HACETTEPE ÜNİVERSİTESİ

Ramazan Erduran - 21821809

22 NİSAN 2022

Regresyon Dönem Ödevi

İçindekiler Tablosu

Verileri İçeri Aktarma

```
library(readr)
hwdata <- read_table2("D:/Hacettepe/Regresyon/Ödev/hwdata.txt")
head(hwdata)

## # A tibble: 6 x 5
##   `"y"` `"x1"` `"x2"` `"x3"` `"x4"`
##   <dbl> <dbl> <dbl> <dbl> <dbl>
## 1  17.6  8.04  3.12  2.39     1
## 2  18.5  7.28  2.07  2.49     1
## 3  23.8  7.67  3.00  0.528    1
## 4  16.8  8.01  2.34  1.89     1
## 5  19.4  7.59  1.92  0.738    1
## 6  17.5  6.49  2.74  2.43     1

names(hwdata) <- c("Y", "X1", "X2", "X3", "X4")
hwdata$X4 <- as.factor(hwdata$X4)
class(hwdata$X4)

## [1] "factor"
```

Verilerimdeki değişkenleri **Y, X1, X2, X3 ve X4** olarak isimlendirdim. **X4** değişkenini ise factor olarak belirttim ki kategorik değişkeni uygun bir şekilde kullanabileyim.

Part 1. Senaryo

Verilerim Y bağımlı değişkeni, X1, X2, X3 bağımsız nitel ve X4 bağımsız 3 katlı kategorik değişkenden oluşuyor.

Araç satma ve kiralamaya ilişkin bir şirketimiz olduğunu varsayalım. 2. el araçlar hem kiralanabiliyor hem de satılabiliyor. Sıfır araçlar ise sadece satılabiliyor kiralamada kullanılamıyor. Ayrıca Borsa İstanbul 100 endeksine bağlı hisselerimizi halkın alım satımına açmış bir şirket olalım. Bizim elimizdeki verilerde Y bağımlı değişkenimiz ise tüm bu süreçten kazandığımız gelir olsun (MilyonTL). 150 aylık bir veri setimiz olduğunu düşünelim:

DEĞİŞKENLER	DEĞİŞKEN SENARYOSU
Y	Elde edilen toplam gelir (MilyonTL)
X1	Araç satışı ve kiralamadan gelen gelir (BinTL)
X2	Şirket hisse fiyatlarındaki değişim (TL)
X3	Enflasyondaki değişim (%)
X4-1	LPG ile çalışan araçlar
X4-2	Benzin ile çalışan araçlar
X4-3	Dizel ile çalışan araçlar

Part 2. Tanımlayıcı İstatistikler

```
summary(hwdata)

##       Y               X1              X2              X3         X4
##  Min.   : 4.006   Min.   : 3.590   Min.   :-1.119   Min.   :-2.2333   1:50
##  1st Qu.:10.019   1st Qu.: 6.425   1st Qu.: 1.302   1st Qu.: 0.4236   2:50
##  Median :13.306   Median : 6.987   Median : 1.976   Median : 0.9662   3:50
##  Mean   :13.930   Mean   : 7.063   Mean   : 2.030   Mean   : 1.0279
##  3rd Qu.:15.971   3rd Qu.: 7.829   3rd Qu.: 2.742   3rd Qu.: 1.7082
##  Max.   :64.394   Max.   :10.163   Max.   : 5.294   Max.   : 3.6371
```

