Regresyon Analizi Ödevi

Veri : Steel Industry Data

**Tanla Çakır:** 20191101026

**Ece Simay Timoçin:** 20191101037

**Muhammet Sina Kiriş:** 20191101011

**Beyza Memiş:** 20191101020

**Alper Ertaş:** 20191101041

Bağımlı değişken : Sanayi enerji tüketimi(kWh) (y)

Bağımsız değişkenler :Gecikmeli akım reaktif güç(kVarh) (x1)

Lider akım reaktif güç(kVarh) (x2)

Gecikmeli akım güç faktörü(%) (x3)

Lider akım güç faktörü(%) (x4)

Yük türü (x5)

n = 25

1. **NORMALLİK TESTİ**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Tests of Normality** | | | | | | |
|  | Kolmogorov-Smirnova | | | Shapiro-Wilk | | |
| Statistic | df | Sig. | Statistic | df | Sig. |
| Enerji\_kWh | ,322 | 25 | ,000 | ,613 | 25 | ,000 |
| a. Lilliefors Significance Correction | | | | | | |

Hipotez:

H0 : Regresyon modeli normal dağılıma uymaktadır.

H1: Regresyon modeli normal dağılıma uymamaktadır.

N < 50 olduğu için Shapiro-wilk kontrol edilecek.

Sig < 0.05 , H0 reddedilir. Regresyon modelinin normal dağılıma uyduğu söylenemez.

1. **ARTIK ANALİZİ**

Önce modelimizin özet istatistiklerini gözlemleyelim:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model Summaryb** | | | | |
| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate |
| 1 | ,942a | ,888 | ,865 | 8,78871 |
| a. Predictors: (Constant), Lider\_akım\_guc\_faktoru\_(%), Gecikmeli\_akım\_reaktif\_guc\_kVarh, Gecikmeli\_akım\_guc\_faktoru\_(%), Lider\_akım\_reaktif\_guc\_kVarh | | | | |
| b. Dependent Variable: Enerji\_kWh | | | | |

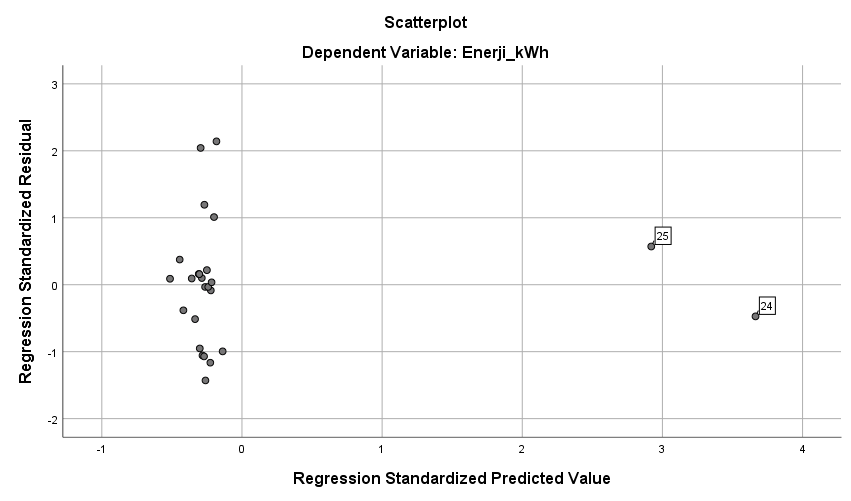
**R^2 = 0.888**

Bağımlı değişken modeldeki bağımsız değişkenler tarafından %89 oranında açıklanıyor.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **ANOVAa** | | | | | | |
| Model | | Sum of Squares | df | Mean Square | F | Sig. |
| 1 | Regression | 12190,438 | 4 | 3047,609 | 39,456 | ,000b |
| Residual | 1544,830 | 20 | 77,241 |  |  |
| Total | 13735,268 | 24 |  |  |  |
| a. Dependent Variable: Enerji\_kWh | | | | | | |
| b. Predictors: (Constant), Lider\_akım\_guc\_faktoru\_(%), Gecikmeli\_akım\_reaktif\_guc\_kVarh, Gecikmeli\_akım\_guc\_faktoru\_(%), Lider\_akım\_reaktif\_guc\_kVarh | | | | | | |

**Regression = 12190,438**

**Residual = 1544,830**



Yukarıdaki grafik yi^ değerlerinin standartlaştırılmış artıklara olan saçılım grafiğidir. Grafikte gözümüze 2 uzak değer çarpıyor. Değerlerin hangileri olduğunu tespit ediyoruz. Daha sonra tek tek veriden çıkararak aykırı değer olup olmadıklarını kontrol ediyoruz.

Önce 24 numaralı veriyi çıkararak sonuçlara bakıyoruz.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model Summaryb** | | | | |
| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate |
| 1 | ,892a | ,795 | ,752 | 8,88324 |
| a. Predictors: (Constant), Lider\_akım\_guc\_faktoru\_(%), Gecikmeli\_akım\_reaktif\_guc\_kVarh, Gecikmeli\_akım\_guc\_faktoru\_(%), Lider\_akım\_reaktif\_guc\_kVarh | | | | |
| b. Dependent Variable: Enerji\_kWh | | | | |

**R^2 = 0,795**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **ANOVAa** | | | | | | |
| Model | | Sum of Squares | df | Mean Square | F | Sig. |
| 1 | Regression | 5829,485 | 4 | 1457,371 | 18,468 | ,000b |
| Residual | 1499,326 | 19 | 78,912 |  |  |
| Total | 7328,811 | 23 |  |  |  |
| a. Dependent Variable: Enerji\_kWh | | | | | | |
| b. Predictors: (Constant), Lider\_akım\_guc\_faktoru\_(%), Gecikmeli\_akım\_reaktif\_guc\_kVarh, Gecikmeli\_akım\_guc\_faktoru\_(%), Lider\_akım\_reaktif\_guc\_kVarh | | | | | | |

**Regression = 5829,485**

**Residual = 1499,326**

SSreg ve SSres değerleri azaldı, 24.verinin artık değer olduğunu söyleyebiliriz.

