Ekonomi-tek Volume / Cilt: 4 No: 2 May / Mayıs 2015, 1-23

İktisat Araştırmalarında Sorgulayıcı Veri Çözümlemesi

Ümit Şenesen[*]

Özet

Sorgulayıcı veri çözümlemesi (SVÇ) yöntemleri iktisat araştırmalarında kullanılan verilerin içerdiği yapıyı önceden inceleyerek anlamaya yarar. Böylece, kurulan modellerin daha uygun yapıda olmasını sağlar. Buna karşın, günümüzde ne yazık ki pek yaygın kullanılmamaktadır. Bu eksiklik, bulguların yanlış yorumlanmasına neden olmaktadır. Bu yazıda çok kısa bir tarihçeden sonra SVÇ'nin ne anlama geldiği açıklanıp belli başlı SVÇ araçları tanıtılacaktır. Bu araçların araştırmalarda nasıl uygulanması, sonuçların nasıl yorumlanması gerektiği iki özgün uygulama örneği eşliğinde açıklanmaktadır.

JEL Kodları: C10, C81, C82

Anahtar kelimeler: Sorgulayıcı veri çözümlemesi, dal-yaprak, beşli-özet, kutu-çizim, serpilme çizimi, lowess ya da loess eğrisi.

[*] İstanbul Teknik Üniversitesi emekli öğretim üyesi. usenesen@gmail.com.

Exploratory Data Analysis in Economic Research

Abstract

Exploratory data analysis (EDA) provides insights into data patterns that can be used in economic research. Its advantage lies in its ability to produce excellent model structures. Unfortunately, EDA methods have become neglected, being rarely used nowadays in practice. As a result, many cases in the empirical research now contain misleading inferences. This paper offers a brief history, followed by a description of EDA and its major tools. Two examples help to explain the use of these tools in research and the interpretation of the findings they yield.

JEL Codes: C10, C81, C82

Keywords: Exploratory data analysis, stem-and-leaf, five-point summary, box-plot, scatter diagram, lowess or loess curve.

1. Giriş

Sorgulayıcı veri çözümlemesi (SVÇ) (exploratory data analysis) araçlarının son yıllarda giriş düzeyindeki istatistik kitaplarında bile (sözgelimi Newbold, vd., 2013, ss. 26-27 ve ss. 49-51) yer almasına karşın, ne yazık ki pek çok iktisatçı tarafından hâlâ kullanılmamaktadır. Bu durum, eksik ya da yanlış sonuçlara varılmasına, belki de kötüye kullanımlara yol açmaktadır. Bu konuda ilk uyarılardan biri 40 yılı aşkın bir süre önce "Anscombe dörtlüsü" diye ün kazanan bir veri kümesiyle çok güzel örneklenmiştir (Anscombe, 1973; bu konuda Türkçe bir yazı için bkz. Şenesen, 2011). Yine aynı dönemde nicel yöntemleri öğrencilerine daha iyi anlatabilmeyi dert edinen bir öbek bilim insanı, Mosteller'in başkanlığında dört ciltlik bir kitap yayımlamıştır (Mosteller, vd.. 1973a, 1973b, 1973c, 1973d). Ama deyim yerindeyse bu işin asıl "kitabını yazan" John W. Tukey, daha 1960'ların başlarında SVÇ yöntemlerini, arabayı çeken öbür at olarak doğrulayıcı yaklaşımın yanına eklemiş (Tukey, 1962), bu konuda çığır açan kitabıyla aşağıda göreceğimiz yepyeni kimi araçlar türetmekle kalmamış, bu alanda düşünce yapısını kökten değiştirmiştir (Tukey, 1977).

Tukey'in bu ünlü sorgulayıcı veri çözümlemesi kitabını, aynı yıl Mosteller ile birlikte, veri çözümlemesini bağlanımla buluşturdukları yapıt izlemiş (Mosteller ve Tukey, 1977), birkaç yıl sonra kendilerine Hoaglin'in de katılmasıyla sorgulayıcı çözümleme, berk olma özelliğiyle yoğrulmuştur (Hoaglin, vd. 1983). Hemen ardından çizelgeler, genel eğilimler ve biçimler de bu çözümlemenin içine karılmıştır (Hoaglin, vd. 1985). Bir yanlış anlamayı önlemek için Tukey, "ne sorgulayıcı ne de doğrulayıcı çözümleme tek başına yeterlidir. Birini öbürünün yerine koymaya kalkışmak deli saçmasıdır. Her ikisi de gereklidir" diye bir uyarı yapmaktan da geri durmamıştır (Tukey, 1980).

Özet bilgiler içeren bir başka kitapçık (Hartwig ve Dearing, 1979) giriş düzeyinde açıklamalarla konuyu toplumsal bilimlere genişletmiştir. Velleman ile Hoaglin (1981) BASIC diliyle hazırladıkları programları da ekleyerek Tukey'in geliştirdiği yeni yöntemleri kolayca kullanılabilir hale getirmiştir. Daha sonra sadece bu konuyu kapsayan üniversite ders kitapları da yazılmıştır (Sözgelimi Marsh, 1988). SVÇ alanının sayısal bilgiyi görselleştirme yönüne vurgu yapan Cleveland, (1985, 1993) "lowess" ya da "loess" eğrisi gibi önemli katkılarda bulunmuştur. Tufte'nin efsaneleşmiş yapıtları (1983, 1990, 1997, 2006) bunları taçlandırmıştır.

2. Sorgulayıcı Veri Çözümlemesi

SVÇ, veriye yaklaşıma yeni bir boyut getirmiştir. Tukey'in öncülüğünde, daha önce yapılan varsayımları en aza indirgeyip uygun yöntemlerin seçilme-

sinde verinin yol gösterdiği bir düşünce yapısı geliştirmiştir. Tukey'in kitabının önsözüne koyduğu temel ilke şudur: "Bir şeyi ne kadar İYİ YAPTIĞINI ölçmeye kalkışmadan önce ne YAPABİLECEĞİNİ anlamak önemlidir." SVÇ, verinin dediklerine kulak vermektir. Kullanılan yöntemlerin anlaşılması da kullanılması da kolaydır. Basit aritmetikle, kolayca çizilen gösterimlerle veri kümelerinin dile gelmesi sağlanır. Sözgelimi, bir veri kümesinin logaritması alındığında kabaca bakışık hale geldiğini söylemek, ham verinin sola çarpık olduğunu belirtmekten daha iyi bir tanımlamadır. Çizimler, zaten bildiklerimizi göstermenin ötesine geçmeli, bizi görmeyi hiç beklemediğimiz özellikleri görmeye zorlamalıdır (Tukey (1977, ss. v-vi). Her araştırmacı toplumsal bilim kuramlarını sınamak için toplanmış verileri, kullanmadan önce iyice tanımalıdır (Hartwig ve Deamimg, 1979, s. 5). SVÇ araçları burada devreye girip daha sonra çok işe yarayabilecek değerli ipuçları verebilir. Her araştırmada önce bu araçları kullanarak işe başlamak doğru olur. Bu nedenle bu tekniklere *ön çözümleme* adı da verilmektedir (Chatfield, 1986).

"Sorgulayıcı veri çözümlemesi, sayısal dedektifliktir. … Bir suçu araştıran biri hem araçlara gerek duyar, hem durumu anlaması gerekir. Eğer elinde kullanabileceği parmak izi tozu yoksa pek çok yüzeyde aradığı izleri bulamayacaktır. Eğer suçlunun nerelerde parmak izi bırakabileceğini kestiremiyorsa bu kez de doğru yerlere bakmayacaktır. Veri çözümleyicisi de aynı durumdadır; hem araçları olması, hem olayı anlaması gerekir." (Tukey, 1977, s. 1). Bulunan ipuçları mahkemede kılı kırk yararak değerlendirilmelidir. Bu da doğrulayıcı çözümlemenin alanına girer.

SVÇ'nin dört önemli öğesi öne çıkar (Vellemean ve Hoaglin, 1981, s. xv):

- *Gösterimler*: Verinin davranışını açığa çıkarır, çözümlemenin yapısını ortaya koyar.

- *Kalıntılar*: Verinin çözümlemeyle açıklanamadan kalan bölümüne odaklanır.

