BOX-COX DÖNÜŞÜMÜ İLE REGRESYON MODELİNİN BİÇİMLENDİRİLMESİNDE BAZI SORUNLAR

Atıf Evren

Abstract:*Box-Cox transformations are generally introduced to achieve a better fit in regression analysis. However there are some problems. In practice the transformations generally do not yield the desired results since the theory is approximate .*

Klasik doğrusal normal regresyon modelinin varsayımlarından biri de varyansların eşitliği (homoskedastisite) durumudur. Bu varsayımdan sapma halinde, yapılan parametre tahminleri, yansız ve tutarlı olma özelliklerini korumalarına karşılık, minimum varyanslı olma özelliğini yitirmektedirler. Bu durumda da hiç şüphesiz klasik güven aralığı yaklaşımı (ve dolayısıyla gerçekleştirilen hipotez testleri) geçerliliğini yitirmektedir. Böyle hallerde parametre tahminleri, ya ağırlıklı en küçük kareler (A.E.K.K.) yöntemi ile gerçekleştirilmekte, ya da değişkenlerde uygun dönüşümlere gidildikten sonra, klasik en küçük kareler yönteminin (E.K.K.) araçlarına geri dönülerek, sorunun çözümü bulunmaya çalışılmaktadır.

İstatistik ve Ekonometri literatüründe, değişkenlerin olasılık dağılımlarının bilinmesi durumunda, ne tür dönüşümlere gitmenin uygun olacağını inceleyen çalışmalar mevcuttur. Çalışmada yeri geldiğinde bu konuya değinilecektir. Bundan başka olasılık dağılımının bilinmediği durumlarda, ne tür bir dönüşüm uygulanacağını ortaya koymak için geliştirilen analitik yöntemler de söz konusudur. Bu dönüşümlerin içinde en sık kullanılanlardan biri de Box-Cox dönüşümleridir. Çalışmada sözkonusu problemle ilgili literatür kısaca özetlendikten sonra, Türkiye'de ücretliler geçim endeksi ve onu belirleyen faktörlerle ilgili bir uygulama yapılarak, Box-Cox dönüşümlerinin model belirlenmesinde (spesifikasyonunda), ortaya çıkan bazı sorunlara, ne ölçüde çözüm getirdiği olabildiğince tartışılacaktır. Modelde bağımlı veya açıklanan değişken olarak ücretliler geçinme indeksi alınacak, reel faiz oranı, dolar kuru, sanayi indeksi gibi değişkenler de bağımsız ya da açıklayıcı değişkenler olarak ele alınacaktır.

Böyle bir model kurulurken elbette herhangi bir iktisadi geçerlilikten değil, daha çok sezgilerden ve "akla yakınlıktan" hareket edilmektedir[1].Yine ortaya çıkan sonuçların istatistiksel olması, bu sonuçlardan yalnızca lojistik destek alabileceğimiz anlamına gelmektedir. Bunun ötesinde istatistik "iktisat teorisinin" kimi çıkarsamalarının yerine ikame edilebilecek bir "gerçekliği" sunmamaktadır.

Çalışmanın uygulama safhasında ise, sözkonusu değişkenler arasındaki ilişkiler regresyon analizinin geleneksel araçları ile incelendikten sonra, Box-Cox dönüşümü ile modelin varsayımlarına daha uygun bir durum yaratılmaya çalışılacak, belirli bir spesifikasyon hatasının giderilmesine çalışılacaktır.

1.1. Heteroskedastisite

Regresyon analizinde hata terimlerinin varyansının bağımsız değişkenden bağımsız olduğu düşünülür:

$$Var(\varepsilon_i) = Var(Y_i / X = X_i) = \sigma^2, i = 1,2,...,n \quad (1.1.1)$$

Bu durum homoskedastisite durumuna denk düşen durumdur. Bağımlı değişkenin koşullu varyanslarının birbirlerine eşit olması, parametre tahminlerinin belirlenmesinde bütün noktalara aynı oranda güvenilebileceği (ya da güvenilemeyeceği) durumunu yansıtmaktadır. Heteroskedastisite durumu ise

$$Var(Y_i / X = X_i) = Var(\varepsilon_i) = \sigma_i^2 \quad (1.1.2)$$

denklemi ile ifade edilebilir.

