DÖNÜŞÜMLER (TRANSFORMATIONS)

İstatistiksel yöntemlerin bir çoğu normallik varsayımına dayanır.

- Deney tasarımı
- Regresyon
- Zaman serileri

alanlarındaki yöntemler buna örnek olarak verilebilir. Bu nedenle analize başlamadan önce veri setinin dağılımının normal olup olmadığının test edilmesi gerekir. Eğer veri setinin dağılımı normal değilse veri seti aşağıda gösterilen alternatif yaklaşımlardan bir tanesi kullanılarak analiz edilmelidir.

Bu alternatif yaklaşımlar

- Parametrik Olmayan Teknikler
- Dayanıklı (Robust) Teknikler
- Dönüşümler

olarak gruplandırılabilir.

Bu bölümde dönüşümler konusu detaylı olarak incelenecektir. Veri setinin dağılımı normal olmadığında uygun dönüşüm yapılarak veri setinin dağılımı normal hale getirilir ve normal dağılım varsayımına dayanan testler kullanılarak analize devam edilir.

Not: Burada önemli olan hangi dönüşümün yapılacağına karar vermektir.

Üssel dönüşümler (Power transformations)

Üssel dönüşümler uygulamada yaygın olarak kullanılır. Sadece pozitif değerler alan gözlemler için kullanılabilir, ancak bu problem negatif değerler alabilen gözlemlere belli bir sabit ekleyerek söz konusu negatif değerleri pozitif hale getirerek ortadan kaldırılabilir.

x bir gözlem olmak üzere, üssel dönüşüm λ parametresi ile ilişkilendirilir. λ değeri belirli bir dönüşümü temsil eder.

Örnek: x^{λ} dönüşümü kullanıldığında

$\lambda = -1$ ise $x^{-1} = \frac{1}{x}$ olduğundan tüm gözlemlerin tersi alınarak dönüşüm yapılır.

$\lambda = 0$ ise $x^{\lambda} = lnx$ olarak tanımlanmaktadır. Bu durumda tüm gözlemlerin doğal logaritması (ln) alınarak dönüşüm yapılır.

Mümkün dönüşümler

$$x^{-1} = \frac{1}{x}\,, x^0 = lnx\,, x^{1/4} = \sqrt[4]{x}\,, x^{1/2} = \sqrt{x}\,, x^2, x^3$$

şeklinde gösterilebilir.

Bu dönüşümlerden,

$x^{-1}, x^0, x^{1/4}, x^{1/2}$ büyük değerli gözlemleri aşağıya çekerken

x^2, x^3 gibi dönüşümler büyük değerli gözlemleri daha da arttırır.

Box-Cox Dönüşümü

$x > 0$ için

$$x^{(\lambda)} = \begin{cases} \dfrac{x^{\lambda} - 1}{\lambda} & \lambda \neq 0 \\ lnx & \lambda = 0 \end{cases}$$

şeklinde tanımlanır.

Uygun λ değeri, aşağıdaki $l(\lambda)$ fonksiyonunu λ'ya göre maksimize eden değerin bulunmasıyla belirlenir. Burada,

$$l(\lambda) = -\frac{n}{2} ln\left[\frac{1}{n}\sum_{j=1}^{n}\left(x_j^{(\lambda)} - \bar{x}^{(\lambda)}\right)^2\right] + (\lambda - 1)\sum_{j=1}^{n} lnx_j$$

dır ve

$$\bar{x}^{(\lambda)} = \frac{1}{n}\sum_{j=1}^{n} x_j^{(\lambda)} = \frac{1}{n}\sum_{j=1}^{n}\left(\frac{x_j^{(\lambda)} - 1}{\lambda}\right)$$

