|  |
| --- |
| Final Ödevi |
| Bilecik Şeyh Edebali Üniversitesi |
| R ve İstatistiksel Programlama Dili |

|  |
| --- |
| MERT YİĞİT Çelebi 26825091238  6.01.2023 |

**Regresyon Analizi**

Regresyon analizi, iki ya da daha çok nicel değişken arasındaki ilişkiyi ölçmek için kullanılan analiz metodudur. Eğer tek bir değişken kullanılarak analiz yapılıyorsa buna tek değişkenli regresyon, birden çok değişken kullanılıyorsa çok değişkenli regresyon analizi olarak isimlendirilir. Regresyon analizi ile değişkenler arasındaki ilişkinin varlığı, eğer ilişki var ise bunun gücü hakkında bilgi edinilebilir.

Örneğin, bir ziraat mühendisi buğday verimi ve gübre miktarı arasındaki ilişkiyi, bir mühendis, basınç ve sıcaklık bir ekonomist gelir düzeyi ve tüketim harcamaları, bir eğitimci öğrencilerin devamsızlık gösterdiği gün sayıları ve başarı dereceleri arasındaki ilişkiyi bilmek isteyebilir. Regresyon, iki (ya da daha çok) değişken arasındaki doğrusal ilişkinin fonksiyonel şeklini, biri bağımlı diğeri bağımsız değişken olarak bir doğru denklemi olarak göstermekle kalmaz, değişkenlerden birinin değeri bilindiğinde diğeri hakkında kestirim yapılmasını sağlar. Genellikle bu iki (veya çok) değişkenlerin hepsinin niceliksel ölçekli olması zorunluluğu vardır.

Regresyonda, değişkenlerden biri bağımlı diğerleri bağımsız değişken olmalıdır. Buradaki mantık eşitliğin solunda yer alan değişkenin sağında yer alan değişkenlerden etkilenmesidir. Sağda yer alan değişkenlerse diğer değişkenlerden etkilenmemektedir. Burada etkilenmemek matematiksel anlamda bu değişkenleri bir doğrusal denkleme koyduğumuzda etki yapması anlamındadır. Çoklu doğrusallık, ardışık bağımlılık sorunları kastedilmemektedir.

Doğru bir regresyon analizi yapabilmemiz için gerekenler önecelikle;

1. Bağımlı ve bağımsız değişkenlerimizi belirlememiz gerekir.
2. Gerekli Hipotezleri kurmamız gerekir.
3. Daha doğru bir sonuç elde etmemiz için değişkenlerimize normallik testi yapmamız gerekir.
4. Regresyon modelimizi oluşturup doğru bir şekilde yorumlamamız gerekir.

Buradaki bilgilerimizden yaralanarak R dilinde bir regresyon modeli yorumlayalım:

Öncelikle verimizi belirleyip veri setini R’ a tanımlamamız gerekir. Ben burada verimi UCİ Machine Learning Repository sitesindeki: <https://archive.ics.uci.edu/ml/datasets/Seoul+Bike+Sharing+Demand>

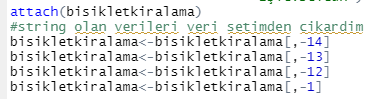
Bisiklet Kiralama verisini kullandım.



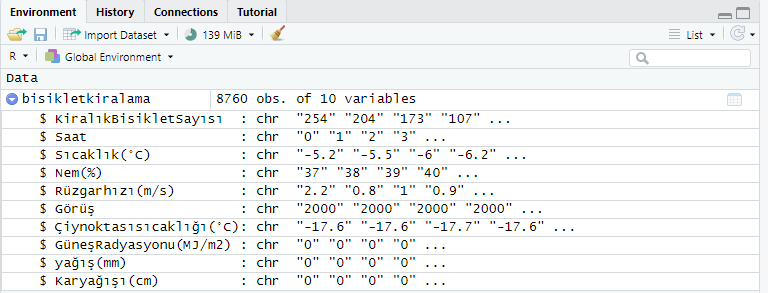
Kodunu kullanarak veri setimi R’a çağırdım. Bura da ilk karşılaştığım sorun verimin sütun isimlerinin içinde birinci satırın da olmasıydı bende bu sorunu;