Heteroskedastisite durumunun çeşitli nedenleri olabilir. İçerisinde öğrenme sürecinin bulunduğu bir kesitten hareketle gözlemler yapılmışsa, bunun bir neden olduğu düşünülebilir. Zaman içerisinde yapılan iş daha iyi yapılacağından hata azalacak, dolayısıyla σ^2 de küçülecektir. Bilinen bir diğer örnek, gelir arttıkça bireylerin marjinal tüketim ya da tasarruf eğilimleri birbirlerinden büyük farklılıklar arzedeceğinden, yine heteroskedastisite durumunun ortaya çıkacak olmasıdır. Bu durumda varyans artan gelir ile

[1] Bilim ile sağduyunun birbirleri ile çelişebildiğini göstermesi bakımından Lewis Wolpert'in Sarmal Yayınevi tarafından dilimize kazandırılmış olan kitabı "Bilimin Doğal Olmayan Doğası"na bakılabilir.

birlikte artma eğilimi gösterecektir. Bir üçüncü neden olarak veri toplamada kullanılan teknoloji/ yöntem vb. faktörlerin eskiliğinde aranmalıdır. Sözgelimi daha gelişmiş bilgisayar donanımlarına sahip bankaların, aylık ya da üç aylık değerlendirme raporları, diğer bankaların söz konusu raporlarına göre daha az hata içerecektir. Sapan değerler(outliers) de heteroskedastisite durumunun ortaya çıkmasına neden olabilmektedir. Örnek büyüklüğünün küçük olması durumunda, özellikle bir sapan değerin veri setinden dışlanması, parametrelerin nokta ve güven aralığı tahminlerinde dramatik değişikliklere yol açmaktadır. Bir diğer neden ise model spesifikasyonunda yapılmış olması olası bir hatadır. Bu arada heteroskedastisite probleminin kesit verilerle çalışılırken daha fazla karşılaşılan bir problem olduğunu da vurgulamak gerekir. Kesit verilerle çalışılırken, genellikle birbirinden büyüklük olarak farklı değerler aynı veri seti içine dahil edildiğinden, heteroskedastisite söz konusu olmaktadır. Zaman serilerinde ise genellikle büyüklük olarak birbirine yakın değerler benzer kategoriler altında toplandığından, heteroskedastisiteye daha az rastlanılmaktadır.[2]

Heteroskedastisite regresyon analizindeki bağımlı değişkenin varyansının, bağımlı değişkenin ya da değişkenlerin ortalaması ile fonksiyonel olarak ilişkili olduğu durumlarda, kaçınılmaz olarak ortaya çıkmaktadır. Bu gibi durumlarda genel olarak bağımlı değişkenin önemli ölçüde normallikten saptığına da tanık olunmaktadır.[3]

Poisson dağılımı için $E(x) = Var(X) = \lambda$ (1.1.3)

Bernoulli dağılımı için $E(x) = p; Var(x) = p(1-p)$ (1.1.4)

Binom dağılımı için $E(x) = np; Var(x) = np(1-p)$ (1.1.5)

olduğu hatırlanacak olursa, bu tür dağılım özellikleri gösteren gözlem değerleri için, heteroskedastisitenin kaçınılmaz olarak ortaya çıkacağını söylemek zor olmayacaktır.

---

[2] Gujarati,D.;Basic Econometrics;Third Edition;Mc Graw Hill;s355-359.

[3] Neter,J.;Wasserman,W.;Applied Linear Statistical Models;Richard D.Irwin, s131.

1.2.Heteroskedastisite'nin Saptanması

Genel bir kural olarak, heteroskedastisite'nin kesit verilerde sıkça karşılaşılan bir durum olduğundan hareket edilebilir. Yine de bu tür önselliklerin yanısıra bazı deneyselliklerden de yararlanmak gerekmektedir.

1.2.1.Grafik Yöntemler:

Grafiksel yöntemlerle kalıntıların kareleri ile bağımlı değişkenin tahmini arasında belirli bir ilişki olup olmadığını, yine kalıntıların kareleri ile bağımsız değişkenler arasında ilişki olup olmadığına bakarak heteroskedastisiteyi saptamaya çalışmak düşünülebilir. (Her iki durumda da iki değişkenli model için benzer grafiksel çıktılar elde edilmesine karşılık, bağımsız değişken sayısının birden fazla olduğu durumlarda , durumun karmaşıklık arzettiği ve bazı dönüşümlerin gerekli olabileceği İstatistik ve Ekonometri yazınında belirtilmektedir.[4])

1.2.2.Park Testi

Park grafiksel yöntemlere paralel olarak varyans ile bağımlı değişkenler arasında bir fonksiyonel ilişkinin test edilmesi gerektiğini vurgulamaktadır. Park

$$\sigma_i^2 = \sigma^2 X_i^{\beta} e^{vi} \quad (1.2.2.1)$$ ya da

$$\ln\sigma_i^2 = \ln\sigma^2 + \beta \ln X_i + v_i$$ (Burada v_i stokastik hata terimidir.) (1.2.2.2)

modelinin geçerliliğini test etmeyi önermektedir. Betanın istatistiksel olarak anlamlı bulunması halinde, modelde heteroskedastisitenin varlığına hükmedilmektedir.