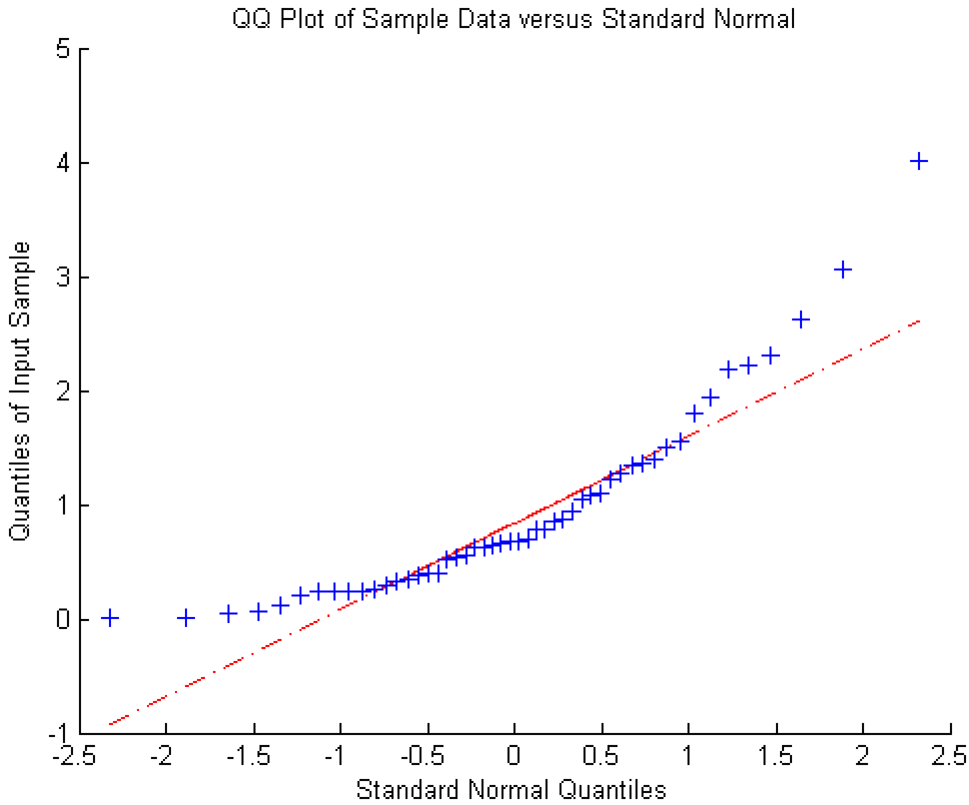
şeklinde tanımlanır.

Verileri dönüştürmek için kullanılacak uygun λ değeri belirlendikten sonra, dönüştürülen verilerin Q-Q grafiği çizilerek dağılımın normale uyup uymadığı kontrol edilir.

Örnek: Aşağıdaki veri setine Box-Cox dönüşümünü uygulayınız.

i	x	i	x
1	0.0001	**26**	0.6824
2	0.0012	**27**	0.6960
3	0.0439	**28**	0.7735
4	0.0612	**29**	0.7846
5	0.1104	**30**	0.8454
6	0.2040	**31**	0.8625
7	0.2298	**32**	0.9307
8	0.2315	**33**	1.0497
9	0.2420	**34**	1.0865
10	0.2438	**35**	1.1022
11	0.2534	**36**	1.2232
12	0.2959	**37**	1.2679
13	0.3286	**38**	1.3481
14	0.3357	**39**	1.3637
15	0.3692	**40**	1.3895
16	0.3885	**41**	1.4969
17	0.3935	**42**	1.5587
18	0.5170	**43**	1.7953
19	0.5356	**44**	1.9355
20	0.5560	**45**	2.1845
21	0.6161	**46**	2.2269
22	0.6167	**47**	2.3061
23	0.6452	**48**	2.6206
24	0.6543	**49**	3.0595
25	0.6789	**50**	4.0151

Öncelikle verilerin Q-Q grafiği aşağıdaki şekilde gösterildiği gibi çizilmiş ve gözlemlerin düz bir doğru etrafında yayılım göstermediği görülmüştür. Sonuç olarak orijinal verinin dağılımının normal olmadığı gösterilmiştir.



Not: Q-Q grafiği verilerin normal dağılıma sahip olması durumunda düz bir doğru şeklinde olur. Eğer normal dağılımdan sapmalar varsa Q-Q grafiğindeki bazı gözlemler doğrudan sapmalar gösterir.

Şimdi uygun λ değerini belirlemek için rastgele olarak seçilmiş bir aralıkta değişen λ değerlerine karşılık gelen $l(\lambda)$ değerlerini elde edelim. Bu örnekte $\lambda \in [-1.0, 1.5]$ olarak alınmıştır. Bu aralığın $l(\lambda)$ fonksiyonunun maksimum olduğu noktayı kapsaması gerekir. Sonuçlar aşağıdaki tabloda gösterildiği gibi elde edilmiştir.