- *Dönüştürmeler*: Logaritma, kare, karekök gibi kolay matematik işlemleri ile verinin davranışını yalınlaştırıp yapısının anlaşılmasını sağlar.

- *Berklik*: Az sayıda olağandışı gözlemin çözümleme sonuçlarını saptırmasını önler.

Bu yazıda ele alınan uygulama örneklerinde kullanacağımız belli başlı SVÇ araçları aşağıda kısaca açıklanmıştır. Bunlar elle kolayca hazırlanabileceği gibi, bazı istatistik paket programları kullanılarak da elde edilebilir. Gerek bu araçlara gerek bu yazıda adı geçmeyen araçlara ilişkin Türkçe ayrıntılı açıklamalar için Şenesen'e (2004) bakılabilir.

Dal-yaprak (Stem-and-leaf)

Sayısal veri kümesini incelemenin belki de ilk adımıdır. Çok kolay düzenlenen bu gösterim, veride daha önce fark edilmemiş, beklenmedik yapıyı ya da örüntüyü yansıtır. Şu soruları yanıtlamayı amaçlar: Gözlem değerleri nerelerde yoğunlaşmıştır? Verilerin yayılma aralığı ne kadardır? Küme bakışık mıdır, çarpık mıdır? Ne kadar çarpıktır? Veriler basık mı dağılmaktadır, yoksa sivri mi? Veri kümesi tek tepeli midir, yoksa çok tepeli mi? Gözlem değerleri arasında belirgin boşluklar var mıdır? Verilerden herhangi biri ya da bir kaçı ötekilerden açık ara uzak mıdır? Bu özellikleriyle sıklık çizimine (histograma) benzediği söylenebilir ama verileri belli değer aralıklarında dikdörtgen kutulara tıkmak yerine her birinin sayısal değerini, kısmen de olsa, korumaya özen gösterir. Aşağıda verilen ilk uygulama örneğinde yoğun biçimde kullanılacaktır.

Beşli Özet (5-point-summary)

Dal-yaprak gösterimi düzenlendikten sonra beşli özeti bulmak çocuk oyuncağıdır. İki uç değer, iki dördebölen ile ortanca, büyüklük sırasına göre alt alta yazılırsa beşli özet oluşur. Dal-yaprak bütün verileri büyüklüklerine göre sıraladığından bu iş kolayca yapılabilir. Eğer dal-yaprak çizimi elle yapılmışsa sıklık sayılarını her iki uçtan başlayarak birikimli sıklık olarak yazmak işi biraz daha kolaylaştırabilir (Tukey, 1977, ss. 30-35). Bu çizim, veri kümesini, her biri verinin dörtte birini içerecek biçimde dörde ayıran gözlem değerlerini yansıtır. Bu da bakışımın ya da çarpıklığın derecesini gözler önüne serer. Verinin alt yarısı ile üst yarısının, hatta her çeyreğinin yayılımlarının karşılaştırmasını sağlar. İlk uygulama örneğimizde kullanılacaktır.

Kutu Çizim (Box-plot)

Beşli özetle bulunanların göze yönelik biçimde sergilenmesi için bu çizim biçilmiş kaftandır. İki dördebölen bir dikdörtgen gibi birleştirilip arasına ortanca değer işaretlenir, dışadüşen yoksa, alt dördebölen en küçük, üst dördebölen en büyük değere bağlanırsa kutu çizim gösterimi tamamlanır. Veri kümesinde dışadüşen(ler) varsa bu(nlar) iki uç çizgisine değmez, ayrık işaretlenir. Beşli özetin sağladığı yararların daha görsel biçimde sunulmasıdır. Verinin bütünündeki ya da ortada kalan yarısındaki bakışım ya da çarpıklık derecesinin bir bakışta anlaşılmasını sağlar. Bu gösterim de ilk uygulama örneğimizde uygulanacaktır.

Lowess ya da Loess Eğrisi

Cleveland'ın (1985, ss. 167-178; 1993) geliştirdiği bu eğrinin adı, *yerel tartılandırılmış serpilme çizimi düzleyicisi* (*locally-weighted scatterplot smoother*) anlamına gelen terimin kısaltılmış biçimidir. Dışadüşenlerden etki-

lenmeyen bu berk eğri iki değişken arasındaki ilişkinin biçimini gösterir. Hesaplanması epeyce uzun olan bu eğrinin çizimi için yazılan öze bir bilgisayar izleci bazı bilgisayar paketlerinde yer almaktadır. İkinci uygulama örneğimizde nasıl kullanıldığı gösterilecektir.

3. Uygulama Örneği 1 - İllere Göre Kişi Başına GSYH Dağılımı

Bu uygulamada TÜİK'in Türkiye'de illere göre kişi başına gayrisafi yurtiçi hasıla (GSYH) verilerini kullanacağız. 1975-2000 arasındaki yıllarda bütün illerin 1987 fiyatlarıyla kişi başına GSYH değerlerini gösteren bu veriler, Çizelge 1'de yer almaktadır ve sayfaya sığdırmak için kırpılmıştır.

Çizelge 1. İllere Göre Kişi Başına GSYH Dağılımı (1987 Fiyatlarıyla)

İl adı	Plaka	1975	1976	1977	1978	1979	…	1998	1999	2000
ADANA	1	1163	1339	1326	1343	1237	…	2107	1981	1933
ADIYAMAN	2	418	527	565	558	597	…	808	768	793
AFYON	3	771	855	871	851	765	…	993	956	1024
AGRI	4	424	490	486	475	481	…	332	368	350
AMASYA	5	869	902	887	884	759	…	1219	1181	1149
ANKARA	6	1279	1316	1407	1417	1371	…	2375	2224	2398
ANTALYA	7	1031	1023	1024	1031	1149	…	1954	1758	1724
ARTVİN	8	849	934	992	1016	857	…	2275	1691	1551
AYDIN	9	1081	1277	1285	1284	1088	…	2038	1772	1905
BALIKESiR	10	1193	1262	1209	1214	1121	…	1723	1575	1668
BiLECiK	11	1119	1216	1280	1233	1422	…	2942	2914	2956
BiNGÖL	12	416	504	463	444	411	…	445	463	452
BiTLiS	13	495	561	592	563	412	…	416	415	362
BOLU	14	1004	1034	1117	1106	990	…	1786	1750	3172
…	…	…	…	…	…	…	…	…	…	…

Böyle büyük bir çizelgenin tamamını inceleyip içerdiği verilerin herhangi bir örüntüsünü görmek neredeyse olanaksız denecek kadar güçtür. Bazı sorgulayıcı veri çözümleme araçlarıyla bu güçlüğün nasıl aşılabileceği anlatılmaya çalışılacaktır. Önce ilk yılı ele alıp dal-yaprak gösterimini düzenleyelim.

Çizim 1'de yıl ile il sayısını belirten başlığın altında yaprak değerlerinin ölçü birimi yer almıştır. Buna göre, ikinci sütundaki sayılar "dal" değerini 1000 TL, onun sağında yer alan her sayı bir ilin "yaprak" değerini 100 TL birimleriyle temsil eder. Dal değerleri kendi içlerinde büyüklük sırasındadır. Benzer biçimde bir satırın yaprak değerleri de büyüklük sırasına dizilmiştir.

Böylece bir satırın dal değeri ile ilk yaprak değeri birleştirilirse o yıl bir ilimizin kişi başına GSYH değeri bulunur. İşe dal-yaprak çiziminin hazırlanması yönünden bakıp, Çizelgedeki ilk il olan Adana'yla başlayalım.

Çizim 1. İllerimizin 1987 Fiyatlarıyla Kişi Başına GSYH Dağılımı, 1975

```
1975  N  = 67
Yaprak = 100 TL

 12   0  444444455555
 30   0  666666677777777777
(13)  0  8888888999999
 24   1  0000000011111
 11   1  22222
  6   1  445
  3   1  6
  2   1
  2   2
  2   2
  2   2  4
  1   2  7
```

Çizelge 1'de 1975 yılı kişi başına GSYH değeri 1163 TL'nin dal değeri 1 (1000TL), yaprak değeri 1 (100 TL)'dir. Onlar ve birler basamaklarındaki 63 TL göz ardı edilerek bu değer, dal-yaprak çiziminde üstten dördüncü satırda dal kısmında 1 (yani 1*1000 TL), yaprak kısmında 1 (yani 1*100 TL) ile yer alır. Çizelgenin ikinci ili Adıyaman'ın 418 TL olan kişi başına GSYH değeri de aynı yolla dal-yaprak gösteriminin ilk satırının dal kısmına 0, yaprak kısmına 4 yazılarak belirtilir. Bütün iller tamamlandığında dal-yaprak çizimi büyük ölçüde hazırdır.