150 Aylık veri setinden elde edilen tanımlayıcı istatistiklere göre;

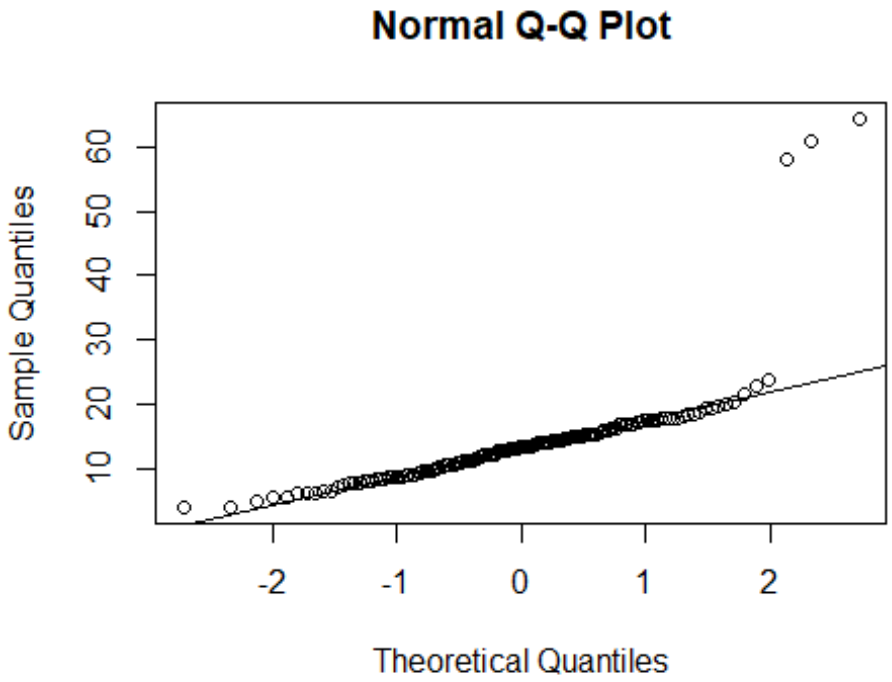
- Ortalama **Y = Elde edilen ortalama toplam gelir** 13.930 MilyonTL/Ay olarak bulunmuş.
- 150 Aylık süre içerisindeki ortalama **X1 = Araç satışı ve kiralamalardan gelen gelir** aylık 7.063 BinTL hesaplanmış.
- Hisse senetlerimizdeki (**X2 = Şirket hisse fiyatlarındaki değişim**) durum ise aylık ortalama 2.030 TL olarak hesaplanmış. (Hissedarlarımız hisselerini satsaydı eksi değer alırdı, tam tersi hisselerimizin alınması ise bize artı değer kazandırırdı bundan dolayı bu değişkenimiz hem + hem - değer alabiliyor.)
- **X4 = Enflasyondaki değişim** değerlerinin ortalamasına baktığımızda ise 150 aylık süreçte enflasyon ortalama %1.0279 artmış.

Part 3. Doğrusallık ve Normallik Testi

Normallik testi:

Normallik testini yapmadan önce görsel olarak görebilmek adına verilerin saçılım grafiği üzerine normal dağılım eğrisi çizip baktım.

```
attach(hwdata)
qqnorm(Y)
qqline(Y)
```



Daha sonra emin olmak için sayısal olarak da kontrol ettim. Bu kısımda shapiro.test() fonskiyonunu kullandım.

```
shapiro.test(Y)

##
##  Shapiro-Wilk normality test
##
## data:  Y
## W = 0.5937, p-value < 2.2e-16
```

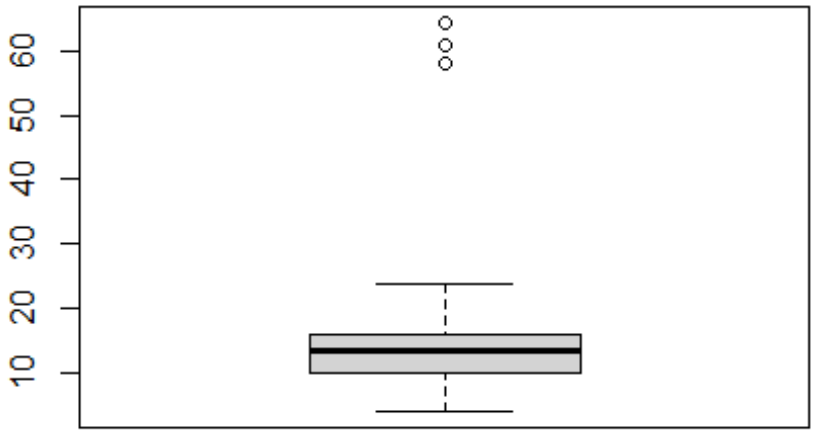
Hipotezlerim;

H0: Verilerin dağılımı ile normal dağılım arasında fark yoktur.
Hs: Verilerin dağılımı ile normal dağılım arasında fark yoktur.