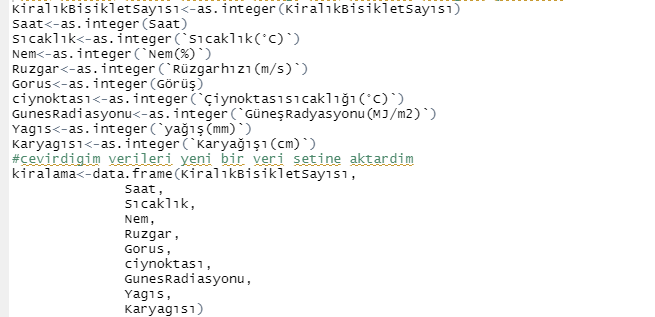


İlk satırı silip verimin sütün isimlerini değiştirdim. Hazır verimin sütün isimlerini değiştirirken Türkçe karakterler kullandım.



Gerekli işlemleri yapabilmek için nitel verilerden kurtuldum. Sonra veri setime baktığımda verilerimin hepsinin karakter olduğunu fark ettim. Veri setimde ki tüm veriler karakter olduğu için hiçbir işlem yapamadım.

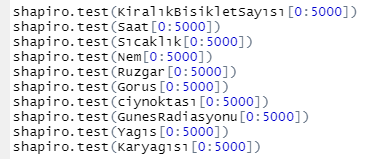


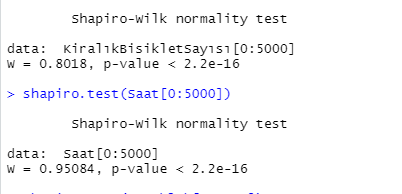


Daha düzgün bir veri seti oluşturmam gerekiyordu, ben de veri setimdeki tüm verileri stringe (sayısal) çevirip yeni bir veri setine aktardım. Artık regresyon analizi yapabileceğim bir veri setine sahibim. Doğru bir kanıya varabilmek için bağımlı ve bağımsız değişkenlerimizi belirleyelim, değişken türü ve ölçme düzeylerini yazalım. Bağımlı değişkenimiz bura da bisiklet kiralama sayısıdır. Saat, sıcaklık, nem, rüzgar hızı, görüş, çiy noktası, güneş radyasyonu, yağış , kar yağışı arttıkça ya da azaldıkça bağıllı değişkenimiz de değişim gözlenmektedir.

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | Değişken Türü | Ölçme Düzeyi |
| Bağımlı Değişken: | Bisiklet Kiralama Sayısı | nicel | Oranlama |
| Bağımsız Değişken: | Saat(sa) | nicel | Oranlama |
| Bağımsız Değişken: | Sıcaklık(\*C) | nicel | Eşit Aralık |
| Bağımsız Değişken: | Nem(%) | nicel | Oranlama |
| Bağımsız Değişken: | Rüzgar(m/s) | nicel | Oranlama |
| Bağımsız Değişken: | Görüş(10m) | nicel | Oranlama |
| Bağımsız Değişken: | Çiy noktası(\*C) | nicel | Eşit Aralık |
| Bağımsız Değişken: | Güneş Radyasyonu(Mj/m2) | nicel | Oranlama |
| Bağımsız Değişken: | Yağış(mm) | nicel | Oranlama |
| Bağımsız Değişken: | Karyağışı(cm) | nicel | Oranlama |

Değişkenlerimizi tanımladıktan sonra verimize normallik testi uygulayalım. Normal dağılıma uyumlu bir veri ise verimden daha doğru sonuçlar elde edebilirim. İlk olarak Shapiro-Wilk testiyapalım. R’da Shapiro-Wilk testi 3 ila 5000 değer veri üzerinden çalıştığı için ilk 5000 veriyi aldım.