Şimdi 25.veriyi çıkardıktan sonraki değerleri inceleyelim:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model Summaryb** | | | | |
| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate |
| 1 | ,908a | ,824 | ,787 | 8,88209 |

|  |
| --- |
| a. Predictors: (Constant), Lider\_akım\_guc\_faktoru\_(%), Gecikmeli\_akım\_reaktif\_guc\_kVarh, Gecikmeli\_akım\_guc\_faktoru\_(%), Lider\_akım\_reaktif\_guc\_kVarh |
| b. Dependent Variable: Enerji\_kWh |

**R^2 =0,824**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **ANOVAa** | | | | | | |
| Model | | Sum of Squares | df | Mean Square | F | Sig. |
| 1 | Regression | 7009,931 | 4 | 1752,483 | 22,214 | ,000b |
| Residual | 1498,939 | 19 | 78,892 |  |  |
| Total | 8508,870 | 23 |  |  |  |
| a. Dependent Variable: Enerji\_kWh | | | | | | |
| b. Predictors: (Constant), Lider\_akım\_guc\_faktoru\_(%), Gecikmeli\_akım\_reaktif\_guc\_kVarh, Gecikmeli\_akım\_guc\_faktoru\_(%), Lider\_akım\_reaktif\_guc\_kVarh | | | | | | |

**Regression = 7009,931**

**Residual = 1498,939**

Regression ve Residual değerleri azaldı , 25.verinin artık değer olduğunu söyleyebiliriz.

**Sonuç =**

24.değer atıldığında R^2’de daha büyük düşüş yaşanmıştır. Modelin anlamlılığı azalmıştır. Aynı zamanda 24. değer regresyon eğrisine diğer artıklara göre daha uzaktadır. Bu nedenle 24. değere aykırı değer diyebiliriz.

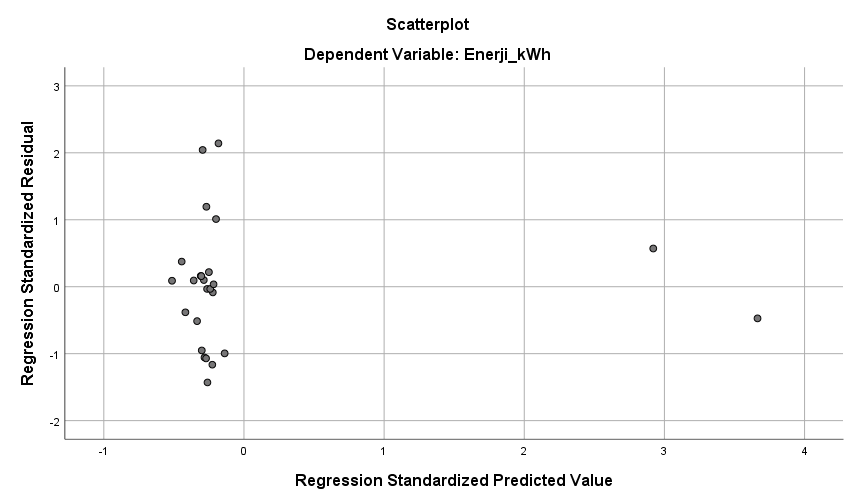
25.değer atıldığında R^2’de çok büyük bir düşüş yaşanmamıştır. Modelin anlamlılığı bozulmamıştır. Uç değer diyebiliriz.

1. **DÖNÜŞÜMLER**

Birinci sorudaki hipotezlerimizin sonucunda veri setimizin normal dağılmadığını görmüştük. Bu sorunu gidermek ve normal dağılıma sahip bir veri seti oluşturmak için çeşitli dönüşümler yapmalıyız.

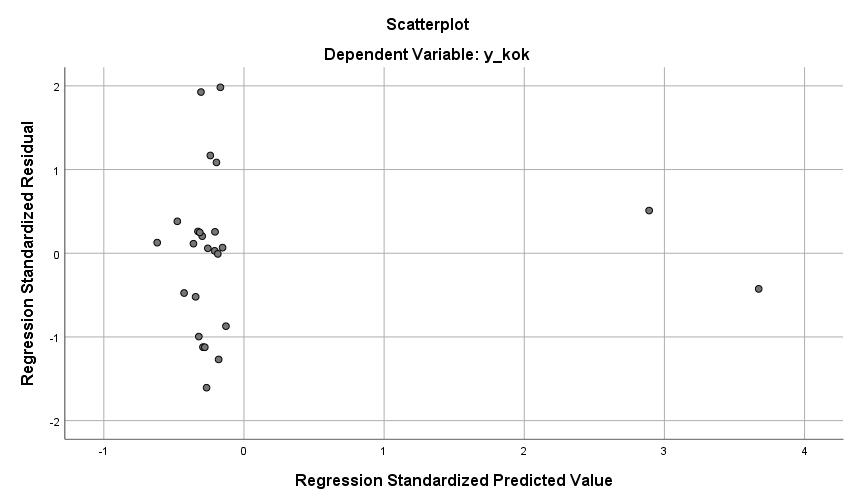
Ln, Kök ve Ters dönüşümleri tek tek yaptıktan; normallik testlerine, histogramlara ve Skewness-Kurtosis değerlerine baktıktan sonra şu yorumlar yapılabilir:

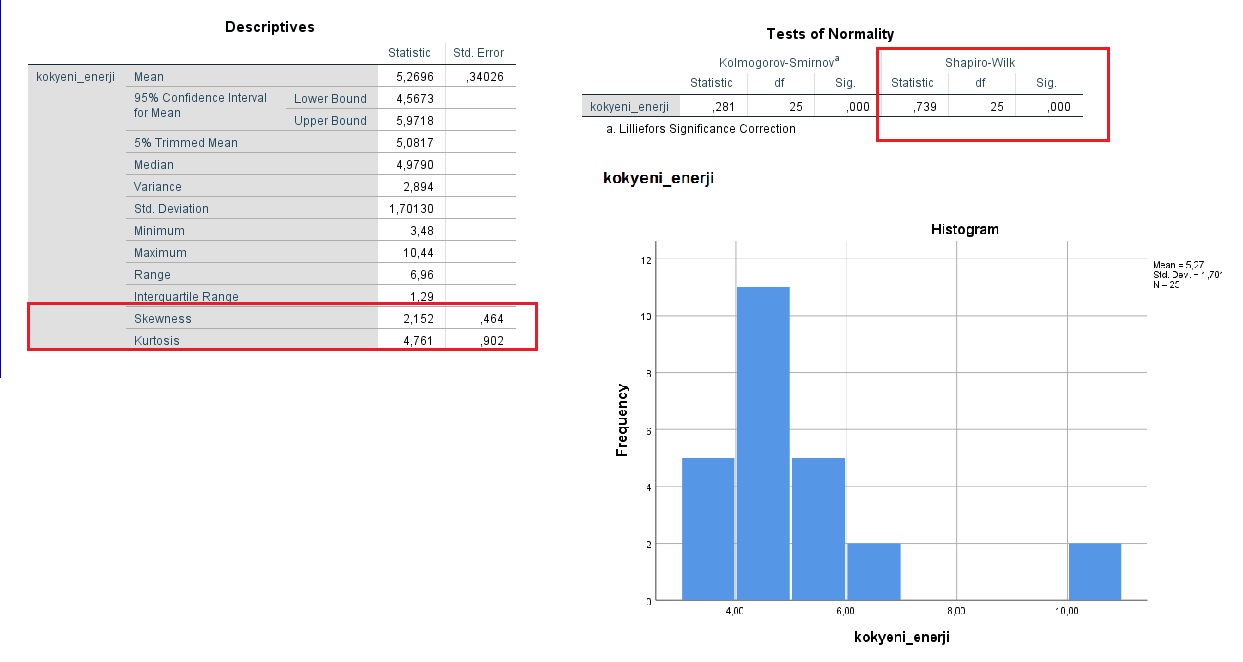
Modelin dönüşüm uygulanmamış artık grafiğini görelim.

****

**Kök Dönüşümü**

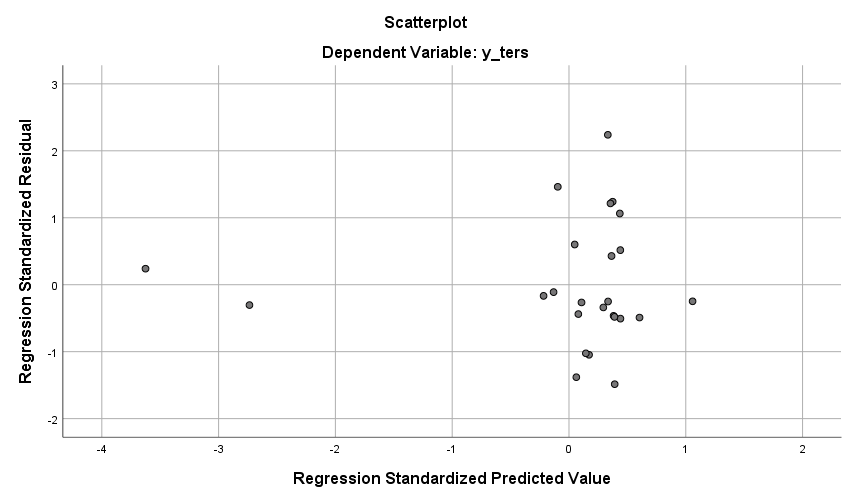
Kök dönüşümü için: Sig değeri 0,000 olduğu, histogramda sağa çarpıklık gözlendiği ve Skewness-Kurtosis değerleri Tabachnick'e göre (-1.5, 1.5) aralığında olmadığı için normal dağılımdan söz edemeyiz.