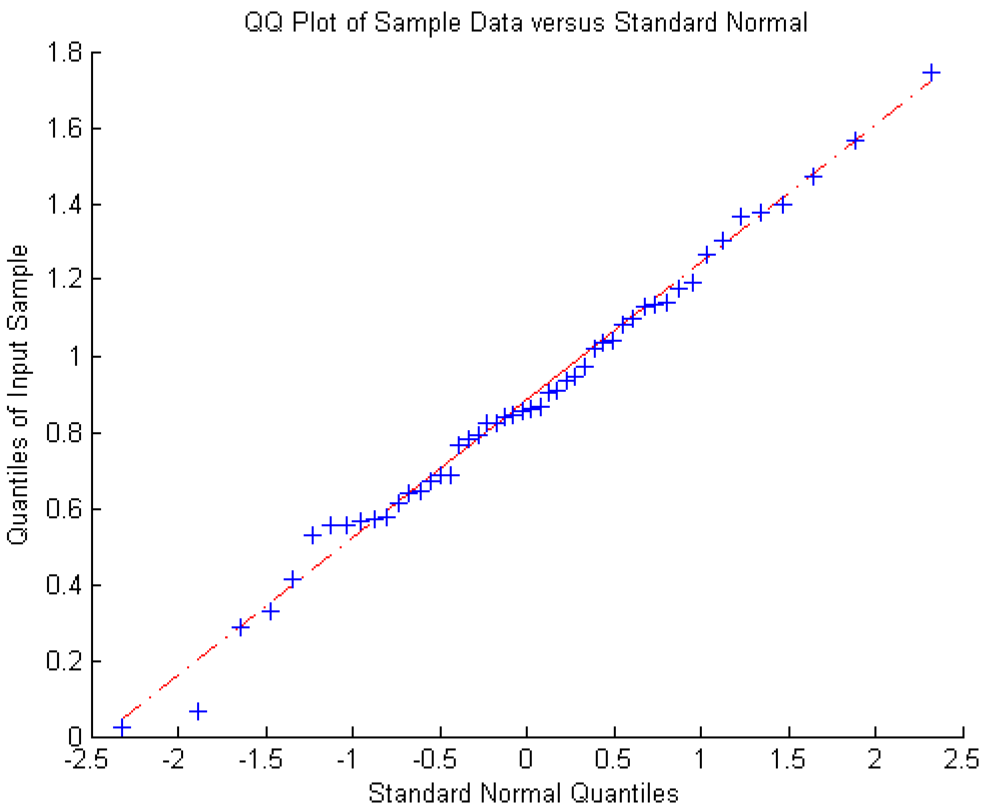
λ	$l(\lambda)$	λ	$l(\lambda)$
-1.0	-291.4442	0.3	26.0313
-0.9	-254.2973	**0.4**	**26.6314**
-0.8	-217.8169	0.5	25.6828
-0.7	-182.1864	0.6	23.6352
-0.6	-147.6599	0.7	20.7669
-0.5	-114.5884	0.8	17.2532
-0.4	-83.4544	0.9	13.2091
-0.3	-54.9101	1.0	8.7139
-0.2	-29.7975	1.1	3.8252
-0.1	-9.0628	1.2	-1.4131
0.0	6.5692	1.3	-6.9660
0.1	17.0618	1.4	-12.8048
0.2	23.1754	1.5	-18.9050

Yukarıdaki tablodan $l(\lambda)$ fonksiyonunu maksimum yapan λ değerinin 0.4 olduğu görülmektedir. Orijinal verilere (x_j), $\lambda = 0.4$ alınarak Box-Cox dönüşümü uygulanırsa

$$x_j^{(0.4)} = \frac{x_j^{0.4} - 1}{0.4}, \quad j = 1,2,\dots,50$$

olarak ifade edilen dönüştürülmüş veriler elde edilir.

Dönüştürülmüş verilerin Q-Q grafiği aşağıdaki gibidir.



Dönüştürülmüş verilerin normal dağılıma sahip olduğu görülür. Çünkü Q-Q grafiğinden de görüleceği gibi veriler düzgün bir doğru etrafında yayılım göstermektedir.

Not: λ için aralık belirlenirken $l(\lambda)$ fonksiyonunun bu aralıkta önce artan sonra azalan olması gerekir. Aksi takdirde kullanılan aralık maksimum noktayı kapsamaz.