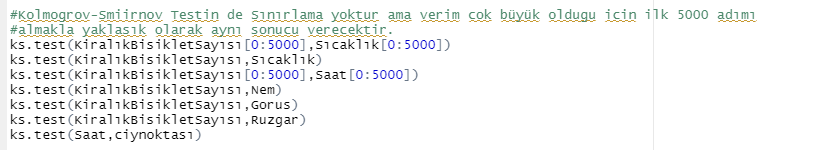


Sonuçlarını elde ettim. Burada diğer bağımsız değişkenlerde yaklaşık olarak aynı değerleri verdiği için çıktılarını almadım. Şimdide hipotezleri kurup 0.05 anlamlık düzeyinde yorumlayalım;

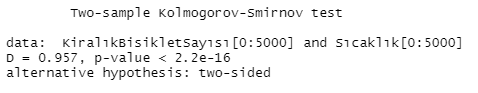
𝐻0: Normal Dağılıma Uygundur. a=0.05

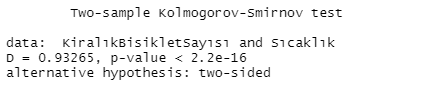
𝐻1: Normal Dağılıma Uygun değildir.

Burada tüm verilerim için p<a olduğundan tüm veriler için H0 Reddedilir. Shapiro-Wilk testine göre %95 güven aralığında verilerimiz normal dağılıma uyumlu değildir.



Kolmogrov-Smirnov testine göre normallik testi yaptığımızda:





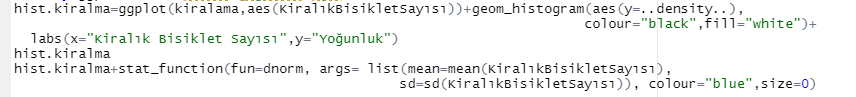
Sonuçlarını elde ettim. Burada elde edeceğim sonuçlar Shapiro-Wilk testi ile aynı kanıya vardım. Verilerimin hepsi için p<a olduğundan Kolmogrov-Smirnov testine göre normal dağılıma uyumlu değildir.

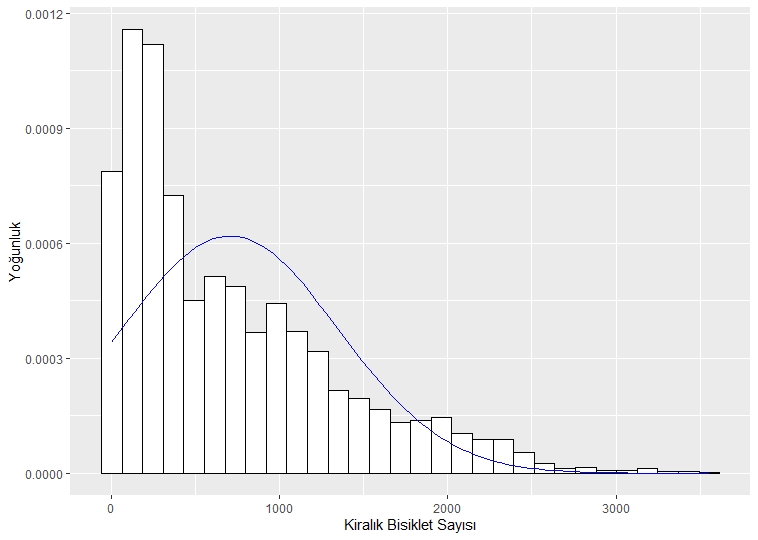
NOT: Büyük verilerde çalışır iken verilerimiz de normalden az bir sapma gözlemleniyor ise verimiz, Shapiro-Wilk ve Kolmogrov-Smirnov testlerine göre ugun çıkmasa bile, verimizin kesin bir şekilde normal dağılıma uygun olmadığını söyleyemeyiz. Dolayısıyla büyük örneklemlerde bu testlere göre normallik varsayımı yapamayız. Böyle durumlarda verinin histogram grafiğine, Q-Q plot ve çarpıklık-basıklık değerlerini incelememiz gerekir.