Şimdi ilk satırdaki illeri sayıp en soldaki sütuna 12 olarak geçirelim. İkinci satırdaki il sayısını buna ekleyerek (12+18=) 30 elde edip ilk sütuna işleyelim. Üçüncü satıra geldiğimizde, ilk üç sütundaki toplam il sayısı (30+13=) 43'ün, o yıldaki toplam il sayısı 67'nın yarısını aştığını görürüz. Bu durumda bütün illerin *ortanca* kişi başına GSYH değeri bu satırdaki illerden biridir. Bu nedenle o satırdaki ilk sütuna birikimli il sayısı değil, sadece o satırdaki il sayısı yazılır. Bunun öbür satırlar gibi birikimli bir sayıyı göstermediğini belirtmek için de 13 sayısı ayraç içine alınır. Sonra öbür uca gidilerek birikimli il sayılarının sol tarafa yazılması sürdürülür.

Bu dal-yaprak bize 1975 yılında illerimizin 1987 fiyatlarıyla kişi başına GSYH değerlerinin nasıl *dağıldığını* gösterir. Önce bu dağılımın *merkezi değeri*nden başlayalım. Aritmetik (ya da başka herhangi bir) ortalama hesaplama zahmetine girmeden ortanca bulmak burada çok kolaydır. 1975 yılında 67 ilin ortancası 34. il olduğuna göre, üçüncü satırda 8 ile gösterilen illerden dördüncüsünün değeri (1987 fiyatlarıyla yaklaşık 800 TL), ortanca değerdir. Bu bilgi

dal-yaprak çiziminden kolayca okunabilir. Bu 800 TL, diğer yılların ortancalarıyla karşılaştırılarak yıllar boyunca ortanca ilin kişi başına GSYH değerinin nasıl geliştiği, başka bir deyişle bu gösterge bakımından Türkiye ekonomisinin nasıl büyüdüğü anlaşılabilir.

Bir sonraki adım dağılımın *yayılım*ı konusunda fikir sahibi olmaktır. Bunu da en büyük ve en küçük değerlerin farkını alarak yapabiliriz. Dal-yaprak çiziminden (çizelgemizin tersine) bu iki uç değeri (küsurları atılmış olarak) hemen okuyabiliriz: 2700 ve 400 TL. Aradaki fark, dağılımın uç değerleri arasındaki yayılımın ölçüsüdür: 2300 TL. *Aralık*, uç değerlere karşı duyarlı olduğu, başka bir deyişle *berk* olmadığı için başka bir yayılım ölçüsü de deneyebiliriz. İki uçtan aynı sayıda gözlem dışlanırsa geriye kalanın yayılımı, dağılımın orta bölgesinin yayılımını gösterir. Bunun için 67 ilin alt ve üst *dördebölen*lerini saptamak gerekir. Bunların konumu da alttan ve üstten (67+1)/4 = 17'dir. İki uçtan da 16 ili dışlayıp, *dördebölenler aralığı*nı (*interquartile range*) şöyle hesaplayabiliriz. En düşük 17. değer 6 (yaklaşık 600 TL) ile en yüksek 17. değer 10 (yaklaşık 1000 TL) arasındaki fark yaklaşık 400 TL bize dördebölenler aralığını verir. İster aralık, ister dördebölenler aralığı ya da her ikisi birden öbür yılların değerleriyle karşılaştırılırsa Türkiye'nin 25 yıl boyunca illeri arasında gelir dağılımının düzeldiği mi, yoksa bozulduğu mu gözlenebilir.

Yukarıda hesapladığımız en küçük, alt dördebölen, ortanca, üst dördebölen ve en büyük beş değeri bir arada Çizelge 2'deki gibi oluşturmuş oluruz:

Çizelge 2. 1975 Yılı Beşli Özeti, 100TL

En Küçük	4
Alt Dördebölen	6
Ortanca	8
Üst Dördebölen	10
En Büyük	27

Buradan da ilk dört dilimin aralıklarının kabaca benzer (200 TL dolayında) olduğunu görürüz. Yani illeri kişi başına GSYH değerlerine göre sıralayıp eşit sayıda il içeren dört dilime ayırırsak, bu dilimler içindeki iki ilin arasındaki en büyük düzey farkının 200 TL dolayında olduğu anlaşılır. Ama en yüksek dörtte birlik dilimin iki sınırı arasındaki farkın 1700 TL olması çarpıcı bir bilgidir. Her dilimin birbirine göre durumu bu basit özetle kolayca gözler önüne serilir.

Merkezi değeri ve yayılımı gördükten sonra dağılımın *biçimi*ne bir göz atalım. Burada bakmamız gereken iki önemli özellik dağılımın *çarpıklık* derecesi ile *tepe sayısı*dır. Dal-yaprak gösterimine bakar bakmaz dağılımın büyük değerlere doğru bir hayli uzadığını, yani sağa çarpıklığını, tepe sayısının da

tek olduğunu hemen görebiliriz. Her yılın çarpıklığı gözle karşılaştırılarak 1975 ile 2000 arasında iller arasında gelir dağılımı çarpıklığının gelişimi izlenebilir. Tepe sayısının artıp artmadığına bakılarak da illerimiz arasında herhangi bir kümeleşme, ayrışma ya da kutuplaşma olup olmadığı gözlenebilir.

İlgilenmemiz gereken son nokta dağılımda *dışadüşen* (*outlier*) olup olmadığıdır. Dışadüşen, veri kümesindeki öbür verilerden hayli uzakta çok küçük ya da çok büyük değerlerdir. Sorgulayıcı veri çözümlemesinde bunu anlamak için basit bir hesaplama yolu kullanılır. Az önce bulduğumuz dördebölenler aralığının bir buçuk katı alınarak bulunan değer, alt dörde bölenden çıkarılıp bir alt eşik, üst dördebölene eklenerek bir üst eşik oluşturulur. Bu eşiklerin dışına düşen her değer dışadüşen sayılır. Örneğimizde alt dördeböleni 6 (altıyüz TL), üst dördeböleni de 10 (bin TL) bulmuştuk. Bunlardan 1.5*(10 – 6) = 6 elde edilir. Bu 3'ten çıkarılırsa sonuç 0'dan küçük olur. Hiçbir ilin kişi başına GSYH değeri eksi olamayacağına göre alt dışadüşen yoktur. 6 + 10 = 16 olduğundan dal-yaprak gösteriminde 16'dan büyük olan 24 (2400 TL) ve 27 (2700 TL) ile gösterilen iller dışadüşen sayılır. Çizelge 1'e kısa bir göz atmayla bu illerin İstanbul ve Kocaeli olduğunu anlayabiliriz. Bu bilgiler sonraki yılların bulgularıyla karşılaştırılırsa dışadüşen illerimizin bu konumlarını sürdürüp sürdüremedikleri ya da yeni dışadüşen illerin ortaya çıkıp çıkmadığı anlaşılabilir.

Bütün bunları görebilmek için öbür yılların dal-yaprak çizimlerini düzenlemek gerekir. Çizim 2'de beşer yıl arayla bu gösterimler sunulmuştur. Burada, Çizim1'deki dal-yapraktan farklı olarak dışadüşen illerin kişi başına GSYH değerlerine, dal-yaprak gösteriminin içinde değil ayrık olarak altında yer verilmiştir.

Çizim 1'deki gösterimlerden beşli özetleri hesaplarsak Çizelge 3'ü elde ederiz.

Çizelge 3. İl Düzeyinde Kişi Başına GSYH Değerlerinin Beşli Özetleri, 1987 Fiyatlarıyla, (100 TL)

Yıl	1975	1980	1985	1990	1995	2000
En Küçük	4	3	2	2	2	3
Alt Dördebölen	6	5	5	7	7	8
Ortanca	8	8	8	10	11	11
Üst Dördebölen	10	11	13	15	16	18
En Büyük	27	35	41	39	40	43

Beşli özet değerlerinden yararlanarak kutu çizimleri kolayca düzenlenebilir. Aynı yıla ilişkin beş değer kutu çizimin gerekli konum noktalarını temsil eder. Çizim 3 bu kutu çizimleri sergilemektedir. Çizim 4'te her yıl için ayrı bir kutu

çizim düzenlenmiştir. Böylece beşer yıl atlayarak değil bütün seyrin izlenmesi olanaklı hale gelmiştir.