[4] Gujarati,a.g.e.;s368-369.

1.2.3.Glejser Testi:

Glejser de Park testine paralel olarak bilinen bazı formlardan hareket ederek, kalıntılar ile bağımsız değişkenin arasında herhangi bir fonksiyonel ilişkinin var olup olmadığının, istatistiksel olarak sınanmasını önermektedir.[5]

1.2.4.Spearman Sıra Korelasyonu Testi:

Bunun için E.K.K. ile kalıntılar elde edilmekte, daha sonra da ana kütle hata payının normal dağıldığı varsayımı altında; kalıntıların mutlak büyüklükleri ile bağımsız değişken (ya da bağımlı değişkenin) gözlem değerleri arasındaki Spearman sıra korelasyonu hesaplanarak, gerekli istatistik testler gerçekleştirilmektedir.

1.2.5.Goldfeldt-Quandt Testi

Bu test varyansların bağımsız değişkenlerden herhangi bir tanesi ile pozitif ilişkili olduğu durumda kullanılmaktadır.

$$Y_i = \beta_1 + \beta_2 X_i + \varepsilon_i \qquad (1.2.5.1)$$ modeli ve

$$\sigma_i^2 = \sigma^2 X_i^2 \qquad (1.2.5.2)$$ varsayılsın. Bu durumda bağımsız değişkenin değerleri esas alınarak gözlem değerleri küçükten büyüğe sıralanmakta, bu sıralamada merkezi konumda bulunan c sayıda gözlem değeri analiz dışında bırakıldıktan sonra, geri kalan n-c gözlem değeri iki guruba bölünmektedir.Daha sonra sözkonusu iki grup için iki ayrı regresyon doğrusu tahmin edilerek, her bir regresyon denklemi için kalıntı kareleri toplamı hesaplanmakta ve birbirleri ile kıyaslanmaktadır.[6]

1.2.6.Breusch-Pagan-Godfrey Testi

Goldfeldt-Quandt testininin başarısı merkezde yer alan gözlem sayısı olan c sayısının seçimine bağlıdır. Başka bir deyişle küçükten büyüğe

[5] a.g.e.;s372
[6] a.g.e.;s374

sıralama yapıldığında, merkezi bir konum işgal eden kaç tane gözlem değerinin analiz dışı bırakılacağı testin sonucunu hiç şüphesiz yakından etkileyecektir. Bunun yanısıra hangi bağımsız değişkene göre sıralama yapılacağı da testin sonucunu etkileyecektir. Bu test bu sakıncayı ortadan kaldırmayı amaçlamaktadır.

1.2.7White Testi:

Heteroskedastisiteye neden olan bağımsız değişkenlere göre gözlem değerlerinin yeniden organize edilmesini şart koşan Goldfeldt-Quandt ya da normallik varsayımına çok duyarlı olan Breusch-Pagan-Godfrey testlerinin aksine White, normallik varsayımına dayanmayan ve gerçekleştirilmesi kolay bir test önermektedir. Heteroskedastisite özelliğinin arandığı örnek model şu olsun:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \varepsilon_i \quad (1.2.7.1)$$

Bu modelden elde edilen kalıntılar e_i ile gösterilsin. White bu durumda

$$e_i^2 = \alpha_1 + \alpha_2 X_{2i} + \alpha_3 X_{3i} \alpha_4 X_{2i}^2 + \alpha_5 X_{3i}^2 + \alpha_6 X_{2i} X_{3i} + v_i$$
(1.2.7.2)

regresyon modelinin oluşturulmasını önermektedir.Burada kalıntı karelerinin, bağımsız değişkenlerin birinci ve ikinci derecelerine, çapraz çarpımlarına göre değişimleri incelenmektedir.[7]

1.3.Genelleştirilmiş ve Ağırlıklı E.K.K. Yöntemi ile Heteroskedastisite'nin Giderilmesi

Çok değişkenli doğrusal regresyon modeli

$Y = X\beta + \varepsilon$ (1.3.1) denklemi ile ifade edilir. Ana kütle hata payı ile ilgili varsayımlar

[7] a.g.e.;s380.