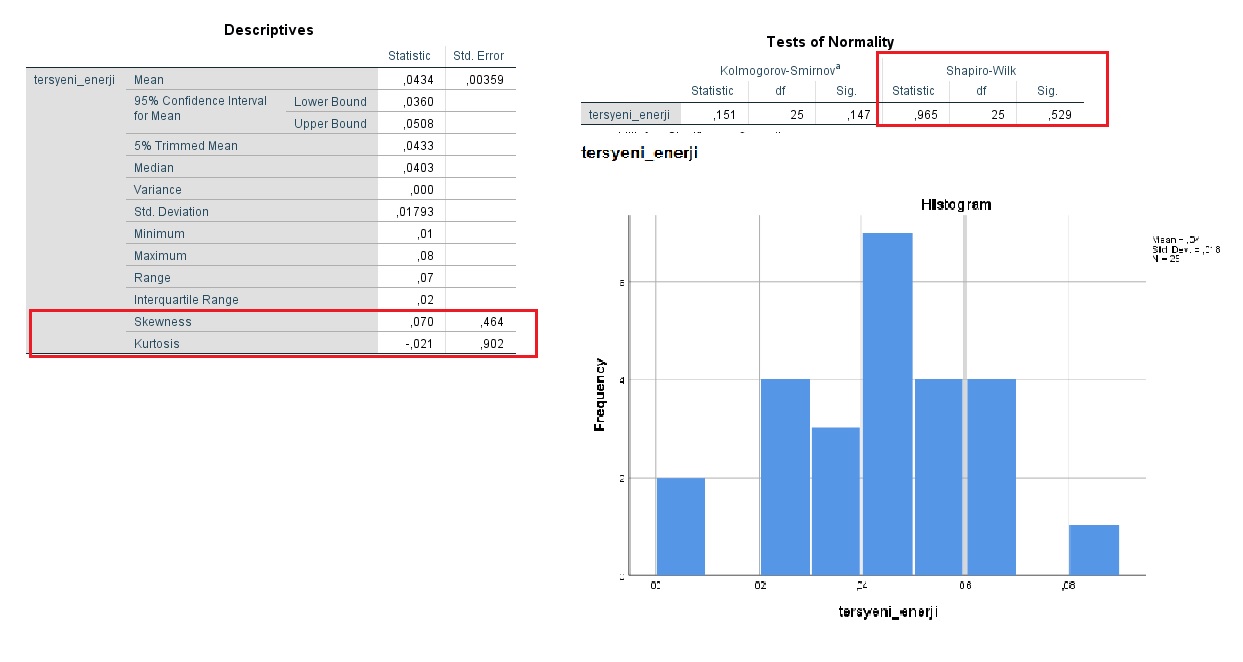
****



**Ters Dönüşüm**

Ters dönüşüm için: Shapiro-Wilk testinde sig=0,529 olduğu, histogramda normal dağılım gözlendiği ve Skewness-Kurtosis değerleri Tabachnick'e göre (-1.5, 1.5) aralığında olduğu için normal dağılım sağlanmıştır.



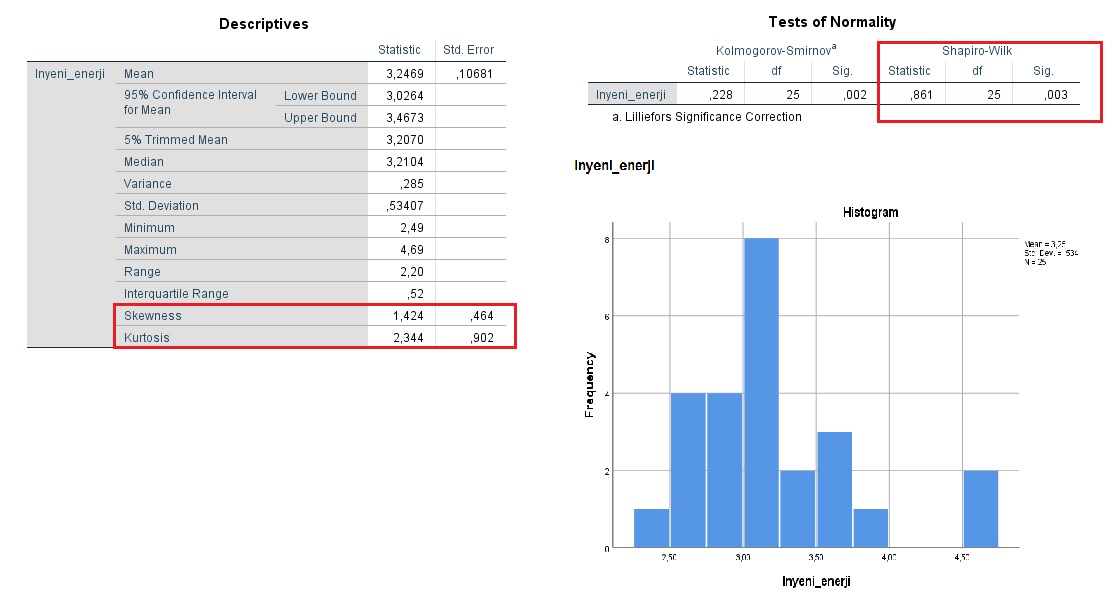


**Ln Dönüşümü**

Ln dönüşümü için: Histograma baktığımızda normal dağıldığını düşünebiliriz fakat sig değeri 0,003 olduğu ve Skewness-Kurtosis değerleri Tabachnick'e göre (-1.5, 1.5) aralığında olmadığı için normal dağılımdan söz edemeyiz.