Shapiro test sonucumda p-value değerim α = 0.05 'ten küçük olduğu için yokluk hipotezimi reddedip verilerimin normal dağılmadığı kanısına vardım.

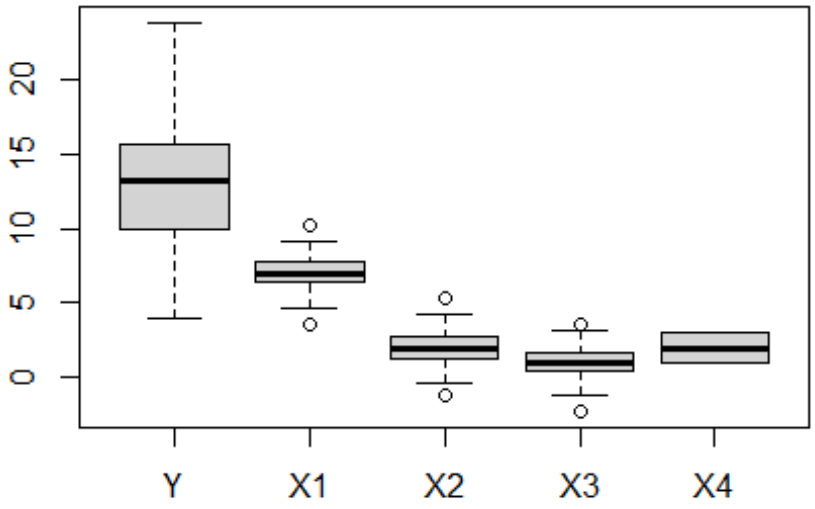
Boxplot grafiği ile aykırı değer var mı ona baktım. Aykırı değerleri temizlemenin normal dağılım varsayımında bana yardımcı olabileceğini düşündüm.

```
boxplot(Y)
```



Aykırı değer temizliği:

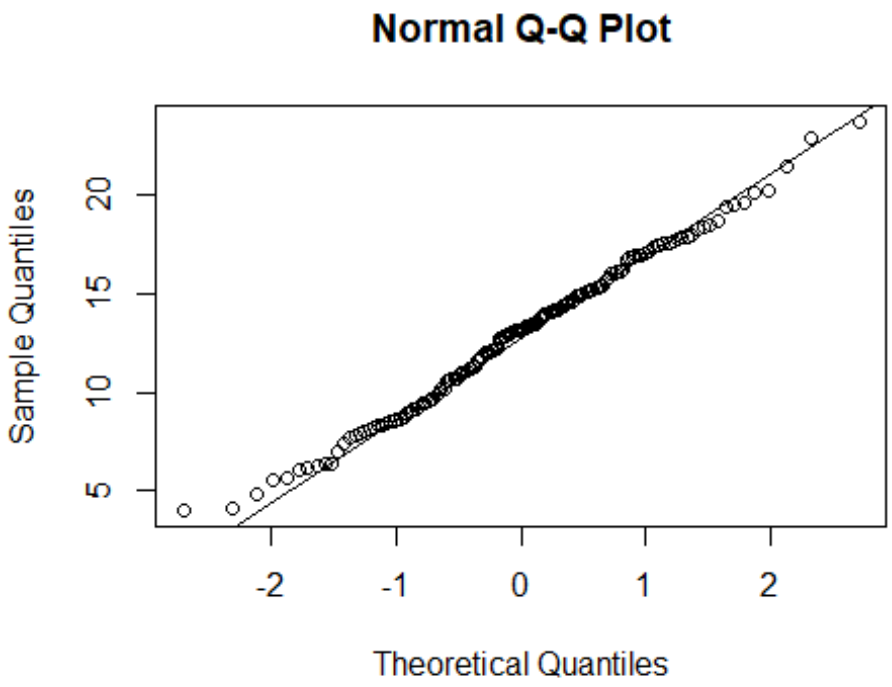
```
Q1 <- quantile(Y, 0.25)
Q3 <- quantile(Y, 0.75)
IQR <- IQR(Y)
hwData <- subset(hwdata, Y > (Q1 - 1.5*IQR) & Y < (Q3 + 1.5*IQR))
boxplot(hwData)
```



```
detach(hwdata)
```