Çizim 2. İl Düzeyinde 1987 Fiyatlarıyla Kişi Başına GSYH Dağılımı, 1975 – 2000 (Beş Yıl Arayla)

```
1975  N  = 67

 12   0  444444455555
 30   0  666666677777777777
(13)  0  8888888999999
 24   1  0000000011111
 11   1  22222
  6   1  445
  3   1  6

Dışadüşen 24; 27
```

```
1980  N  = 67

  6   0  333333
 17   0  44455555555
 29   0  666666677777
(13)  0  8888889999999
 25   1  0000111111
 15   1  2222333
  8   1  4445
  4   1
  4   1  9

Dışadüşen 21; 21; 35
```

```
1985  N  = 67

  6   0  223333
 17   0  44445555555
 31   0  666666777777777
(10)  0  8889999999
 26   1  001
 23   1  2223333
 16   1  444455
 10   1  6666777
  3   1
  3   2
  3   2  2

Dışadüşen 25; 41
```

```
1990  N  = 71

  5   0  22333
 13   0  44555555
 23   0  6666777777
 29   0  899999
(14)  1  00000001111111
 28   1  2222333
 21   1  44455555
 13   1  7
 12   1  88999
  7   2  001
  4   2  33
  2   2  5

Dışadüşen 39

Dal    = 1000 TL
Yaprak =  100 TL
```

```
1995  N  = 76

  6   0  223333
 12   0  444555
 24   0  666677777777
 34   0  8888999999
(10)  1  0000111111
 32   1  2233333
 25   1  4455
 21   1  66777
 16   1  8888
 12   2  0011111
  5   2  3
  4   2  5
  3   2  6
  2   2  9

Dışadüşen 40
```

```
2000  N  = 81

  5   0  33333
 11   0  445555
 20   0  666677777
 31   0  88888899999
(10)  1  0000111111
 40   1  2233333
 33   1  4455
 29   1  66666777
 21   1  88899
 16   2  011
 13   2  3
 12   2  45
 10   2  6666
  6   2  89
  4   3  1

Dışadüşen 34; 37; 43
```

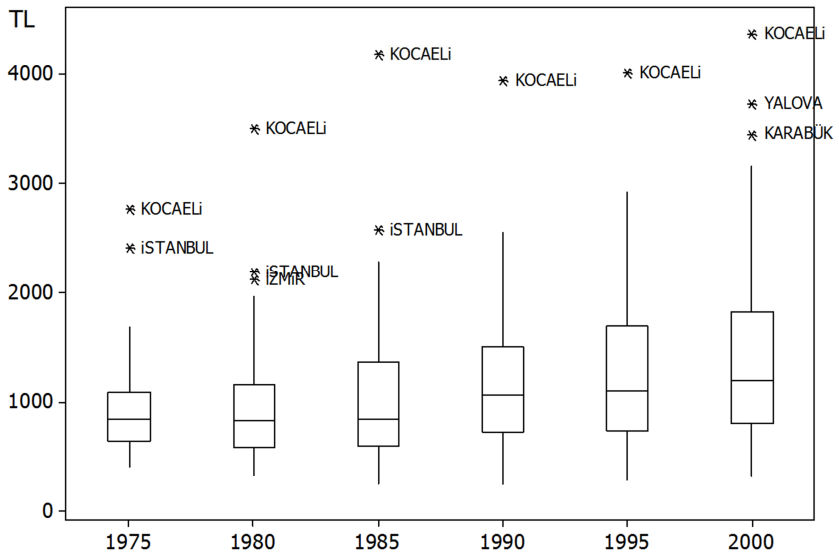
Bu gösterimlere bakarak özetle şunları söyleyebiliriz: 1975'te 800 dolayında L olan *ortanca* değer, 1985'e kadar bu düzeylerde dalgalanmış, 1990'da 1000, 2000'de de 1100 TL'nin üstüne çıkmıştır. Kısacası ilk on beş yıllık durakla-manın ardından son dönemde yükselme görülmektedir.

- 1975'te en yüksek ve en düşük kişi başına GSYH sahibi iki il arasındaki farkı gösteren 2300 TL dolayındaki *aralık* değerinin, beşer yıl arayla 3200, 3900, 3700, 3800, 4000 TL biçiminde oynadığı, bazen düşse de uzun dönemde büyüme eğilimi gösterdiği söylenebilir. 1975'te en yüksek ve en düşük kişi başına GSYH oranı 2300/300 ≈ 7–8 kat iken bu oran zamanla 14-15 katına kadar çıkmıştır. Yani iki uçtaki iller arasında gelir uçurumu artmaktadır.

- Ama bundan da vahim olanı en küçük değerlerin zaman içindeki seyridir. 1975'te 400 TL dolayında olan bu değer, 1980'e gelindiğinde 300 TL, 1985'te 200 TL düzeylerine düşmüştür. Yani en düşük kişi başına GSYH

değerine sahip ilin geliri on yıl içinde belki de yarıya inmiş; bu düzeyden ancak 2000 yılında yaklaşık 300 TL'ye yükselerek kurtulabilmiştir.

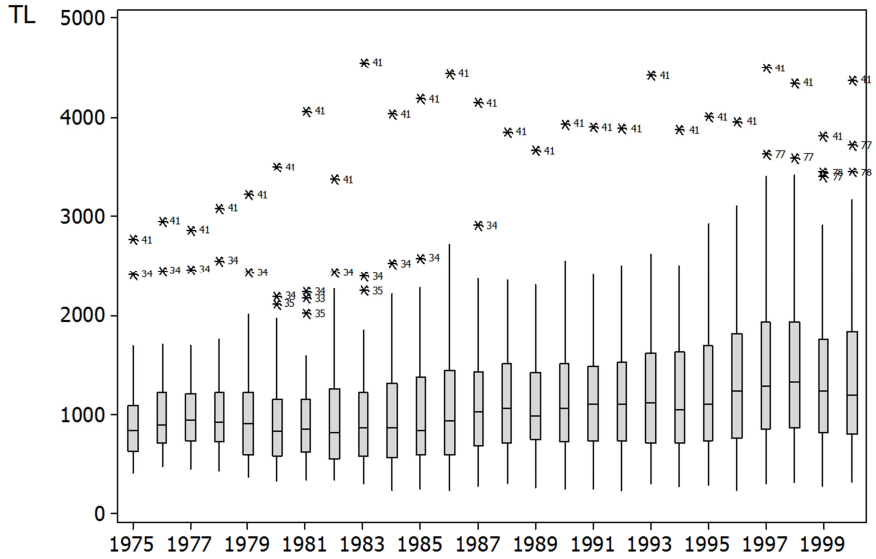
Çizim 3. İl Düzeyinde Kişi Başına GSYH Değerleri Kutu Çizim, Beş Yıl Arayla



- Aralık göstergesinin uçtaki değerlere olan aşırı duyarlılığı düşünülerek *dördebölenler aralığı*nın gelişimi incelenebilir. Başlangıçta 400 TL dolayındaki bu yayılımın beşer yıl arayla aldığı değerler, 600, 800, 900, 900, 1000 TL dolayında olmuştur. Yani illerimizin yarısını oluşturan ortadaki öbek içinde sabit fiyatlarla en düşük ve en yüksek kişi başına GSYH'ye sahip iller arasındaki uçurum 25 yılda oynamalarla kabaca %65 kadar büyümüştür.

- Gerek dal-yaprak gösterimleri, gerek kutu çizimleri bize 1975'te zaten çarpık olan iller arası gelir dağılımının zamanla çok daha fazla çarpıldığını göstermektedir. Daha da kötüsü, 1975'te tek olan tepe sayısının zamanla ikiye, belki de üçe çıkmasıdır. Yani illere göre gelir dağılımı gittikçe daha çarpık olmakla kalmıyor, aynı zamanda illerimiz kişi başına GSYH göstergesi bakımından kendi aralarında öbekleşiyor, kutuplaşıyor izlenimi vermektedir. Bu iki ya da üç öbeğe giren illerimizin hangileri olduğuna bakıldığında bugünkü siyasi çözümsüzlüğün belki de en önemli nedenlerinden biriyle karşılaşırız.

