodol)*. 58: 267-288.
- Tutz G. ve Ulbricht J. (2009). Penalized Regression with Correlation-Based Penalty. *Stat Computing*. 19(3): 239-253.

Vinod, H. D. ve Ullah, A. (1981). *Recent Advances In Regression Methods*.

World Bank Group, (2021) <https://databank.worldbank.org/data/source/world-development-indicators> Erişim tarihi (18.10.2021).

Yüzbaşı, B. ve Ahmed, S. E. (2020). Ridge Type Shrinkage Estimation of Seemingly Unrelated Regressions And Analytics of Economic and Financial Data from “Fragile Five” Countries, *Journal of Risk and Financial Management*, 13(6): 131.

Zou, H. ve Hastie, T. (2005). Regularization and Variable Selection via the Elastic Net. *J R Stat Soc: Ser B (Stat Methodol)*. 67(2): 301-320.



EKLER

YILLAR	DYY	KBGB	EGD	MHİH	MHİT	GHNT	AKH	CHD
1974	0,179776	3,123331	28,99235	5,728116	11,27471	10,52844	4,005628	-1,57585
1975	0,255412	4,699301	21,30924	4,421347	11,23177	11,01458	2,939482	-3,69228
1976	0,019501	7,953686	15,61266	4,859086	10,66569	11,41885	1,914972	-3,9567
1977	0,046015	1,095635	24,08902	3,815204	10,70719	12,23137	1,673574	-5,35135
1978	0,05219	-0,74658	47,54115	4,146912	7,288703	12,12466	1,508895	-1,94176
1979	0,083898	-2,82932	76,72087	3,218027	5,881718	11,74904	1,89498	-1,58064
1980	0,026167	-4,62558	93,00322	5,161932	11,92789	11,55317	3,010643	-4,95426
1981	0,133727	2,496909	44,05707	8,236932	12,9047	9,840527	3,505067	-2,72522
1982	0,08521	1,229841	28,22683	11,86405	15,01687	9,943577	3,315448	-1,47491
1983	0,074581	2,633971	26,25825	12,47321	16,55784	9,388169	2,453052	-3,11779
1984	0,188365	4,403882	48,23683	15,6066	19,6735	8,335644	3,012173	-2,39874
1985	0,147245	2,077817	53,05448	15,86072	18,96625	7,515493	2,549269	-1,50666
1986	0,165064	4,888764	36,00689	13,31244	16,10202	7,588005	2,157722	-1,93455
1987	0,131922	7,402176	33,61223	15,58084	17,75744	7,822874	2,318384	-0,9246
1988	0,389641	0,447776	69,01813	18,65397	17,55286	7,612643	1,95481	1,756687
1989	0,618797	-1,48631	75,40483	16,20292	17,78075	9,343364	2,858787	0,875463
1990	0,474527	7,383071	58,2444	13,3651	17,57789	10,96102	2,154287	-1,74215
1991	0,5399	-0,96726	59,16411	13,84113	16,63488	12,41925	1,878985	0,166636
1992	0,532629	3,324226	65,19944	14,39224	17,34513	12,92502	1,898281	-0,61467
1993	0,376312	5,931883	68,37943	13,6738	19,34328	12,89387	1,620139	-3,57052
1994	0,465222	-6,17586	104,7491	21,36213	20,38372	11,65737	2,010098	2,013158
1995	0,522167	6,177188	86,00754	19,89161	24,35103	10,7858	1,962995	-1,37947
1996	0,39785	5,686793	77,22351	21,54266	27,82666	11,57077	1,951778	-1,34288
1997	0,424053	5,886122	81,45486	24,58172	30,38861	12,25939	2,210871	-1,38963
1998	0,34062	0,806059	143,6397	20,66623	19,72918	10,61322	1,940809	0,724723
1999	0,305399	-4,75059	54,27996	18,81091	18,85401	12,66856	1,768041	-0,36078
2000	0,357998	5,322136	49,36078	19,8812	22,47324	11,92866	1,662395	-3,61644
2001	1,661453	-7,14785	52,97976	27,18194	22,68953	12,64592	1,380909	1,863682
2002	0,450358	4,890331	37,57482	25,10512	22,87728	12,83567	0,805816	-0,26056
2003	0,541017	4,252446	23,28928	23,00618	23,21892	12,59042	0,231728	-2,4012
2004	0,681136	8,286033	12,42751	23,62475	25,20108	12,3152	0,196637	-3,47245
2005	1,981204	7,559697	7,085416	21,87627	24,26609	12,18597	0,270191	-4,14372
2006	3,623502	5,620645	9,376206	22,39483	26,36804	12,90753	0,32905	-5,59385
2007	3,235842	3,797987	6,199758	21,88953	25,96129	13,38969	0,30763	-5,42257
2008	2,576506	-0,38292	12,04615	23,56629	26,98196	13,59629	0,316563	-5,11706
2009	1,322249	-6,02791	5,446449	23,3702	23,41687	15,65786	0,28247	-1,74965
2010	1,171054	6,919607	7,009284	21,19413	25,50033	14,87456	0,234108	-5,74265
2011	1,92927	9,509983	8,200847	22,9937	30,31047	13,60036	0,224497	-8,87045
2012	1,560831	3,093102	7,422673	24,36088	28,46993	14,12612	0,213615	-5,44656

2013	1,416083	6,664884	6,28011	23,79301	28,73429	14,01838	0,198479	-5,83201
2014	1,420412	3,168091	7,389519	25,20554	28,56076	14,01998	0,185206	-4,13738
2015	2,228697	4,328128	7,844917	24,53128	26,55727	13,80517	0,161399	-3,16018
2016	1,590791	1,642463	8,130478	23,08351	25,24468	14,73321	0,13637	-3,10903
2017	1,285454	5,794264	10,98242	26,03942	29,72275	14,38026	0,122003	-4,75124
2018	1,647274	1,428915	16,48528	31,20486	31,34254	14,68642	0,144146	-2,79338
2019	1,216924	-0,40135	13,90621	32,7414	29,94146	15,52729	0,106379	0,887674