Aykırı değerlerini temizlediğim verilerin normal dağılıp dağılmadığını kontrol etmek için tekrardan test yaptım:

```
qqnorm(hwData$Y)
qqline(hwData$Y)
```



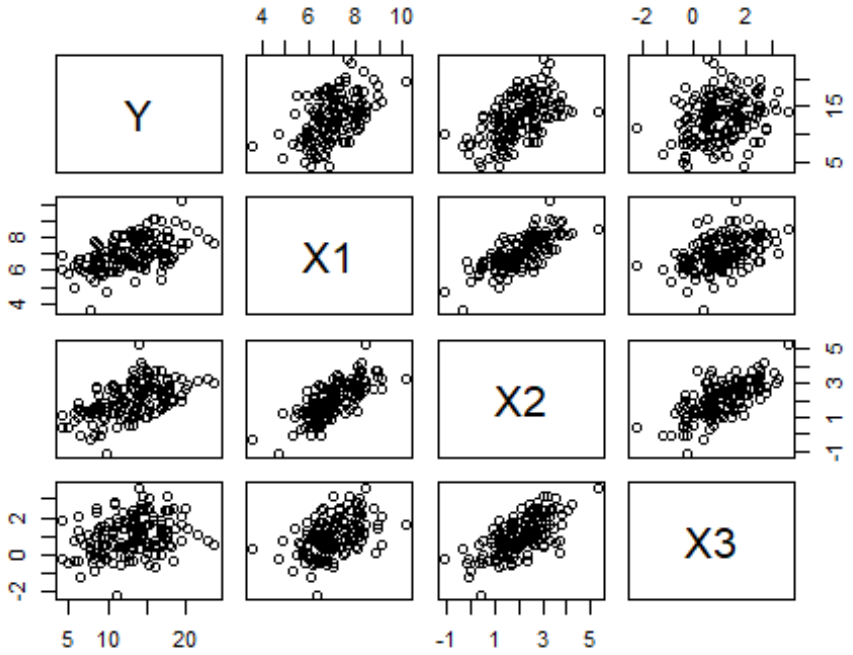

```
shapiro.test(hwData$Y)

##
##  Shapiro-Wilk normality test
##
## data:  hwData$Y
## W = 0.99303, p-value = 0.6957
```

Gerek grafikten gerekse shapiro.test() sonucunda görebildiğimiz üzere, p-value: 0.6957 > α=0.05 için yokluk hipotezi reddedilemez, yani %95 güvenle söyelenebilir ki verilerimiz normal dağılıyor.