Q-Q plot ve çarpıklık-basıklık değerlerini incelememiz için R’a ggplot2 paketini yüklememiz ve çağırmamız gerekir.



Şimdi de çarpıklık grafiğini oluşturup yorumlayalım:

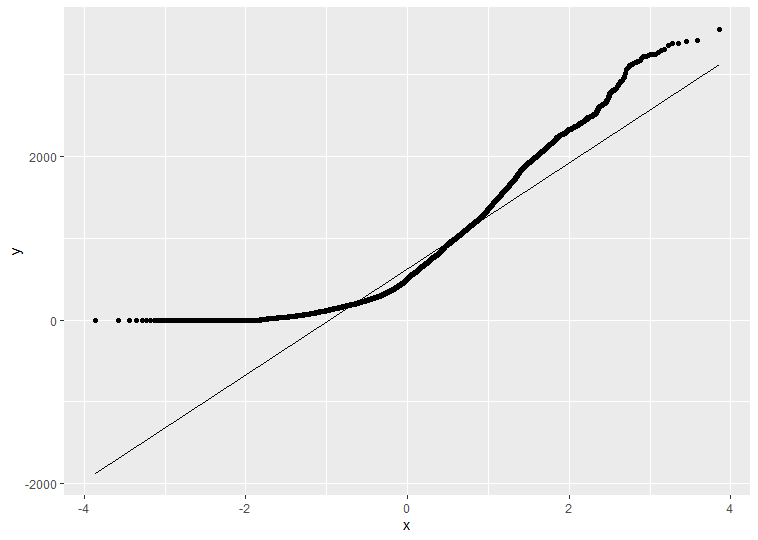




Buradan verimizin Sağa Çarpık (Sola Yığılımlı) normal dağılımlı bir grafik oluşturduğunu söyleyebiliriz. Verimiz için Mod < Medyan < Aritmetik Ortalamadır. Verimiz çok büyük olduğu için buradan vardığımız kanı yüzde yüz doğru değildir.

Q-Q plot grafiğini oluşturup yorumlayalım:

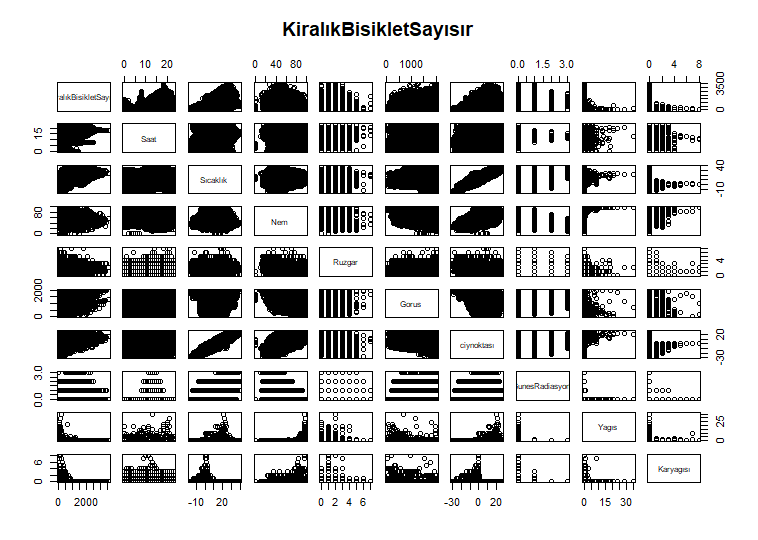




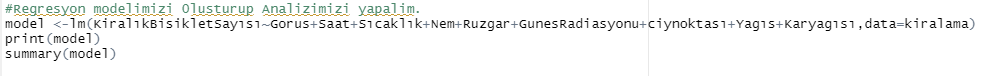
Buradan verimiz büyük bir kısmı çizgi üzerinden seyretmediğini söyleyebiliriz. Bu nedenle Q-Q plot grafiğine göre normal dağılıma uyumlu değildir. Verimiz çok büyük olduğu için buradan çıkardığımız sonuç bize tam olarak doğru bir kanıya varmamızı sağlamaz. Bu neden ile verimizi normal dağılıma uygun olarak kabul edeceğiz.