tablo içeren bir resim

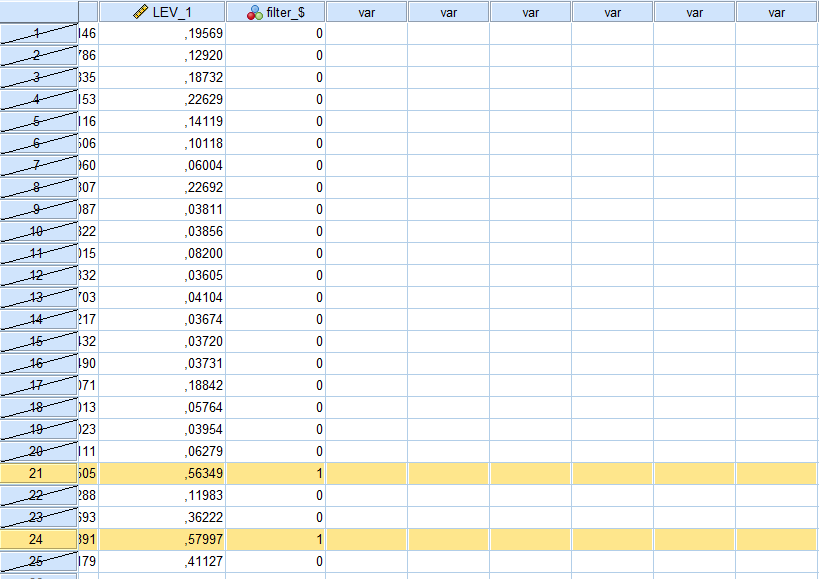
Açıklama otomatik olarak oluşturuldu



**d) ETKİNLİK ANALİZİ**

Öncelikle bağımlı değişken üzerinde değişim yapılmamış haliyle etkinlik analizi yapacağız.

**Kaldıraç (Centered Leverage Value)**

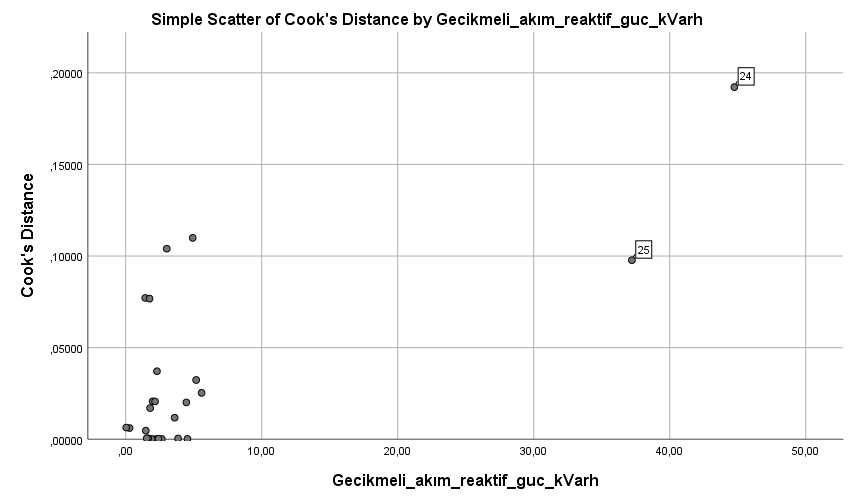
****

Teorik olarak, 2(p/n)’den büyük olan H şapka matrisinin köşegen elemanları hii’ler, bize potansiyel aykırı değerleri verirler. Çünkü bu elemanlar verilerin geri kalanından yeterince uzaktır. Yukarıdaki analizde görüldüğü gibi 2(p/n)’den büyük olan gözlemler 21 ve 24’tür. Bu gözlemlerin bir uç değer ya da etkili gözlem olduğunu henüz göremeyiz ama bir kaldıraç noktası olduğunda hemfikir olabiliriz.

**Covratio**

Etkinlik analizi yaptığımızda, “COVRATIO” için verilere COV>1+((3p)/n) formülünü uyguladığımızda 4.(1,73316), 8.(1,76225), 21.(3,24301), 23.(2,16072),24.(2,92847),25.(2,02553) gözlemlerin 1+((3p)/n)’den büyük olduğu gözlemlenmiştir. Bu sebeple, bu gözlemler etkin değer olarak kabul edilebilir. Bu gözlemler kestirimin keskinliğini arttırma eğilimindedir.

**Cook’s Distance**



Cook’un D uzaklığı grafiğini çıkarttığımızda 24 ve 25. değerler gözümüze çarpıyor.

24.gözlemin cook değerine baktığımızda : ,19221 olarak gözlemliyoruz. Bu gözlemin etkili olduğunu söyleyemeyiz.

25.gözlemin cook değerine baktığımızda : ,09772 olarak gözlemliyoruz. Bu gözlemin etkili olduğunu söyleyemeyiz.

**DFFITS**

Analiz yapıldığında |DFFITS| > 2((p/n)\*\*1/2) olan;

yani |DFFITS| > 0,98 olan gözlemler aşağıdaki gibidir:

1.Gözlem = -1,03462

3.Gözlem = -1,32822

5.Gözlem= 2,32276

9.Gözlem = 1,52163

11.Gözlem= 1,23411

16.Gözlem= -1,05277

17.Gözlem= -3,02901

18.Gözlem= 2,03580

20.Gözlem= -1,00239

21.Gözlem= 1,19259

24.Gözlem= -6,78397

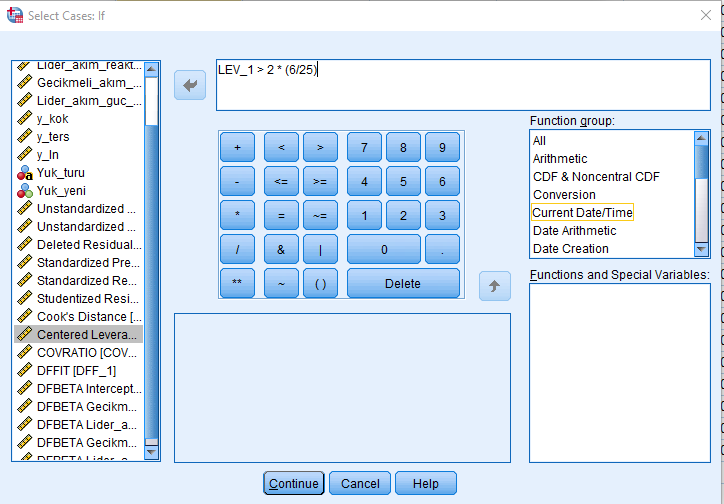
25.Gözlem= 4,12689

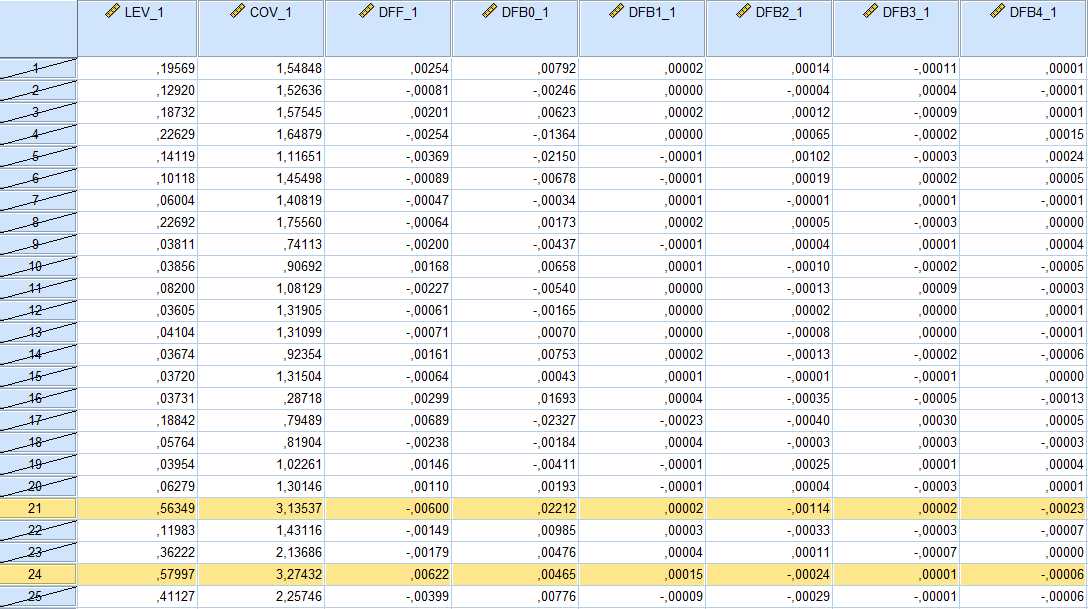
Bu gözlemler etkililik bakımından incelenmelidir.

Şimdi de bağımlı değişken üzerinde ters dönüşüm yapılmış haliyle etkinlik analizi yapacağız.

**Kaldıraç (Centered Leverage Value)**

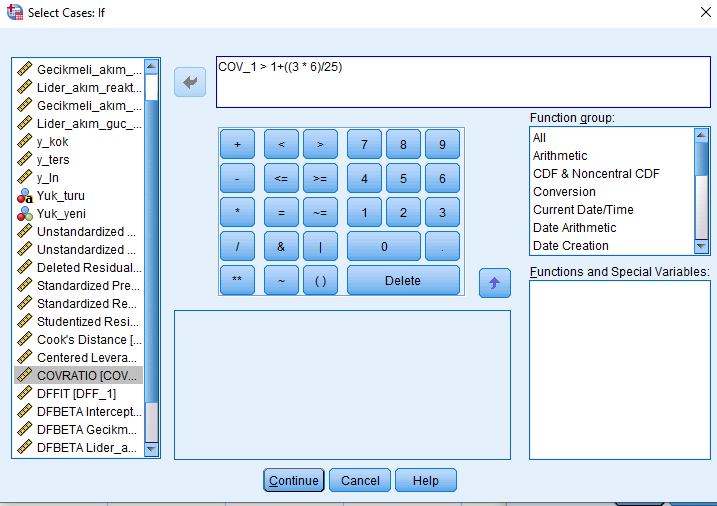
Yukarıda yaptığımız gibi kaldıraç noktalarını belirlemek için 2(p/n)’den büyük olan hii’leri tespit edeceğiz. Analizde görüldüğü gibi 2(p/n)’den büyük olan gözlemler yine 21 ve 24’tür. Bu gözlemlerin bir uç değer ya da etkili gözlem olduğunu henüz göremeyiz ama bir kaldıraç noktası olduğunda hemfikir olabiliriz.