Doğrusallık:

X4 Değişkeni kategorik bir değişken olduğu için doğrusallık incelemesi yapılamaz bundan dolayı grafikte yer vermedim.

```
pairs(hwData[1:4])
```



Dağılım grafiklerine baktığımızda X1 X2 ve X3 değişkenleri ile Y arasında belli derecede doğrusal bir ilişki olduğunu söyleyebiliriz. Ancak burada bağımsız değişkenlerimiz arasında önemli derecede doğrusal bir ilişki de olabilir. Bundan dolayı Bağımsız değişkenlerim arasındaki doğrusallığa baktım:

```
result <- lm(hwData$X1~hwData$X2)
a<-summary(result)
sqrt(a$r.squared)

## [1] 0.7391816

result <- lm(hwData$X1~hwData$X3)
a<-summary(result)
sqrt(a$r.squared)

## [1] 0.4812473

result <- lm(hwData$X2~hwData$X3)
a<-summary(result)
sqrt(a$r.squared)

## [1] 0.6856835
```

Doğrusallığını izlediğim ve bağımsız değişkenlerimizin kendi aralarında bağımsız olma konusunda tabloki sonuçları elde ettim:

DEĞİŞKENLER	İLİŞKİ DERECELERİ
X1 ile X2	*0.7391816*
X1 ile X3	*0.4812473*
X2 ile X3	*0.6856835*

Ben burada modelden çıkarıp çıkarmama veya herhangi bir işlem yapma konusunda önce regresyon analizi yapıp modeldeki anlamlılıklarına bakıp daha sonra bu aralarındaki doğrusallığın sorun olup olmadığına başka işlemler ile karar vereceğim.

Part 4. Artık İncelemesi

```
sonuc<-lm(hwData$Y~hwData$X1+hwData$X2+hwData$X3+hwData$X4)
info <- ls.diag(sonuc)
info

## $std.dev
## [1] 1.965013
## $hat
##   [1] 0.03383974 0.04602691 0.04612694 0.03652923 0.02846964 0.04744609
##   [7] 0.02641277 0.02756351 0.04100784 0.04577654 …
## $std.res
##   [1] -0.37565145  1.81537184  2.08384333 -0.15003766  1.29464400  0.89868623
##   [7]  0.91565324  1.80250255  0.73193698  0.52498066 …
## $stud.res
##   [1] -0.37450443  1.83044070  2.10917294 -0.14951660  1.29778149  0.89806946
##   [7]  0.91512531  1.81715741  0.73072637  0.52362772 …
## $cooks
##   [1] 8.237524e-04 2.650060e-02 3.499798e-02 1.422497e-04 8.186062e-03
##   [6] 6.704645e-03 3.790966e-03 1.534878e-02 …
## $dfits
##   [1] -0.070088356  0.402062463  0.463814409 -0.029113229  0.222159228
##   [6]  0.200431216  0.150730131 …
## $correlation
##             (Intercept)   hwData$X1   hwData$X2    hwData$X3  hwData$X42
## (Intercept)  1.000000000 -0.95544850  0.46481666 -0.005912092 -0.05152794
## hwData$X1   -0.955448501  1.00000000 -0.63805564 …
##             hwData$X43
## (Intercept) -0.06104148
## hwData$X1   -0.04979925
## hwData$X2   -0.07922420
## hwData$X3    0.10558969
## hwData$X42   0.48316678
## hwData$X43   1.00000000
##
```

```
## $std.err
## (Intercept) 1.4467199
## hwData$X1   0.2490761
## hwData$X2   0.2843687
## hwData$X3   0.2244825
## hwData$X42  0.4026622
## hwData$X43  0.3976858
## $cov.scaled
##             (Intercept)    hwData$X1    hwData$X2    hwData$X3   hwData$X42
## (Intercept)  2.09299837 -0.344289471  0.191226445 -0.001920030 -0.030017056
## hwData$X1   -0.34428947  0.062038891 …
## $cov.unscaled
##             (Intercept)     hwData$X1    hwData$X2     hwData$X3    hwData$X42
## (Intercept)  0.542048558 -0.0891647189  0.049524176 -0.0004972530 -0.0077738722
## hwData$X1   -0.089164719  0.0160669458 …
```

Aykırı Değer (ri):