$E(\varepsilon) = 0$ (1.3.2)

$Var(\varepsilon) = \sigma^2 I$ 'dır. (1.3.3)

Burada **Y** nx1'lik bağımlı değişkenin gözlem değerleri vektörünü; **X** nxp'lik bağımsız değişkenlerin gözlem değerleri matrisini, β px1'lik parametreler vektörünü, ε ise modele stokastik olma niteliğini kazandıran nx1'lik ana kütle hata payı vektörünü ifade etmektedir. Yine σ^2 değeri bilindiği varsayılan bağımlı değişkenin ana kütle varyansını , **I** da nxn'lik birim matrisi göstermektedir. Burada (1.3.3) no'lu denklemin homoskedastisite ile ilgili varsayımı yansıtmaktadır. Bunun yerine hata paylarının dağılımı ile ilgili olarak; ana köşegende yer alan elemanlarının birbirinden farklı olduğu bir köşegen matris **V**, ve

$Var(\varepsilon) = \sigma^2 V$ (1.3.4) modeli düşünülsün. Dolayısıyla artık

$\hat{\beta} = (X'X)^{-1} X'Y$ (1.3.5) E.K.K. denklemlerinin artık en uygun çözümü veremeyeceği açıktır.Bu durumda bazı dönüşümlere gittikten sonra E.K.K. yöntemi uygulanmaktadır. $\sigma^2 V$ ana kütle hata paylarının varyans kovaryans-matrisi olduğu için pozitif belirli ve tekil olmayan (nonsingular) bir matristir. Dolayısıyla **V**'nin nxn'lik karekök matrisi **K** ;

$K'K = KK = V$ (1.3.6) sözkonusu olmaktadır.

$Z = K^{-1} Y$ (1.3.7)

$B = K^{-1} X$ (1.3.8) ve

$g = K^{-1} \varepsilon$ (1.3.9) dönüşümleri uygulanacak olursa

$Y = X\beta + \varepsilon \; ; \Rightarrow \quad K^{-1} Y = K^{-1} X\beta + K^{-1} \varepsilon$ (1.3.10) ya da

$Z = B\beta + g$ (1.3.11) elde edilir. Yine

$$E(g) = K^{-1}\varepsilon = 0 \quad (1.3.12)$$

$Var(g) = \sigma^2 K^{-1} V K^{-1} = \sigma^2 K^{-1} KK K^{-1} = \sigma^2 I$ (1.3.13) olarak hesaplanmaktadır.[8]

Böylelikle ortalaması sıfır varyansı sabit ve birbirleriyle korelasyonsuz elemanlardan oluşan **g** kalıntılar vektörü elde edilmektedir.Dönüşüm sonrasında E.K.K. yönteminin uygulanması ile elde edilen normal denklemler;

$$(X'V^{-1}X)\hat{\beta} = X'V^{-1}Y \quad (1.3.14)$$

Parametre tahminlerinin varyans-kovaryans matrisi de

$V(\hat{\beta}) = \sigma^2 (B'B)^{-1} = \sigma^2 (X'V^{-1}X)^{-1}$ (1.3.15) yardımı ile hesaplanmaktadır.

Eğer $w_1, w_2, \ldots, w_n$ elemanları ile köşegen bir formda olan **W** matrisi

$W = V^{-1}$ olmak üzere düşünülürse A.E.K.K normal denklemleri

$$(X'WX)\hat{\beta} = X'WY \quad (1.3.16)$$ ve parametre tahminleri de

$$\hat{\beta} = (X'WX)^{-1} X'WY \quad (1.3.17)$$ ile bulunacaktır.

Burada w_i değerleri tartılar ya da ağırlıklar olarak adlandırılmaktadır. Bağıl olarak küçük w_i ya da tartı değerine sahip olanların

[8]Montgomery,D.C.;Peck,E.A.;Introduction to Linear Regression Analysis;John Wiley, Second Edition, s 379.

varyansı; büyük w_i değerine sahip olanların varyansından büyüktür. Bu da sezgilere uygun bir durumdur. A.E.K.K. yönteminin kullanılabilmesi için w_i tartılarının bilinmesi gerekir. Kimi zaman geçmiş bilgiler bu konuda yardımcı olurken kimi zaman de teorik bilgiler tartıların belirlenmesinde belirleyici olmaktadır.[9]

Bazen de kalıntı analizi tartıların belirlenmesinde yol gösterici olmaktadır. Örneğin kalıntılar bağımsız değişkenlerin bir fonksiyonu olarak davranıyorlarsa, bu durumdan yararlanılabilir. Eğer kalıntıların varyansı bağımsız değişkenin / değişkenlerin gözlem değerleri ile $Var(\varepsilon_i) = \sigma^2 X_{ij}$ biçiminde bir ilişki arzediyorsa; tartılar $w_i = 1/X_{ij}$ olacak şekilde belirlenebilir. Yine de tartıların belirlenmesi oldukça zahmetli bir iştir.[10] Bundan başka ağırlıklı ya da genelleştirilmiş E.K.K. yöntemlerinin, hata terimlerinin serisel olarak korelasyonlu olduğu durumlarda da kullanıldığı vurgulanmaktadır.[11]