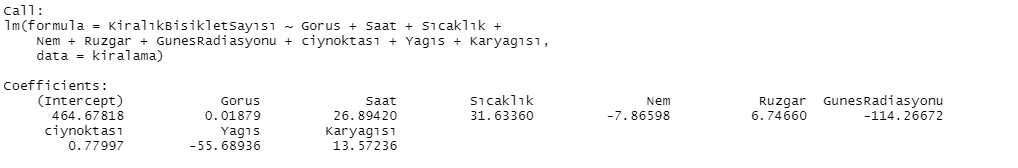
Verilerimiz arasındaki işkiyi görüntülemek için :





Artık regresyon modelimizi oluşturup gerekli analizimizi yapabiliriz. Regresyon modelimizi oluşturmak ve gerekli analizimizi yapmak için :

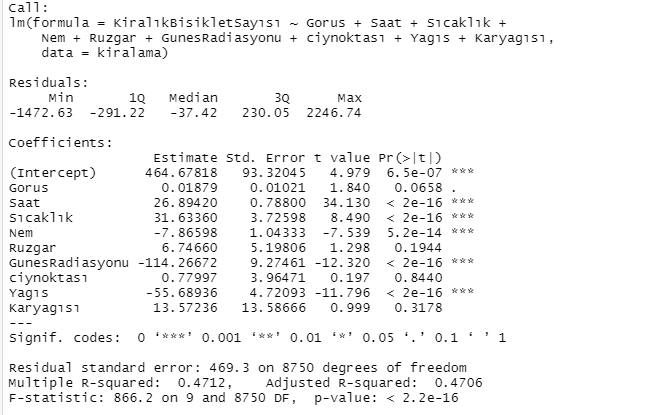




Buradan modelimiz :

Y=464.67818+ 0.01879\*x1+26.89420\*x2+31.63360\*x3-7.86598\*x4+6.74660\*x5-114.26672\*x6+0.77997\*x7-55.68936\*x8+13.57236\*x9

x1=Görüş, x2=Saat, x3=Sıcaklık, x4=Nem, x5=Rüzgar, x6=Güneş Radyasyonu, x7=Çiğ Noktası, x8=Yağış, x9=Kar yağışı



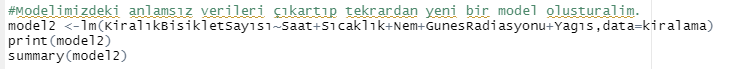
Buraya baktığımızda Pr(>| t | ) verilerimizin bazılarında “ \*\*\* “ gözlemlenir iken bazı verilerimizde hiçbir şey gözlemlenmemektedir. Bur da modelimizin anlamlılığından bahsetmektedir yani modelimizin anlamlılığını test etmek istediğimizde :

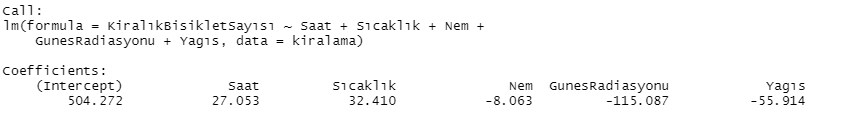
𝐻0:En az bir değişken Anlamsızdır. a=0.05

𝐻1:Model Anlamlıdır.

Görüş, Rüzgar, Çiy Noktası ve Kar Yağışı değişkenlerimizin için p> a olduğundan H0 kabul edilir. Modelimiz için anlamsız verilerdir.