(-2,+2) aralığında olmayan Standartlaştırılmış artıklar, (-3,+3) arasında olmayan Student tipi artıklar aykırı değerdir.

```
aykiri <- list()
for (i in info$std.res){
  if (i<-2 & i>2){
    aykiri <- append(aykiri,which(info$std.res == i))
  }
}
aykiri

## list()

for (i in info$stud.res){
  if (i<-3 & i>3){
    aykiri <- appned(aykiri,which(info$std.dev == i))
  }
}
aykiri

## list()
```

aykiri isimli listemiz boş döndü. Gerek Standartlaştırılmış hatalar gerekse Student tipi hatalarda aykırı değer varlığına dair incelememizin sonucu olarak söyleyebiliriz ki verilerimizde aykırı değer yoktur. Normallik testinden sonra aykıır değer temizliği yapmıştık zaten :)

Cook Uzaklığı(Di):

Tüm gözlemler üzerinden bulunan β kestirimi ile i. gözlem çıkarıldıktan sonraki β kestirimi arasındaki farkın bir ölçüsüdür. Aykırı değerim olmadığı için aykırı değerime ilişkin Cook uzaklığını da bulamam.

Part 5. Kestirim Denklemi

Regresyon Analizi:

```
sonuc<-lm(hwData$Y~hwData$X1+hwData$X2+hwData$X3+hwData$X4)
regression_model <- summary(sonuc)
regression_model
## Call:
## lm(formula = hwData$Y ~ hwData$X1 + hwData$X2 + hwData$X3 + hwData$X4)
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.1262 -1.1206  0.2143  1.2867  3.9992
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   5.4312     1.4467   3.754 0.000254 ***
## hwData$X1     1.0369     0.2491   4.163 5.44e-05 ***
## hwData$X2     2.3206     0.2844   8.160 1.67e-13 ***
## hwData$X3    -1.1229     0.2245  -5.002 1.66e-06 ***
## hwData$X42   -3.7608     0.4027  -9.340  < 2e-16 ***
## hwData$X43   -6.3051     0.3977 -15.855  < 2e-16 ***
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## Residual standard error: 1.965 on 141 degrees of freedom
## Multiple R-squared:  0.7644, Adjusted R-squared:  0.756
## F-statistic: 91.48 on 5 and 141 DF,  p-value: < 2.2e-16
```

Yukarıdaki sonuçlardan yola çıkarak;
Elde edilen toplam gelirdeki değişimin %76,44'ü "*Araç satışı ve kiralamadan gelen gelirler*", "*Şirket hisse fiyatlarındaki değişim*", "*Enflasyondaki değişim*" ve "*Araçların yakıt türleri*" değişkenlerince açıklanabilmektedir. Açıklanamayan kısım için farklı etkenler(değişkenler) olabilir.

Model anlamlılığı:

H0: $\beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5$
Hs: En az bir β_j farklıdır.

p-value: 2.2e-16 < α=0.05 için küçük olduğu için %95 güvenle söyleyebiliriz ki modelimiz anlamlıdır.

Regresyon model denklemi:

Model denklemi:

$$y_i = \underset{(1.4467)}{5.4312} + \underset{(0.2491)}{1.0369}\boldsymbol{X1} + \underset{(0.2844)}{2.3206}\boldsymbol{X2} - \underset{(0.2245)}{1.1229}\boldsymbol{X3} - \underset{(0.4027)}{3.7608}\boldsymbol{X42} - \underset{(0.3977)}{6.3051}\boldsymbol{X43} \pm 1.965$$

Burada ilk başta şunu söyleyebiliriz ki kılavuz değişkenimiz LPG ile çalışan araçlar değişkenidir. (X42=0 , X43=0)

Part 6. Katsayı Anlamlılıkları ve Yorumları

Kısmi F testleri:

H0: $\beta_0 = 0$
Hs: $\beta_0 \neq 0$

p-value: 0.000254 α=0.05 için H0 reddedilir ve %95 güvenle söylenebilir sabit değerimizin modele katkısı anlamlıdır. Ayrıca değerlerimiz Çok ütopik durumlar olmadıığı sürece (enflasyonun%0 , araç satışının %0 olması vs gibi) ki ben burada bunların 0 olamayacağını varsayarak "sabit değerimiz denklem gereği katsayıdır" yorumunu yapıyorum.