1.4.Varyans-Stabilize Edici Dönüşümler

A.E.K.K. yöntemine almaşık olarak bağımlı değişkende dönüşüme gitmek de literatürde sıklıkla uygulanmaktadır.Genel olarak bağımlı değişkenin sahip olduğu olasılık dağılımına göre şu dönüşümler önerilmektedir:

σ^2 ile E(Y) arasındaki İlişki	Önerilen Dönüşüm
σ^2 *sabit*	$y' = y$ (dönüşüm yok)
$\sigma^2 \propto E(y)$	$y' = \sqrt{y}(Poisson)$
$\sigma^2 \propto E(y)(1-E(y))$	$y' = \sin^{-1}(\sqrt{y})(binom)$
$\sigma^2 \propto E(y)^2$	$y' = \ln(y)$ (log)

[9] Montgomery,Peck,a.g.e.; s381.
[10] a.g.y.
[11] a.g.y.

$\sigma^2 \propto E(y)^3$	$y' = y^{\frac{-1}{2}}$
$\sigma^2 \propto E(y)^4$	$y' = y^{-1}$

Dönüşümün şiddeti sözkonusu eğriselliğin derecesine bağlı olmaktadır. Örneğin; karekök alınarak gerçekleştirilen bir dönüşüm değişken üzerinde genellikle sınırlı bir etki yaparken bağımlı değişkenin çarpma işlemine göre tersinin alındığı bir dönüşüm ise oldukça büyük bir etkide bulunmaktadır.[12]

1.5 1Box-Cox Dönüşümleri

Bütün yukarıdakilere rağmen uygulanacak dönüşümü bulmak her zaman sorunsuz olmamaktadır. Bu durumda Box-Cox dönüşümü bir almaşık olarak devreye sokulabilir.Bağımlı değişkenin dönüştürülmüş hali $Y^{(\lambda)}$ ile gösterilsin. Box-Cox yöntemi ile önerilen dönüşüm;

$$Y^{(\lambda)} = \begin{cases} \dfrac{Y^{\lambda}-1}{\lambda Y^{\lambda-1}}, \lambda \neq 0 \\ Y \ln Y, \lambda = 0 \end{cases}$$ (1.5.1.1) biçimindedir.

Bu dönüşümlerin bir diğer ifade ediliş biçimi de Bates-Watts'da ifade edildiği gibi[13] ya da Seber'de ifade edildiği gibi

$$y^{(\lambda)} = \begin{cases} \dfrac{y^{\lambda}-1}{\lambda}, \lambda \neq 0 \\ \log y, \lambda = 0 \end{cases}$$[14] (1.5.1.2) biçiminde olmaktadır.

---

[12] a.g.e.;s98-99

[13] Bates,D.M.;Watts,D.G.,Nonlinear Regression Analysis and Its Applications;John Wiley; s28

[14] Seber,G.A.F.;Wild,C.J.;Nonlinear Regression;John Wiley;s70

Burada $\dot{Y}$ gözlem değerlerinin geometrik ortalamasıdır. Dolayısıyla uygulama ile ilgili bir kısıtlama hemen ortaya çıkmaktadır. Bu kısıtlama da bağımlı değişken y'nin bütün değerlerinin pozitif olması gerekliliğidir. Bu kısıtlılığı ortadan gidermek için, yukarıdaki formüle bir almaşık da şu şekilde verilmektedir:

$$y^{(\lambda)} = \begin{cases} \dfrac{(y+\lambda_2)^{\lambda_1}-1}{\lambda_1}, \lambda_1 \neq 0 \\ \log(y+\lambda_2), \lambda_1 = 0 \end{cases} \quad (1.5.1.3)$$

Ancak burada da literatürde sözü edilen kimi problemlerden söz etmek gerekmektedir. Seber bu problemleri kısaca şöyle özetlemektedir: λ_2'nin maksimum olabilirlik tahmini $-y_{\min}$ 'dir ve lambdanın maksimum olabilirlik fonksiyonunun alabileceği değer artı sonsuz olmaktadır. Dolayısıyla uygun bir tahmin yöntemine ihtiyaç bulunmaktadır. (1.5.1.2) türünden bir dönüşüm bir dağılıma simetri getirmek için gündeme getirilmesine karşılık ,$\lambda>1$ olması halinde gerçekleştirilen bir dönüşüm dağılım eğrisinin sağ kuyruğunun uzamasına, ve $\lambda<1$ olması halinde gerçekleştirilen bir dönüşüm ise dağılım eğrisinin sol kuyruğunun uzamasına neden olmaktadır.