Çizim 4. İl Düzeyinde Kişi Başına GSYH Değerleri Kutu Çizimleri, 1975-2000



İllerimizin kişi başına GSYH değerlerinin her yıl büyük değerlere doğru yöneldiğini, yani sağa çarpık olduğunu gördük. Bu veriler daha bakışık hale getirilebilirse karşılaştırmalarımız bize neler gösterebilir? Bu amaçla dönüştürme uygulamak istersek ne yapabiliriz? Çizelge 4, bu soruya yanıt bulma konusunda yol göstericidir.

Bir veri kümesi, özgün değerlerinde bakışıksa herhangi bir dönüştürmeye gerek yoktur. Çarpıksa üstel dönüştürmeyle bakışık hale çevrilebilir. Bunun için özgün verilerin kuvveti alınır. Eğer veriler sola çarpıksa kare, küp gibi birden büyük kuvvetlerden biri kullanılır. Böylece sola doğru uzayan, yani küçük değerlerde seyrek olup sağa doğru gittikçe araları kapanıp sıklaşan değerler, diyelim kareleri alınarak daha bakışık konuma getirilebilir. Çünkü sola çarpık dağılımlarda küçük değerlerin kareleri arasındaki farklar ile büyük değerlerin kareleri arasındaki farklar birbirine yaklaştırılabilir.

Sözgelimi, 2 ile 5 arasındaki fark 3 iken 10 ile 11 arasındaki fark 1'dir. Yani küçük değerler daha seyrek, büyük değerler daha sıktır. Kareleri alınırsa $2^2 = 4$ ile $5^2 = 25$ arasındaki fark 21'dir. $10^2 = 100$ ile $11^2 = 121$ arasındaki fark da 21'dir. Öyleyse başlangıçtaki çarpıklık giderilip bakışım sağlanmıştır. Eğer çarpıklık sola doğruysa kullanılacak kuvvet de 1'den küçük olmalıdır. Buna da 2, 3, 10, 15 değerlerini içeren bir kümeyi örnek verelim. 2 ile 3'ün

farkı 1, 10 ile 15'in farkı 5'tir; yani küçük değerler daha sık, büyük değerler daha seyrektir. Bu sayıların sıfırıncı kuvvetlerini almak çözümsüzlük yaratır ama logaritma almak *limitte* üssü 0 almakla aynı sonucu verir. Öyleyse veri kümemizdeki değerlerin doğal logaritmasını alarak dağılımı bakışık hale getirmeye çalışalım. ln 2 = 0,6931, ln 3 = 1,0986'dır. Aradaki fark da 0,4055 kadardır. ln 10 = 2,3026 ile ln 15 = 2,7081 arasındaki fark da aynıdır. Burada dağılım bakışık hale gelmiştir.

Çizelge 4. Dönüştürme Merdiveni

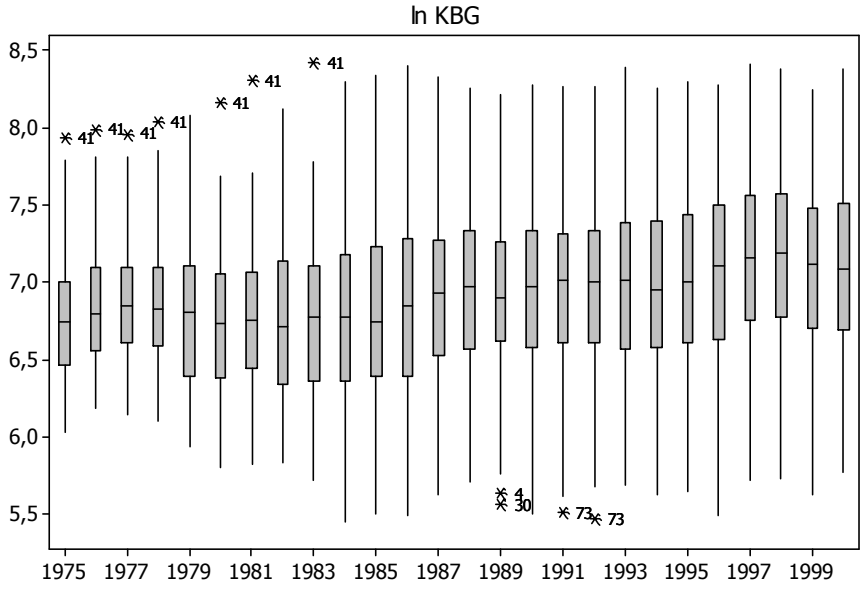
Çarpıklığın yönü	***r***	Y^r	**Dönüştürme**
Sola	3	Y^3	Küp
"	2	Y^2	Kare
"	3/2	$\sqrt{Y^3}$	3/2. kuvvet
Bakışık	1	Y	Yok
Sağa	1/3	$\sqrt[3]{Y}$	Küpkök
"	½	$\sqrt{Y}$	Karekök
"	0	log Y	Logaritma
"	– ½	$-1/\sqrt{Y}$	Karekökten sonra ters
"	– 1	$-1/Y$	Ters
"	– 2	$-1/Y^2$	Kareden sonra ters

Şimdi bu yöntemi kendi verilerimize uygulayalım. Beşli özetlerdeki değerlerin logaritmasını alıp da kutu çizimlerini düzenlersek Çizim 5'teki gösterime ulaşırız.

Buradan görüldüğü gibi gerek kutuların boyları (dördebölenler aralığı) gerek bütün yayılım aralığı ilk çizime göre bakışıma çok daha yakındır. Bu çizimdeki dikey uzunluklar, Çizim 4'teki gibi iki ilin aynı yıldaki kişi başına GSYH değerleri arasındaki farkı değil, bu iki değerin oranını gösterir. Burada büyüklük karşılaştırmaları yapılırken bunu unutmamak gerekir. Bir kutunun üst sınırıyla ortanca arasındaki uzaklık, ortanca ile kutunun alt çizgisi arasındaki uzaklığa eşitse, bu noktalarda yer alan illerin kişi başına GSYH değerleri arasındaki fark değil, bu değerlerin birbirine oranı aynıdır.

Öyleyse, 1980'lerin sonu ile 1990'ların başındaki birkaç yılda dışadüşen konumundaki illerin durumunu nasıl yorumlayabiliriz? Bu iller o yıllarda geri kalan illerden TL ile ölçülmüş fark olarak değil ama oran olarak çok geriye düşmüştür. Yani 1989'da Ağrı ile Hakkâri'nin, 1991 ve 1992'de ise Şırnak'ın kişi başına GSYH değeri öbür illerin yüzde olarak çok altındadır.

Çizim 5. Logaritma Dönüştürmesiyle Bakışık Hale Gelen Kişi Başına GSYH Değerleri



Burada ayrıntıları biraz göz ardı edip kaba büyüklükler kullanarak basit, kolay, hızlı yöntemlerle yaptığımız bu sorgulayıcı çözümleme örneği, bize bu araçların ne kadar kullanışlı, ne kadar yararlı olduğunu açıkça göstermiştir. Bundan sonraki veri çözümleme süreci burada elde edilen ipuçlarının izini sürüp doğruluklarını sınamaktır.

4. Uygulama Örneği 2 – Kişi başına gelir ile ortalama ömür ilişkisi

Bu örnekte 163 ülkenin 2014 yılında satınalma gücü eşleniği kişi başına gayrisafi milli gelirleri (KBG) (bin $) ile ortalama ömürleri (yıl) arasındaki ilişki Dünya Bankası verileri kullanılarak incelenecektir. Baştan ve sondan 10'ar ülkenin verileri Çizelge 5'te gösterilmiştir.

163 ülke verisi kullanılarak tahmin edilen bağlanım doğrusunun (regression line) bulguları aşağıda Çizelge 6'da verilmiştir. Buna göre en düşük gelirli ülkelerde bile ortalama ömür 66 yılın üzerindedir. Bir ülkenin kişi başına geliri 1000 $ arttığında bu ülkede ortalama ömür de 0,30843 yıl ya da yaklaşık 113 gün uzamaktadır. Katsayıların *p*-değerleri çok küçük olduğundan her biri anlamlı sayılmalıdır. Denklemin açıklama gücü 0,45 dolayında olup yatay kesit veri için birçok yazar tarafından yeterli görülebilir.