H0: $\beta_1 = 0$
Hs: $\beta_1 \neq 0$

p-value: 5.44e-05 α=0.05 için H0 reddedilir ve %95 güvenle söylenebilir ki araç satışı ve kiralamadan gelen gelirin modele katkısı anlamlıdır.

Diğer tüm değişkenler sabit tutulduğunda araç satış ve kiralamadan gelen gelir 1 birim (1000TL) artarsa bizim gelirimizi ortalama olarak 1.0369 MilyonTL arttırır.

H0: $\beta_2 = 0$
Hs: $\beta_2 \neq 0$

p-value: 1.67e-13 α=0.05 için H0 reddedilir ve %95 güvenle söylenebilir ki şirket hisse fiyatlarındaki artış veya azalış, elde edilen toplam geliri açıklamada önemli bir değişkendir.

Diğer tüm değişkenler sabit tutulduğunda şirketimizin hisse fiyatlarındaki 1 Lot 'luk (1 Birim) bir artış toplam gelirimizi ortalama olarak aylık 2.3206 MilyonTL arttırır.

H0: $\beta_3 = 0$
Hs: $\beta_3 \neq 0$

p-value: 1.66e-06 α=0.05 için H0 reddedilir ve %95 güvenle söylenebilir ki enflasyon etmeninin modele katkısı anlamlıdır.

Diğer tüm değişkenler sabit tutulduğunda enflasyondaki %1 'lik bir artış gelirimizi aylık ortalama olarak 1.1229 MiyonTL azaltır.

H0: $\beta_4 = 0$

Hs: $\beta_4 \neq 0$

p-value: 2e-16 α=0.05 için H0 reddedilir ve %95 güvenle söylenebilir ki araç yakıt türü kategorik değişkenindeki LPG'li araçlar ile benzinle çalışan araçlar arasında farklılık vardır.

H0: $\beta_5 = 0$

Hs: $\beta_5 \neq 0$

p-value: 2e-16 α=0.05 için H0 reddedilir ve %95 güvenle söylenebilir ki X4 kategorik değişkenindeki LPG'li araçlar ile dizel ile çalışan araçlar arasında farklılık vardır.

Araç yakıt türlerine baktığımızda LPG ile çalışan araçlar, benzin ve dizel ile çalışan araçlardan farklıdır. Ayrıca amacımız elde edilen geliri maksimize etmeye çalışmak olduğundan dolayı, bizim için LPG'li araçlar diğer araçlardan daha karlıdır diyebiliriz.

Part 7. Belirtme Katsayısı

O ay elde edilen toplam gelirin %76,44 'ü araç satışı ve kiralamadan gelen gelir, şirket hisselerindeki fiyat değişimi, enflasyondaki değişim ve kiralanan veya satılan araçlara ilişkin yakıt türü gibi değişkenlerce açıklanabilmektedir. Geriye kalan %23,56 'lık kısım ise elimizde olmayan etmenlerce açıklanabilmektedir.