Bir kez daha yukarıda ifade edilen teorinin yaklaşık olduğu, $\lambda \neq 0$ için $y^{(\lambda)}$'nın (yukarıdan veya aşağıdan sınırlı olduğu için) normal dağılmadığını belirtmek yararlı olabilir. Son olarak Box-Cox dönüşüm ailesinden gelen dönüşümlerin $y^{(\lambda)}$ nin dağılımının şeklini etkilemesinin yanısıra, $y^{(\lambda)}$'nın varyansı ile beklenen değeri arasındaki ilişkiyi de etkileyebileceğini vurgulamak gerekmektedir. [16]

[15] a.g.e., s72

[16] Seber;a.g.e.;s72

1.5.2 Dönüşümün Tipinin Saptanması

Kalıntı kareleri toplamı KKT(λ) ile gösterilecek olursa, λ'nın en büyük olabilirlik tahmin değeri; KKT(λ)'yi minimize eden değeri vermektedir.[17]

Bu yönteme göre genellikle 10-20 λ değeri için KKT(λ) hesaplanmakta ve bu değerlerden hareketle yapılan enterpolasyon işlemi ile λ'nın en büyük olabilirlik tahmin değeri tahmin edilmeye çalışılmaktadır. Dikkat edilmesi gereken bir nokta, λ'nın kolay yorumlanabileceği bir biçimde seçilmesi gerekliliğidir.. Örneğin $\lambda=0.675$ gibi bir sonucu yorumlamak hiç şüphesiz fiziksel olarak oldukça güçtür. Onun yerine olabildiğince yorumu kolay λ değerleri ya da dönüşümleri seçilmelidir. Analiz sonrasında optimum nokta olarak λ=1 sonucuna ulaşılması dönüşüm yapılmayacağına; λ=2 sonucu bağımlı değişkenin karesinin alınacağına, λ=0.5 sonucu karekök dönüşümünün gerçekleştirileceğini ... λ=0 sonucu ise logaritmik dönüşümün gerçekleştirilmesi gerekliliğine işarettir.

Son olarak λ'nın %100(1-α)'lık güven aralığı tahmininden sözetmek de uygun olacaktır. Bu aralık

$$SS^{*} = KKT(\lambda)(1+\frac{t^{2}_{\alpha/2,\nu}}{\nu})$$ (1.5.2.1) formülü ile verilmektedir.

Burada $\nu = n - p$ olmak üzere regresyon doğrusu için geçerli olan serbestlik derecesini göstermektedir. Eğer hesaplanan güven aralığı $\lambda=1$ değerini içerirse dönüşüm yapılmayacağına hükmedilmektedir.[18]

1.6.Uygulamada Karşılaşılan Bazı Sorunlar

Box-Cox Dönüşümleri ile ilgili veriler Türkiye Cumhuriyeti Merkez Bankası (TCMB) Elektronik Veri Sisteminden elde edilmiştir. Aşağıdaki tabloda "ücret" olarak kısaltılan değişken 1997 yılı baz alınarak hesaplanan ücretliler

[17] Montgomery,...,.s103

[18] a.g.e.,s103-104

geçinme indeksidir. "sanayi" ise yine 1997 yılı baz alınarak hesaplanan DİE aylık sanayi üretim indeksi rakamlarını yansıtmaktadır. "reelfaiz" yıllık ortalama reel faizi, "dolar" da döviz alış (TL/Dolar) kurunu ifade etmektedir. Ham verilerle ilgili tablo şu şekildedir:

Tablo-I:Ham Veriler

	Zaman	Ücret	Sanayi	Reel Faiz	Dolar
Oc.96	1	38.63	89.1	17.4	56
Şub.96	2	40.67	77.1	19.4	63893
Mar96	3	42.99	89.7	18.2	68105
Nis.96	4	46.32	82.8	12.4	72361
May.96	5	49.12	91.4	9.7	76517
Haz.96	6	51.33	90.5	8.7	79478
Tem.96	7	52.7	90.5	9.4	82495
Ağus.96	8	54.07	88	9	84695
Eyl.96	9	56.61	90.6	8.4	88627
Ek.96	10	63.38	100	7.2	93487
Kas.96	11	65.93	99.3	6.3	98233
Ar.96		67.88	94.6	6.5	104443
Oc.97	13	72.3	88.7	7.9	111925
Şub.97	14	74.23	80.7	6.8	118820
Mar.97	15	78.88	99	8	124435
Nis.97	16	84.19	88.3	11.6	130717
May.97	17	89.79	101.1	10.4	136723
Haz.97	18	93.42	99	9.7	143739
Tem.97	19	97.9	101.6	7.1	152600
Ağus.97	20	105.45	101.2	6.5	162989
Eyl.97	21	110.79	109.9	5.9	169744
Ek.97	22	123.77	113	4.5	177478
Kas.97	23	131.6	107.9	4.5	186484
Ar.97	24	137.67	109.6	3.1	199050
Oc.98	25	145.06	90.6	2.3	211141
Şub.98	26	148.97	93.5	3.9	222694
Mar.98	27	153.89	105.2	6	234795
Nis.98	28	164.05	89.7	7.2	244957
May.98	29	171.57	105	8.7	251309
Haz.98	30	177.36	101	8.9	259891
Tem.98	31	183.01	101.8	7.6	267602
Ağus.98	32	185.89	100.3	13.1	273256