Modelimizi anlamlı bir hale getirebilmemiz için modelden anlamsız olan verileri çıkarmamız gerekir. Bu neden ile anlamsız değişkenlerimizi çıkartıp yeni bir model oluşturmak için:

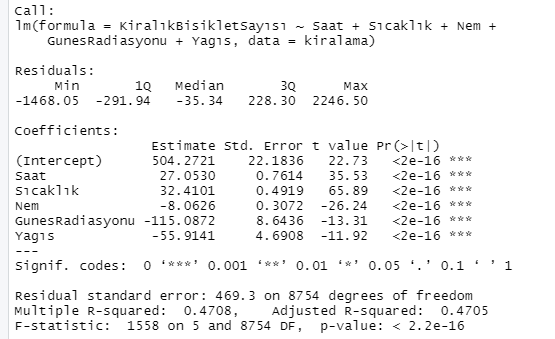




Buradan modelimiz :

Y=504.272+ 27.053\*x1+32.410\*x2-8.063\*x3-115.087\*x4-55.914\*x5

x1=Saat, x2=Sıcaklık, x3=Nem, x4=Güneş Radyasyonu, x5=Yağış



Bur da oluşturduğumuz yeni model için anlamlılık testi yapalım.

𝐻0:En az bir değişken anlamsızdır. a=0.05

𝐻1:Model Anlamlıdır.

Bütün p değerleri için p < a olduğundan H0 hipotezi reddedilir. Buradan %95 güven düzeyinde modelimiz anlamlıdır.

Residual Standart error (Artık Standart Hata) değerini bağımlı değişkenin ortalamasına böldürerek modelimizin artık hatalarını elde edebiliriz. Bu veri için artık standart hata oranı yaklaşık olarak %66.605 dir. Buradan modelimizn hata oranının oldukça yüksek olduğu kanısına varabilirz.

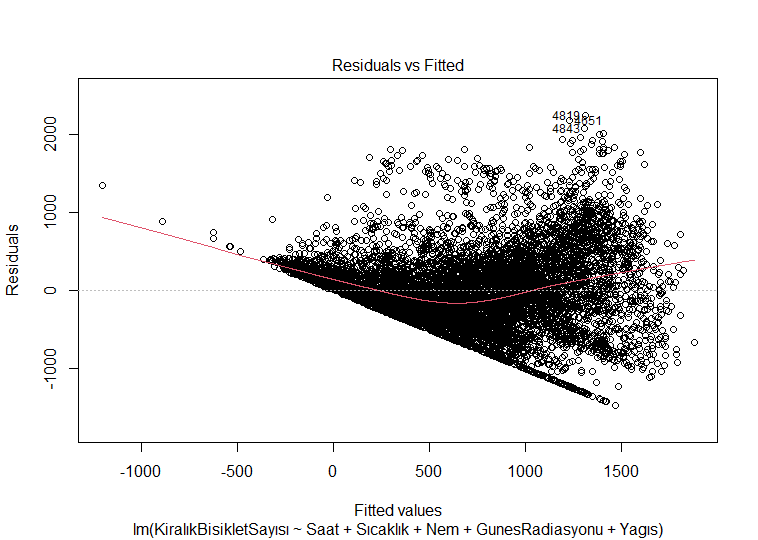
Adjusted R-Squared yani düzeltilmiş regresyon karesine bakarak bağımlı değişken, bağımsız değişkenlerin %47.05 ini açıklar.

F-statistic: p-value değerine bakarak modelimiz anlamlı olup olmadığını kontrol edebiliriz.Buradan p değerimiz 0.05 den oldukça küçük olduğu için modelimiz anlamlıdır.