Çizelge 5. 163 Ülkeden Baştan ve Sondan 10'ar Ülkenin Satınalma Gücü Eşleniği Kişi Başına Gelir (KBG, bin $) ile Ortalama Ömür Değerleri (Yıl), 2014

Ülke	Ömür Yıl	KBG bin $		Ülke	Ömür Yıl	KBG bin $
Afganistan	60	2		ABD	79	56
Arnavutluk	78	11		Uruguay	77	20
Antigua-Barbuda	76	21		Özbekistan	68	6
Avustralya	82	45		Vanuatu	72	3
Avusturya	81	47		Venezuela, RB	74	18
Azerbaycan	71	17		Vietnam	76	5
Bahama	75	22		West Bank-Gaza	73	5
Barbados	75	15		Yemen	64	4
Belarus	73	18		Zambiya	60	4
Belize	70	8		Zimbabwe	57	2

Çizelge 6. 163 Ülke Verisi ile Tahmin Edilen Bağlanım Doğrusu (1)

Ömür (yıl) = 66,0 + 0,308 Kb gelir (bin $)

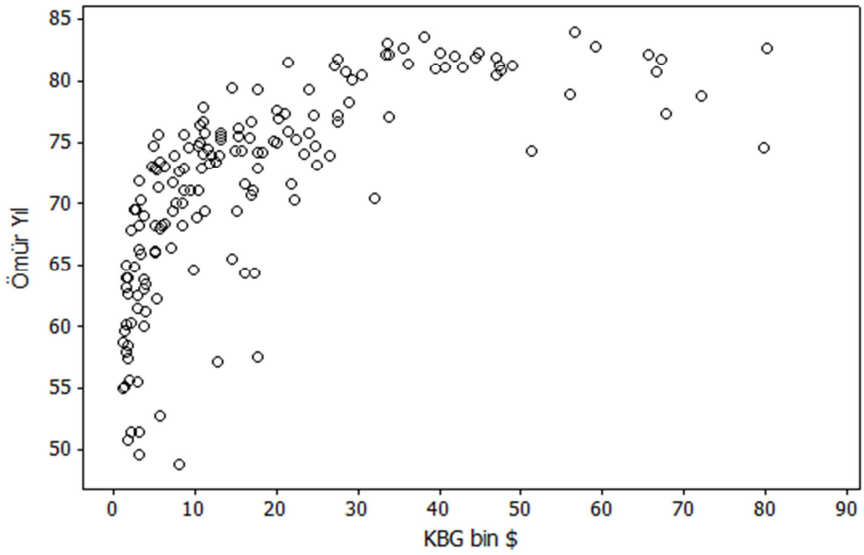
Değişken	Katsayı	StHata	*T*	*p*
Sabit	66,0375	0,6841	96,54	0,000
Kb gelir	0,30843	0,02683	11,50	0,000

S = 6,09731 R-kare = 45,1% R-kare(düz) = 44,7%

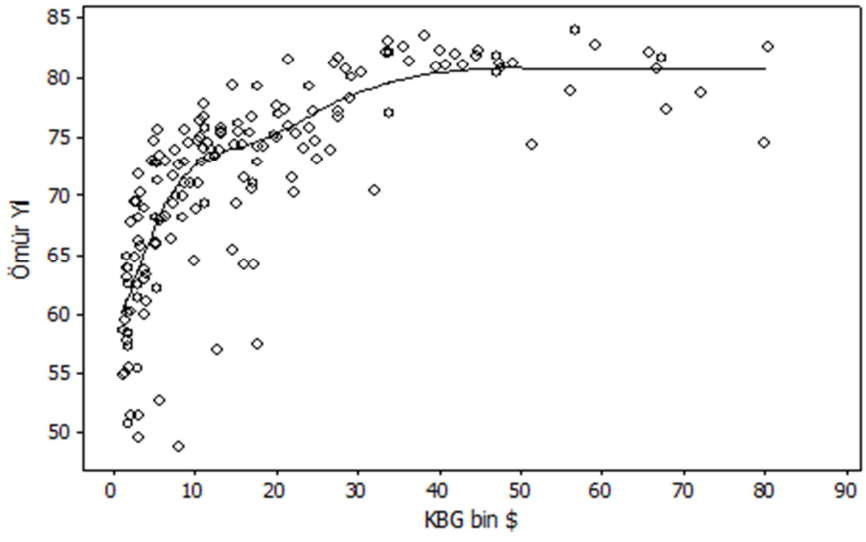
Acaba gerçek durum burada gösterilmek istendiği kadar tozpembe midir? Basit bir serpilme çizimi, hiç de öyle olmadığını yüzümüze vurabilir? Nitekim Çizim 5, bağlanım doğrusunun en önemli varsayımlarından biri olan doğrusallığın sağlanamadığının açık bir kanıtıdır. İlişki doğrusal değil eğriseldir. Dolayısıyla yukarıda dile getirilen yorumlar gerçeğin yakınından bile geçmemektedir. Yani tahmin edilen denklem hiçbir işe yaramamaktadır.

Düşük gelirli ülkelerde gelir çok az artsa bile ortalama ömrün hızla uzaması beklenirken, yüksek gelirli ülkelerde büyük gelir artışlarının bile ortalama ömür üzerinde pek sınırlı bir etki yaratabildiği, hatta kişi başına gelir 50 bin $ aştıktan sonra etkisini bütünüyle yitirdiği Çizim 6'daki *lowess eğrisi'*nden açıkça görülebilmektedir. Bu eğriye uyabilecek bağlanım modellerini aramadan, serpilme çiziminin açığa çıkardığı başka bir noktaya dikkat çekelim.

Çizim 5. KBG Değerleri ile Ortalama Ömür İlişkisinin Serpilme Çizimi, 2014

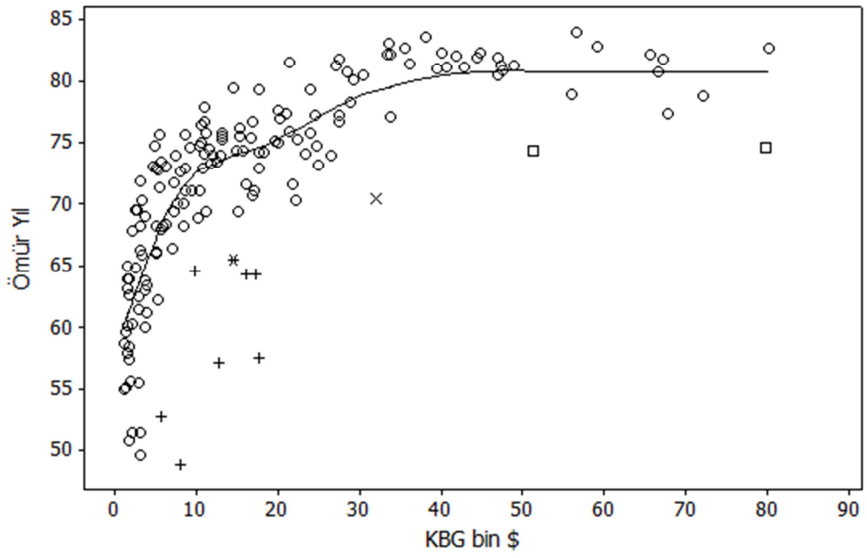


Çizim 6. KBG Değerleri ile Ortalama Ömür İlişkisinin Lowes Eğrisi, 2014



Serpilme çiziminde bazı ülkeler lowess eğrisinin epeyce altında kalmaktadır. Bunlar gelir ya da ömür değişkeni bakımından dışadüşen olmayabilir ama serpilme çizimine yansıyan ilişki eğrisinin dışadüşenleri olarak göze çarpmaktadır. Çizim 7'de ilişkinin dışında kalan bu ülkelerden + ile işaretlenen yedisi (Botswana, Ekvator Ginesi, Güney Afrika, Gabon, Namibya, Nijerya, Swaziland) Sahra-altı ülkeleri, * ile işaretlenen Türkmenistan, × ile işaretlenen Trinidad ve Tobago, □ ile işaretlenenler de Suudi Arabistan ile Kuveyt'tir. Bu ülkelerin hepsinde, gelir düzeylerine yakışmayacak kadar düşük ortalama ömür uzunlukları gözlenmektedir. Serpilme çizimi, verilerdeki bu sakat yapıyı da ortaya koymuştur. Lowess eğrisi berklik özelliği sayesinde bu ülkelerin dışadüşen konumundan etkilenmemektedir.