Eyl.98	33	196.61	108.3	16.8	274625
Ek.98	34	216.83	110.5	21.6	277950
Kas.98	35	225.98	105.2	25.5	293540
Ar.98	36	233.44	99.1	28.7	306170
Oc.99	37	235.23	81.7	31.8	320509
Şub.99	38	239.79	86.8	32.1	340335
Mar.99	39	251.37	92.4	30.1	359127
Nis.99	40	264.55	92.8	28	378448
May.99	41	272.02	100.7	28	394090

Bu değişkenlerin herbirinin zamana göre artış eğilimi göstereceği düşünülerek yukarıdaki tabloya bir de "zaman" değişkeni eklenmiştir. Daha sonra bütün değişkenlerin zaman ile olan ilişkisi incelenmiş, sözkonusu ilişkiler istatistiksel olarak anlamlı bulunduğu için, zamanın trend etkisi STATISTICA yardımı ile arındırılmaya çalışılmıştır. Burada bir yorumlama sorunu ile karşılaşmamak için doğrusal trend modeli ile yetinildiğini vurgulamak gerekir. Trend etkisi giderildikten sonra elde edilen değerler (kalıntılar) ise şöyledir:

Tablo-II:Zamanın Trend Etkisi Giderildikten Sonra Elde Edilen Değerler

	Zaman	Ücret'	Sanayi'	Reel Faiz'	Dolar'
Oc.96	1	28.031	-1.311	11.946	33434.89
Şub.96	2	24.144	-13.605	13.586	29025.46
Mar.96	3	20.538	-1.299	12.025	25191.04
Nis.96	4	17.941	-8.493	5.865	21400.61
May.96	5	14.815	-.188	2.805	17510.19
Haz.96	6	11.099	-1.382	1.445	12424.77
Tem.96	7	6.542	-1.676	1.785	7395.344
Ağus.96	8	1.986	-4.47	1.025	1548.92
Eyl.96	9	-1.401	-2.165	.065	-2565.5
Ek.96	10	-.557	6.941	-1.495	-5751.93
Kas.96	11	-3.934	5.947	-2.755	-9052.35
Ar.96	12	-7.91	.953	-2.915	-10888.8
Oc.97	13	-9.416	-5.241	-1.875	-11453.2
Şub.97	14	-13.413	-13.636	-3.335	-12604.6
Mar.97	15	-14.689	4.47	-2.496	-15036
Nis.97	16	-15.306	-6.524	.744	-16800.5
May.97	17	-15.632	5.982	-.816	-18840.9
Haz.97	18	-17.929	3.588	-1.876	-19871.3
Tem.97	19	-19.375	5.893	-4.836	-19056.7

Ağus.97	20	-17.751	5.199	-5.796	-16714.2
Eyl.97	21	-18.338	13.605	-6.756	-18005.6
Ek.97	22	-11.284	16.411	-8.516	-18318
Kas.97	23	-9.381	11.016	-8.876	-17358.4
Ar.97	24	-9.237	12.422	-10.636	-12838.9
Oc.98	25	-7.774	-6.872	-11.796	-8794.28
Şub.98	26	-9.79	-4.266	-10.557	-5287.7
Mar.98	27	-10.796	7.14	-8.817	-1233.13
Nis.98	28	-6.563	-8.655	-7.977	882.45
May.98	29	-4.969	6.351	-6.837	-811.974
Haz.98	30	-5.106	2.057	-6.997	-276.397
Tem.98	31	-5.382	2.563	-8.657	-611.821
Ağus.98	32	-8.429	.769	-3.517	-3004.24
Eyl.98	33	-3.635	8.474	-.177	-9681.67
Ek.98	34	10.659	10.38	4.263	-14403.1
Kas.98	35	13.882	4.786	7.803	-6859.51
Ar.98	36	14.416	-1.608	10.643	-2275.94
Oc.99	37	11.279	-19.303	13.383	4016.638
Şub.99	38	9.913	-14.497	13.322	15796.21
Mar.99	39	15.566	-9.191	10.962	26541.79
Nis.99	40	22.82	-9.085	8.502	37816.37
May.99	41	24.364	-1.479	8.142	45411.94