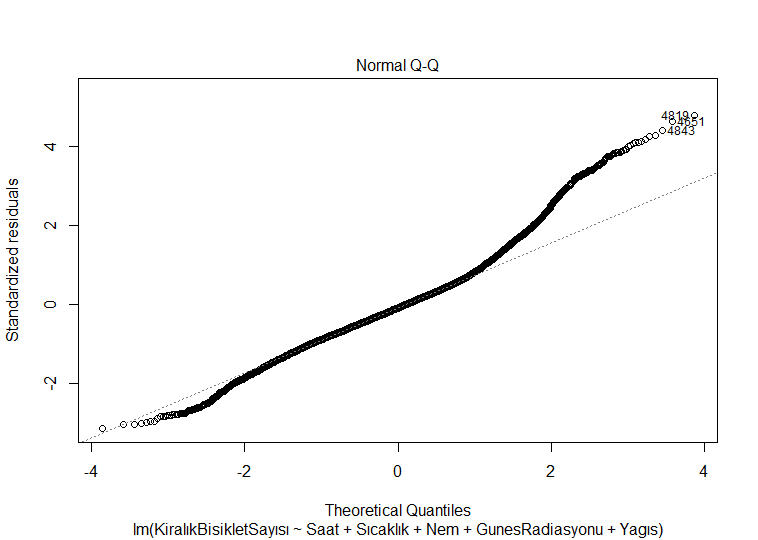
Şimdi de oluşturduğumuz modelin grafiğini yorumlayalım bunun için öncelikle:



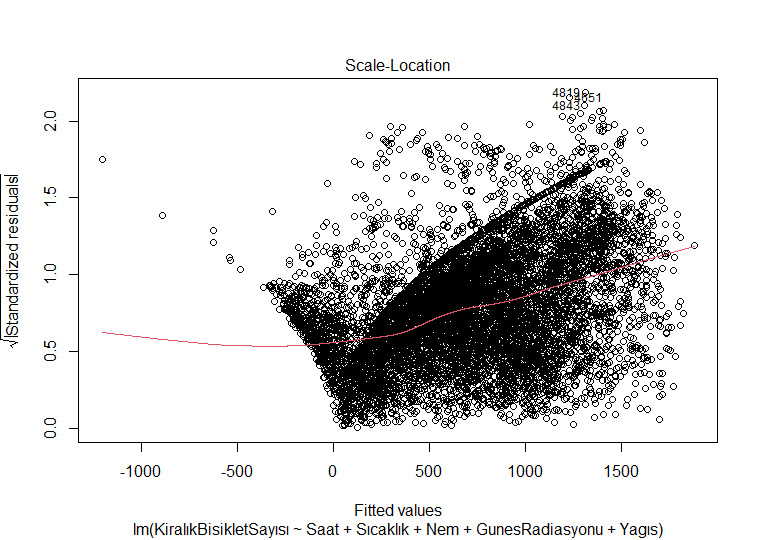
Kodunu yazıp konsolumuzda enter tuşuna basarak 4 tane grafik elde edebiliriz.



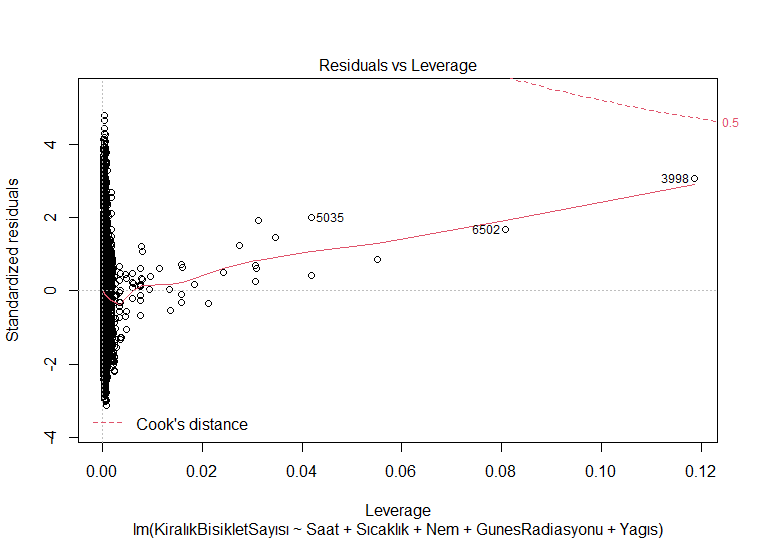
Artıklar ve uydurulmuş grafik nispeten düz ve homoskedastik görünmektedir. Bununla birlikte, sol altta bu tuhaf kesinti var, bu da bana homoskedastisiteyi sorgulamama neden oluyor.