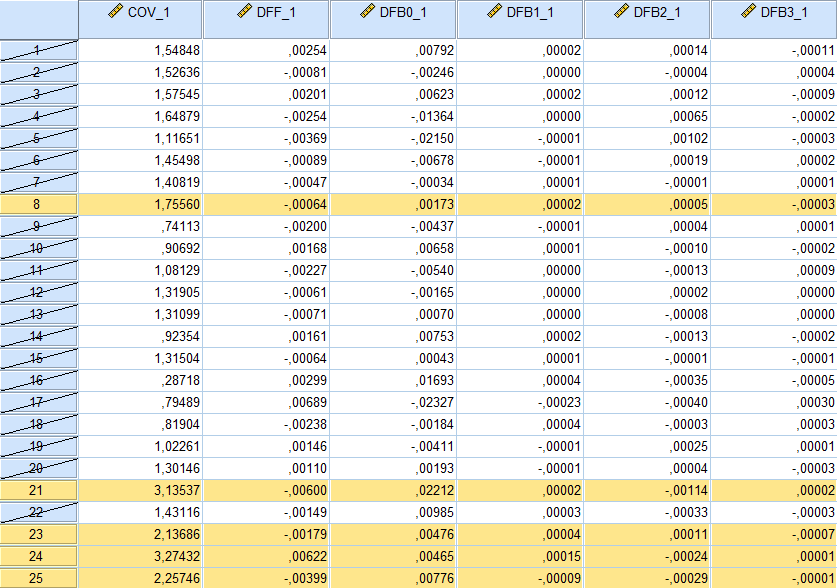
****



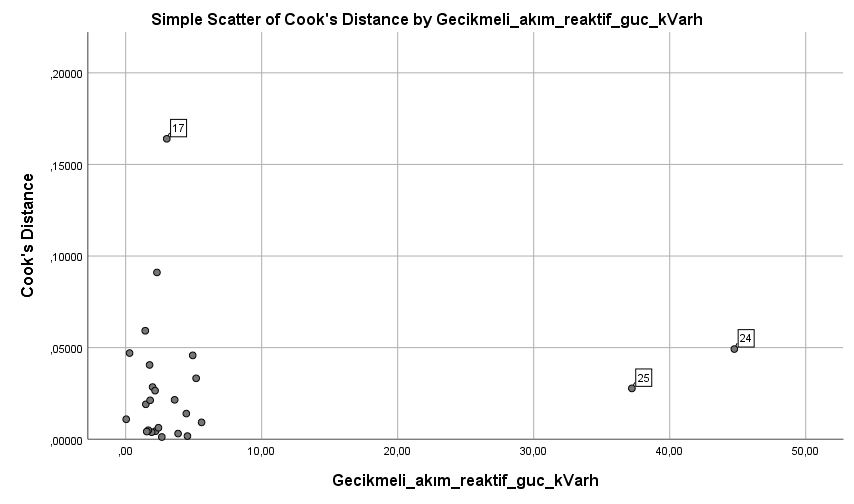
**Covratio**

Etkinlik analizi yaptığımızda, “COVRATIO” için verilere COV>1+((3p)/n) formülünü uyguladığımızda 8.( 1,75560), 21.(3,13537), 23.(2,13686),24.(3,27432),25.(2,25746) gözlemlerin 1+((3p)/n)’den büyük olduğu gözlemlenmiştir. Bu sebeple, bu gözlemler etkin değer olarak kabul edilebilir. Gözlemler kestirimin keskinliğini arttırma eğilimindedir.



****

**Cook’s Distance**



Cook’un D uzaklığı grafiğini çıkarttığımızda 24 , 25 ve 17. Değerler gözümüze çarpıyor.

24.gözlemin cook değerine baktığımızda : ,04923 olarak gözlemliyoruz. Bu gözlemin etkili olduğunu söyleyemeyiz.

25.gözlemin cook değerine baktığımızda : ,02776 olarak gözlemliyoruz. Bu gözlemin etkili olduğunu söyleyemeyiz.

17. gözlemin cook değerine baktığımızda : ,16399 olarak gözlemliyoruz. Bu gözlemin etkili olduğunu söyleyemeyiz.

**DFFITS**

Ters dönüşümde analiz yapıldığında |DFFITS| > 2((p/n)^1/2) olan;

yani |DFFITS| > 0,98 olan hiçbir gözlem saptanamamıştır.

**e) DUMMIES**

Hafif yük türünü referans alarak yük türünü dahil ettiğimizde model özeti aşağıdaki gibi gözükmektedir :tablo içeren bir resim

Açıklama otomatik olarak oluşturuldu

tablo içeren bir resim

Açıklama otomatik olarak oluşturuldu

Katsayılarda sig. değerleri < 0.05 olan değişkenlere bakıyoruz, anlamlılıklarını test ediyoruz.

H0: Regresyon modeli anlamsızdır.

H1: Regresyon modeli anlamlıdır.

Gecikmeli akım reaktif güç dışında diğer değişkenlerin anlamsız olduğunu tespit ediyoruz. Analizimize anlamsız olan değerleri çıkararak devam ediyoruz.

tablo içeren bir resim

Açıklama otomatik olarak oluşturuldu

Anlamsız değerleri çıkardıktan sonra R^2de çok bir değişim olmadığını gözlemliyoruz.

Eski R^2 değeri = 0,892

Yeni R^2 değeri = 0,877

tablo içeren bir resim

Açıklama otomatik olarak oluşturuldu

Anlamsız değişkenler çıktıktan sonra katsayılar tablosu yukarıda verildiği gibi olmaktadır. Bunun da yorumunu :

Kestirim Modeli : yi^ = 18,667 + 2,080xi

(1,950) (0,163) (hatalar)

Gecikmeli akım reaktif gücündeki bir birimlik artış , gözlenen enerji gücünde ortalama 2,080 birimlik değişime neden olur.