Burada zamanın trend etkisinin giderildiğini göstermesi bakımından örneğin detrend edilmiş ücret değişkeni ücret', detrend edilmiş sanayi indeksi değişkeni sanayi',...,biçiminde gösterilmiştir. Değişkenler bu şekilde zamanın etkisinden arındırıldıktan sonra çok değişkenli regresyon analizine gidilmiştir. Burada ücretliler geçinme indeksi bağımlı, diğer değişkenler ise bağımsız değişkenler olarak ele alınmıştır. Elbette böyle bir modelin kurulmasında herhangi bir iktisadi yasallıktan değil, sezgilerden ve bazı istatistiklerden yola çıkılmıştır. Kurulan regresyon modeli ile ilgili istatistiklere bakıldığında modelin biçimlendirilmesi ile ilgili hata yapıldığı görülecektir. (Durbin-Watson d=0.759;Serisel Korelasyon=0.62) Bu durum Box-Cox dönüşümleri ile modelin yeniden biçimlendirilmesi için iyi bir neden olarak düşünülmüştür.

Dönüşüm yapılmadan önce not edilmesi gereken son bir nokta daha vardır. Bağımlı değişkenin(ücret') aldığı değerlere bakıldığında bunlardan bir kısmının sıfırdan küçük olduğu görülecektir. Daha önceden de değinildiği gibi bu durumda Box-Cox dönüşümlerinin bir diğer versiyonunun kullanılmasını zorunluluğu bulunmaktadır. Bunu sağlayabilmek için (bağımlı değişkenin aldığı

değerlerin minimumunun -19.375 olmasından hareketle) bağımlı değişkenin her bir değerine 19.5 eklenerek yeni bir seri elde edilmiş ve bu serinin değerleri bağımlı değişkenin gözlem değerleri olarak düşünülmüştür.Daha sonra ise dönüşümü niteleyen lambdaya 25 değer verilmiş ve kalıntı kareleri toplamının minimize edilmesine çalışılmıştır. Elde edilen sonuçlar ve bazı istatistikler şöyledir:

Tablo-III:Uygulanan Dönüşümler ve Elde Edilen Bazı İstatistikler:

Lambda Değeri	Kalıntı Kareleri Ortalaması	Durbin-Watson d	Serisel Korelasyon	Açıklama
-3				Model istatistiksel olarak anlamlı değil (M.İ.O.A.D.)
-2.75				M.İ.O.A.D.
-2.5				M.İ.O.A.D.
-2.25				M.İ.O.A.D.
-2				M.İ.O.A.D.
-1.75				M.İ.O.A.D.
-1.5				M.İ.O.A.D.
-1.25				M.İ.O.A.D.
-1				M.İ.O.A.D.
-0.75				M.İ.O.A.D.
-0.50				M.İ.O.A.D.
-0.25	0.783	1.05	0.47	
0	0.38	0.7	0.65	
0.25	1.15	0.511	0.748	
0.5	1.7	0.487	0.763	
0.75	2.78	0.58	0.71	
1	4.89	0.759	0.62	
1.25	9.28	0.95	0.52	
1.5	18.98	1.097	0.43	
1.75	46.28	1.25	0.345	
2	95.76	1.14	0.373	
2.25	228.8	1.08	0.383	
2.5	558.95	1.01	0.4	
2.75	1383.98	0.934	0.425	
3	3454.4	0.867	0.445	

Sonuç

Yukarıdaki tabloda, olası bir biçimlendirme hatasının göstergeleri olan Durbin-Watson d , ve serisel korelasyon istatistiklerine bakıldığında en optimum çözümün lambda=1.75 ile lambda=2 aralığında olduğu görülecektir. Şüphesiz bu aralık daha ayrıntılı bir biçimde taranarak optimuma ulaşmak sözkonusu olabilir. Yukarıdaki tabloya bakıldığında; lambda değerlerinin simgelediği dönüşümler kalıntı kareleri ortalamasını minimum kılan dönüşümler olmaktan uzaktır. Dolayısıyla Box-Cox dönüşümünün verdiği çözüm ile modelin biçimlendirilme hatası ile ilgili istatistiklerinin vereceği çözümler birbirini tutmamaktadır. Bu oldukça önemli bir sorundur. İkinci olarak vurgulanması gereken sorun, lambdanın alacağı optimum değerin pratiğe tercüme edilmesinde karşılaşılması muhtemel güçlüktür. Sözgelimi bir an için lambda=1.78 gibi bir sonucun optimum sonuç olduğu varsayılsın. Böyle bir sonuç istatistiksel olarak son derece anlamlı olmasına karşılık, araştırmanın yapıldığı bilim dalı ya da disiplin açısından fazla bir anlam ifade etmeyebilir. Bu da daha basit, ama yorumlanması daha kolay modellere geri dönülmesine neden olacaktır.