Bu grafik de çoğunlukla ilgili verilerimiz doğrusal bir çizgi üzerinde olduğu için modelimizin normal dağılıma uygun olduğunu söyleyeiliriz.



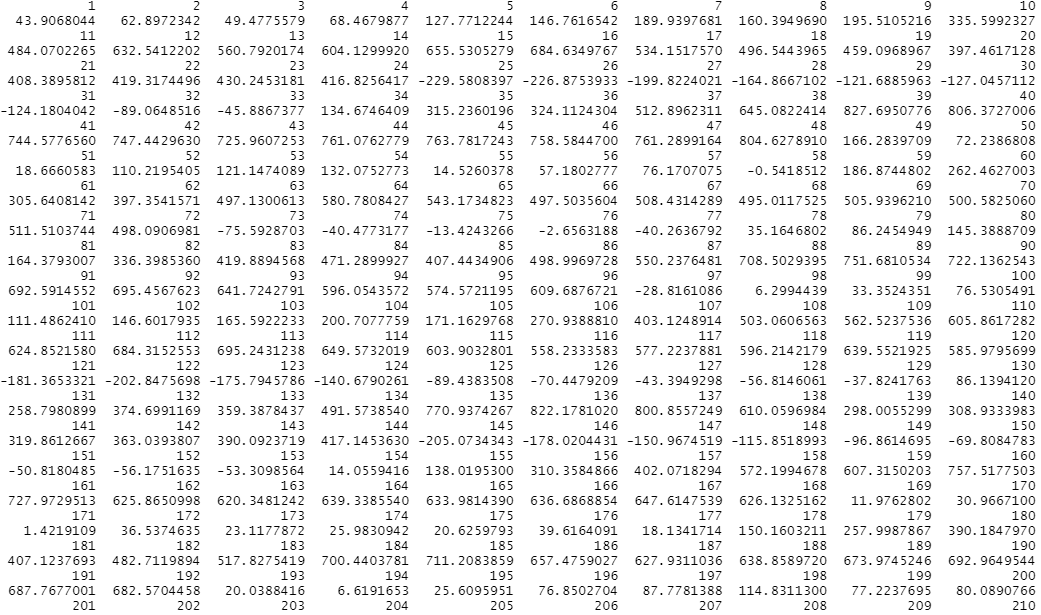
Ölçek-konum grafiği, x ekseni boyunca regresyon modelinin uygun değerlerini ve y ekseni boyunca standart artıkların karekökünü gösteren bir grafik türüdür. Modelimizde artıklar, tüm uygun değerlerde kabaca eşit değişkenlik ile kırmızı çizgi etrafında rastgele dağıldığı için artıkların dağılımı tüm uygun değerlerde kabaca eşit değildir.



Standartlaştırılmış kalıntıların yayılmasının kaldıraç veya takılanın duyarlılığı olarak nasıl değiştiğine bakıyoruz. İlk olarak, bu aynı zamanda heteroskedastisite ve lineer olmamayı tespit etmek için de kullanılabilir. Standartlaştırılmış artıkların yayılımı, kaldıracın bir fonksiyonu olarak değişmemelidir:  burada azalıyor gibi görünmektedir ve heteroskedastisiteye işaret etmektedir.

Şimdi de verimizin öngörü değerlerini bulduralım. Bunun için:





Verim çok büyük olduğu için R’ ın konsol ekranında ilk 1000 öngörü değerini gördüm. Buradan öngörü değerlerimin gerçek değerlerine benzer olmadığını gözlemledim. Modelimin yeterince anlamlı olmadığı kanısına vardım.