Çizim 7. Dünyadaki Gelir- Ömür İlişkisine Uymayan Ülkeler, 2014

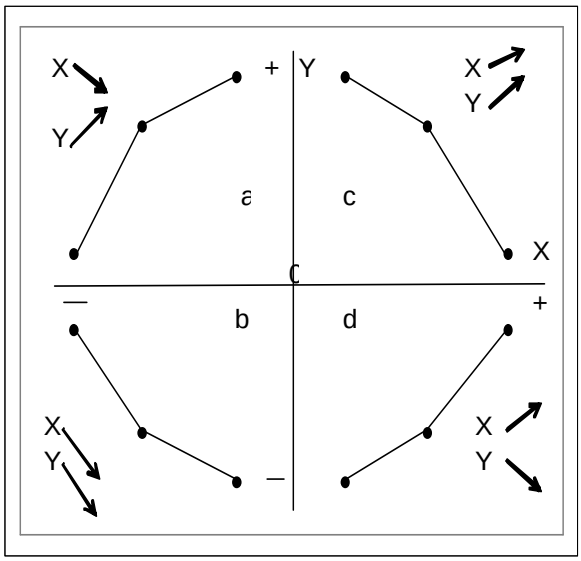


Şimdi yeniden fonksiyon kalıbına dönüp doğrusal olmayan bu ilişkiyi nasıl modelleyeceğimize bakalım. Buraya kadar gördüğümüz serpilme çizimlerinde lowess eğrisinin eğiminin hep artı işaretli olduğunu ve eğiminin önce yüksek olup gitgide azaldığını görebiliyoruz. Bu özellikler Çizim 8'de sol üst taraftaki gibi bir ilişki biçimiyle karşı karşıya olduğumuzu gösteriyor. Burada, kırık çizginin alt ucu *X* ekseninin – işaretli, üst ucu da *Y* ekseninin + işaretli yönüne doğru uzanmaktadır. Öyleyse eğrisel bir ilişki elde etmek için ya yatay eksendeki değişkenin birden küçük kuvvetlerini ya da dikey eksendeki değişkenin birden büyük kuvvetlerini denemeliyiz (Tukey, 1977, s. 198). Gerek

olursa her iki değişken birden dönüştürülebilir. Çizelge 4'te dönüştürmede kullanılabilecek kuvvetler sıralanmıştı.

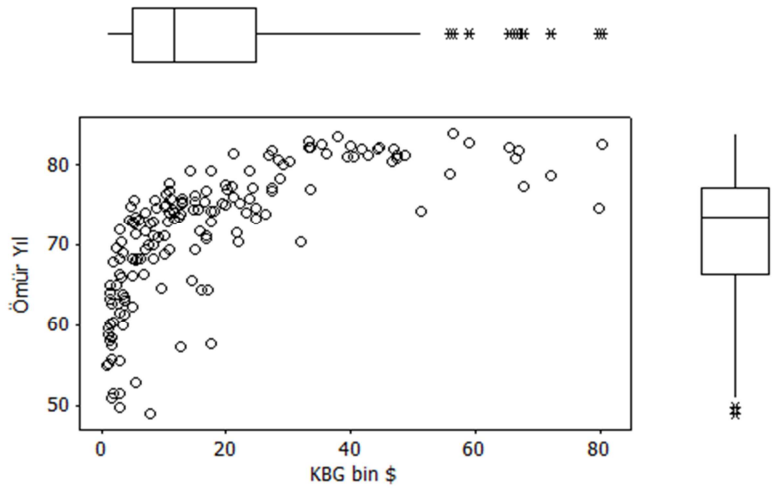
Hangisinden işe başlayacağımıza karar vermek için, serpilme çiziminin iki ekseninde iki değişkenin kutu çizimlerinin de yer aldığı Çizim 9'a bir göz atalım. Beklendiği gibi ülkeler temelinde dünyada sola doğru çok çarpık bir dağılım görülmektedir. Böyle gelir dağılımları genellikle logaritma (birden küçük üs) alınarak bakışık hale getirilebilir. Öbür değişken, ortalama ömrün çarpıklığı, ise gelire göre hem ters yönde hem daha azdır.

Çizim 8. Doğrusal Olmayan İlişkilerin Dönüştürülmesi İçin İpuçları

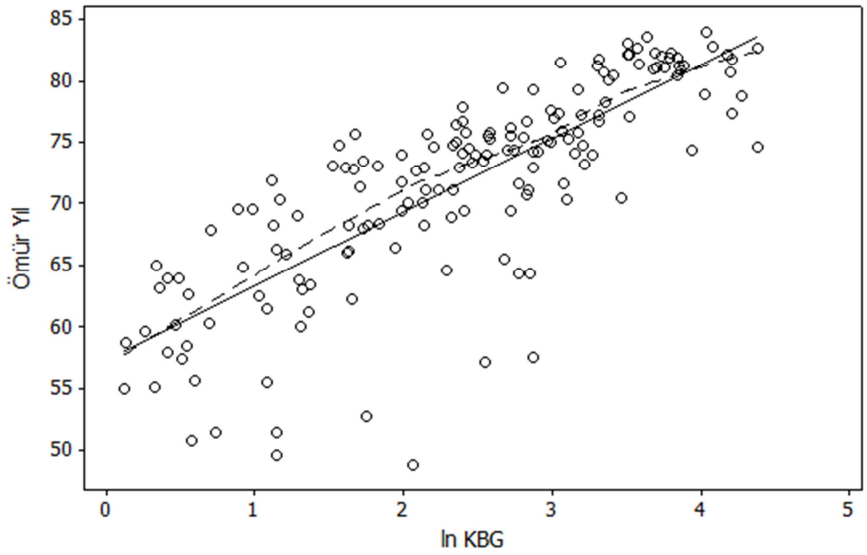


Bu durumda daha çarpık olan gelir değişkenini, logaritmasını alarak, dönüştürmekle işe başlayabiliriz. Dönüştürmeden sonraki lowess eğrisi (kesikli çizgi) ile en küçük kareler bağlanım doğrusunun birbirine epeyce yaklaştığını Çizim 10'da görebiliriz. Aralarındaki farkın, ortalama ömrü beklenenden düşük olan ülkelerin varlığı nedeniyle ortaya çıktığı düşünülebilir. Lowess eğrisi bunlardan etkilenmez ama dışadüşen gözlemler bağlanım doğrusunu kendilerine doğru çeker. Yine de orta bölgede görülen hafif farklılık dışında ilişkinin doğrusallığa çok yaklaştığı söylenebilir.

Çizim 9. Değişkenlerin Kutu Çizimlerini de Gösteren Serpilme Çizimi



Çizim 10. Ortalama Ömür ile ln KBG İlişkisinin Serpilme Görünümü



Ömür değişkenini dönüştürmeye gerek kalmadan hayli doğrusallaşmış bir ilişkiye ulaşabildik. Ayrıca ömür değişkeninin karesini, küpünü alarak başka dönüştürmeler de denenmiş ama daha iyi bir sonuç elde edilememiştir.

Bu durumda en küçük kareler bağlanım denkleminin tahmini Çizelge 7'de verildiği gibidir:

Çizelge 7. 163 Ülke Verisi ile Tahmin Edilen Bağlanım Doğrusu (2)

Ömür = 57,4 + 5,96 ln kb gelir

Değişken	Katsayı	StHata	*t*	*p*
Sabit	57,4162	0,9280	61,87	0,000
ln kb gelir	5,9647	0,3530	16,90	0,000

S = 4,94063 R-kare = 63,9% R-kare(düz) = 63,7%

Bu denklemin açıklama gücü bir önceki doğrusal denkleme göre hayli artmıştır. İlkinde 0,451 olan R^2, bu denklemde 0,639'a kadar yükselmiş, açıklama gücü %40'tan fazla artmıştır. Kalıntıların standart hatası 6 yılın üstündeyken şimdi 5 yılın altına inmiştir. Katsayıların *p*-değerleri yine çok küçüktür. Eğim katsayısının *t* değeri de yükselmiştir.