**Y ters dönüşümlü yaptığımız zaman:**

tablo içeren bir resim

Açıklama otomatik olarak oluşturuldu

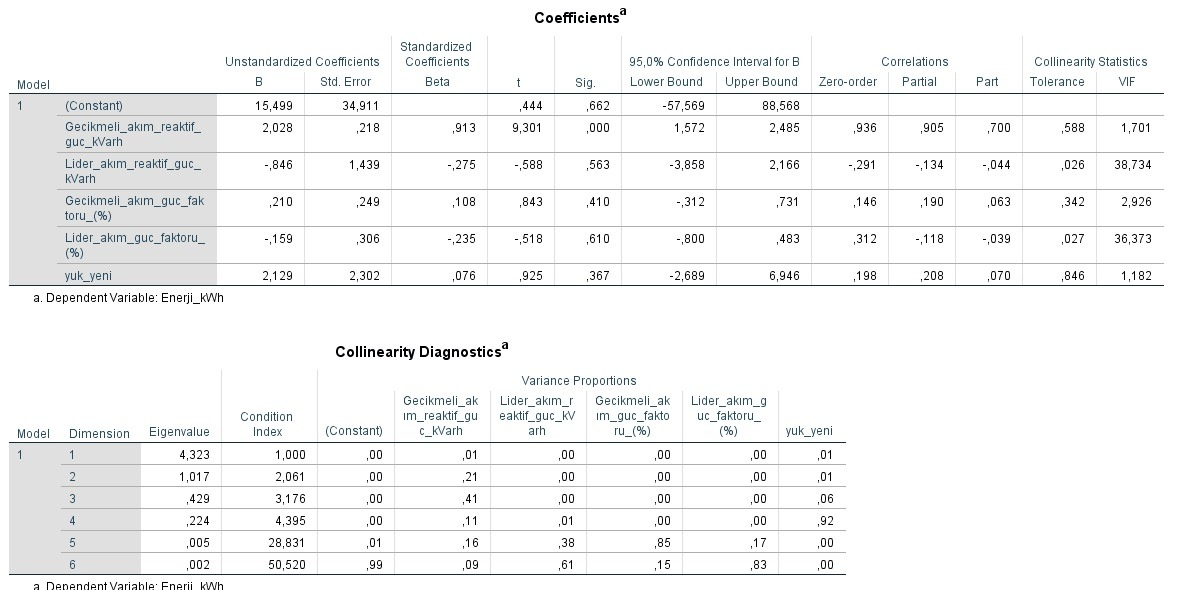
tablo içeren bir resim

Açıklama otomatik olarak oluşturuldu

Değişkenlerin sig. değerine baktığımız zaman hepsinin 0,05’ten büyük olduğunu yani H0 hipotezinin hepsi için kabul edildiğini görürüz. Anlamsız değişkenler için bir analiz yapamayız.

**f) ÇOKLU-BAĞLANTI**

Dönüşüm yapılmadığında;



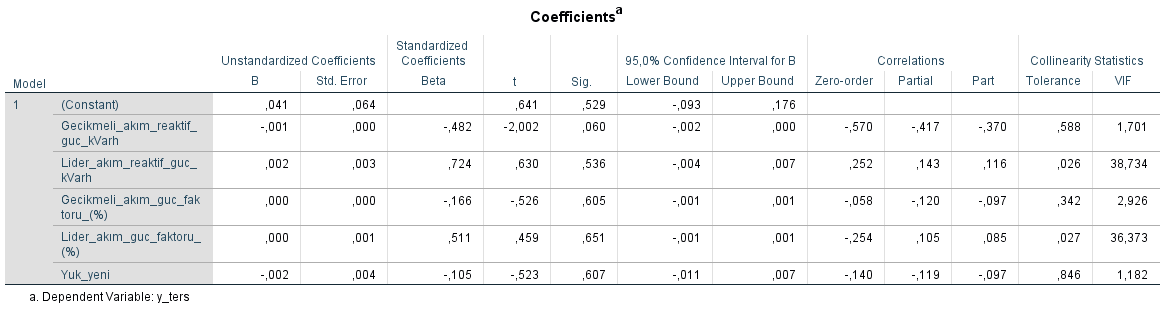
Dönüşüm yapılmamış y değerleriyle:

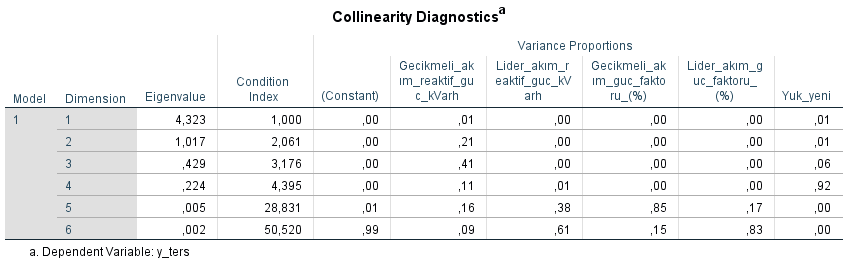
VIF değerleri > 10 ise çoklu bağlantı problemi vardır diyebiliriz. Buradan yola çıkarak;

“Lider akım reaktif güç” ve “Lider akım güç faktörü” bağımsız değişkenlerinde çoklu bağlantı problemi vardır.

Aynı kararı “Koşul indisleri > 30” olma durumuna göre de yorumlayabiliriz.

Dönüşüm yapıldığında;



****

Dönüşüm bağımlı değişken üzerinde yapıldığı için bağımsız değişkenler üzerinde herhangi bir değişim olmamıştır. Bu sebeple çoklu bağlantı değerleri aynı kalmıştır.