Yeni denkleme göre en düşük gelirli ülkelerde ortalama ömrün beklenen değeri 60 yılın biraz altındadır. Bir ülkede kişi başına gelir %1 arttığında ortalama ömrü uzatma beklentisi, bir yılın %6'sı kadar, yani yaklaşık 22 gün kadardır. Elbette bu artış düşük gelirli ülkelerde, zenginlere göre daha önemlidir. 4.000$ olan kişi başına gelirini 400$ (%10) arttırabilen bir ülkede ortalama ömrün yaklaşık 218 gün uzaması beklenirken, 40.000$ düzeyindeki bir ülkede ortalama ömrün aynı ölçüde uzaması için kişi başına gelirin 4.000$ yükseltilmesi gerekir.

Lowess eğrisinin ilk gösterildiği Çizim 6'ya dönüp de bu eğriye bir kez daha dikkatle bakınca şunu görürüz: Lowes eğrisi 10.000$ gelir düzeyine kadar doğrusal yükselmekte, 40.000$'ı aştıktan sonra da eğimi sıfıra yakın bir doğruya dönüşmektedir. 10.000-40.000$ gelir aralığında ise tam doğrusal olmasa da doğrusal yakın biçimde yükselmektedir. Bu özelliğe dayanarak hiç dönüştürme yapmadan, üç gelir düzeyindeki ülkeler için ayrı ayrı bağlanım doğruları bulmak olası mıdır? Yapay (dummy) değişkenler kullanarak bunu denediğimizde sonucun hiç de fena çıkmadığını görürüz. Aşağıda Çizelge 8'deki sonuçlara göre düzeltilmiş R^2 = 0,636, logaritma kullanıp bulduğumuz

dönüştürmeli denklemdeki düzeltilmiş R^2 = 0,637 ile neredeyse aynıdır, yani iki denklemin açıklayıcı gücü eşit denecek kadar yakındır.

Çizelge 8. 163 Ülke Verisi ile Tahmin Edilen Bağlanım Doğrusu (3)

Ömür = 58,5 + 1,50 KBG + 9,72 $D_{10\text{-}40}$ - 1,16 $D_{10\text{-}40}$*KBG

+ 25,7 D_{40+} - 1,56 D_{40+}*KBG

Değişken	Katsayı	StHata	*t*	*p*
Sabit	58,509	1,22	47,97	0,000
KBG	1,5	0,2418	6,2	0,000
$D_{10\text{-}40}$	9,718	1,996	4,87	0,000
$D_{10\text{-}40}$*KBG	-1,1586	0,2526	-4,59	0,000
D_{40+}	25,669	4,956	5,18	0,000
D_{40+}*KBG	-1,5632	0,2562	-6,1	0,000

S = 4,94721 R-kare = 64,7% R-kare(düz) = 63,6%

Katsayılarının *p*-değerleri çok düşük olan bu denklemle üç gelir düzeyi için üç ayrı denklem hesaplayabiliriz:

10.000$'ın altı: Ömür = 58,509 + 1,50 KBG

10.000 – 40.000$: Ömür = 58,509 + 9,718 + (1,50 – 1,1586) KBG

= 68,227 + 0,3414 KBG

40.000$'üstü: Ömür = 58,509 + 25,669 + (1,50 – 1,5632)KBG

= 84,178 - 0,5632 KBG

Buna göre en düşük gelirli ülkelerde ortalama ömür 59 yıl kadardır. 10.000'a kadar her 1.000$ gelir artışının ortalama ömrü artırma beklentisi yaklaşık 1,5 yıldır.

10.000-40.000$ arası ülkelerde en düşük ortalama ömür 69 yaşına yakındır. 40.000'a kadar her 1.000$ gelir artışının ortalama ömrü artırma beklentisi 0,3414 yıl ya da yaklaşık 4 aydır.

40.000 doların üstündeki ülkelerde ortalama ömür 84 yılın biraz üstünde başlamakta ama gelir artışlarıyla ortalama ömür uzamamaktadır, hatta denkleme bakılırsa her 1.000$ gelir artışı ortalama ömrü yaklaşık 23 gün kısaltmaktadır. Suudi Arabistan ile Kuveyt'in gelirlerine göre düşük ortalama ömre sahip olmasının bu azalmada ciddi payı vardır. Çünkü 40.000$ üstü kişi başına gelire sahip 20 ülke içindeki ağırlıkları yüzde ondur.

Son iki denklem aynı olguyu farklı biçimlerde ama hemen hemen aynı güçle açıkladıklarına göre her ikisinden de yararlanılabilir.

Kaynakça

Anscombe, F. J., (1973), "Graphs in Statistical Analysis," *American Statistician* **27** (1), ss. 17–21.

Chatfield, C., (1986), "Exploratory Data Analysis" *European Journal of Operational Research*, 23, ss. 5-13.

Cleveland, W. S., (1985), *The Elements of Graphing Data*. Wadsworth.

Cleveland, W. S., (1993), *Visualising Data*. Hobart Press, AT&T Bell Laboratories.

Hartwig, F. ve B. E. Dearing, (1979), "Exploratory Data Analysis" *SAGE University Paper*, No. 16.

Hoaglin, D. C., F. Mosteller ve J. W. Tukey, (1983), *Understanding Robust and Exploratory Analysis*. John Wiley.

Hoaglin, D. C., F. Mosteller ve J. W. Tukey, (1985), *Exploring Data Tables, Trends, and Shapes*. John Wiley

Marsh, C., (1988), *Exploring Data: An Introduction to Data Analysis for Social Scientists*. Polity Press.

Mosteller, F., W. H. Kruskal, R. S. Peters, G. R. Rising ve R. F. Link, (1973a), *Statistics by Data: Exploring Data*. Addison-Wesley.

Mosteller, F., R. S. Peters, W. H. Kruskal, G. R. Rising ve R. F. Link, (1973b), *Statistics by Data: Weighing Chances*. Addison-Wesley.

Mosteller, F., R. S. Peters, W. H. Kruskal, G. R. Rising ve R. F. Link, (1973c), *Statistics by Data: Detecting Patterns*, Addison-Wesley.

Mosteller, F., R. S. Peters, W. H. Kruskal, G. R. Rising ve R. F. Link, (1973d), *Statistics by Data: Finding Models*, Addison-Wesley.

Mosteller, F., J. W. Tukey, (1977), *Data Analysis and Regression*, Addison Wesley.

Newbold P., W. L. Carlson, B. M. Thorne, (2013), *Statistics for Business and Economics*, Pearson.

Şenesen, Ü., (2004), *İstatistik: Sayıların Arkasını Anlamak*, Literatür Yayınevi.

Şenesen, Ü., (2011), "İktisatta Nicel Yöntemlerin Kötü(ye) Kullanılması: Örnekler, Öneriler", E. Uygur (Ed.) *Küresel Bunalım ve İktisat Eğitimi* içinde, ss. 11-127.

Tufte, E. R., (1983), *The Visual Display of Quantitative Information*, Graphic Press.

Tufte, E. R., (1990), *Envisioning Information*, Graphic Press.

Tufte, E. R., (1997), *Visual Explanations*, Graphic Press.

Tufte, E. R., (2006), *Beatiful Evidence*, Graphic Press.

Tukey, J. W., (1962), "The Future of Data Analysis" *The Annals of Mathematical Statistics*, 33 (1), ss. 1-67.

Tukey, J. W., (1977), *Exploratory Data Analysis*, Addison Wesley.

Tukey, J. W., (1980), "We Need Both Exploratory and Confirmatory" *The American Statistician*, 34 (1), ss. 23–25.

Velleman, P. F. ve D. C. Hoaglin, (1980), *Applications, Basics, and Computing of Exploratory Data Analysis* (*ABC of EDA*), Duxbury. The Internet-First University Press 2004'te bu kitabın tıpkıbasımını yayımlamıştır. Kitap, Cornell Üniversitesi'nin şu adresinden ücretsiz indirilebilmektedir: http://dspace.library.cornell.edu/handle/1813/62.