Part 8. Güven Aralıkları

Katsayılara ilişkin güven aralıkları:

```
confint(sonuc, level = .99)
```

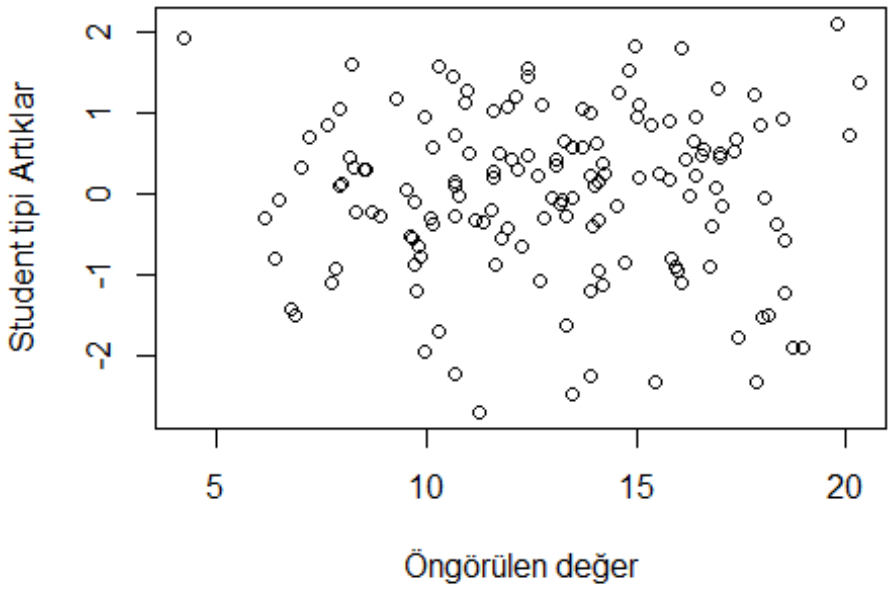
```
##                0.5 %     99.5 %
## (Intercept)  1.6535808  9.2087778
## hwData$X1    0.3865251  1.6872737
## hwData$X2    1.5780604  3.0631175
## hwData$X3   -1.7090249 -0.5367113
## hwData$X42  -4.8122487 -2.7094279
## hwData$X43  -7.3435561 -5.2667236
```

%99 güvenirlikle hesaplanan bu güven aralıkları modelimizdeki değişkenlerimizin aylık gelirimize olan etkilerinin ortalama olarak hangi aralıkta olduğunu göstermektedir.

- **(X1)** Araç satışı ve kiralamadan gelen gelir 1 birim (1000TL) artarsa eğer diğer değişkenler sabit tutulduğunda bizim gelirimizi 386.52,1TL ile 1.687.273,7TL arttırmaktadır.
- **(X2)** Hisse senetlerindeki durumda ise yine aynı şekilde diğer değişkenlerin sabit tutulması koşuluyla hisse senetlerimizdeki 1 birimlik artış (1Lot başına 1TL 'lik artış) aylık gelirimizi 1.578.060,4TL ile 3.063.117,5TL arasında bir değerde arttırmaktadır.
- **(X3)** Diğer tüm değişkenler sabit tutulduğunda enfasyondaki %1 lik bir artış bizim gelirimizi 1.709.024,9TL ile 536.711,3 TL arasında bir değerde azalmaktadır.
- **(X4)** Kategorik değişkenlerimiz hususunda ise diğer değişkenlerimiz sabitse, aracın benzinle çalışması aylık geliri 4.812.248,7TL ile 2.709.427,9TL arasında, dizel ile çalışması aylık geliri 7.343.556,1TL ile 5.266.723,6TL azaltmaktadır.

Part 9. Değişen Varyanslılık İncelemesi

```
plot(predict(sonuc), info$stud.res, ylab= "Student tipi Artıklar", xlab="Öngörülen değer")
```



Student tipi artıklar ile öngörülen değerlere ilişkin saçılım grafiğini incelediğimizde yapının rastgele olduğunu söyleyebiliriz. Burdan yola çıkarak Artıkların rastgele dağıldığı yorumunu yapabiliriz.

```
library(lmtest)
bptest(sonuc)

##
##  studentized Breusch-Pagan test
##
## data:  sonuc
## BP = 9.069, df = 5, p-value = 0.1063
```

Ayrıca Breusch-Pagan testi sonucuna göre;

H0: Değişen varyanslılık yoktur.
Hs: Değişen varyanlılık vardır.

p-value = 0.1063 > α=0.05 için yokluk hipotezi reddedilemez yani değişen varyanslılık sorunu yoktur.

Part 10. Öz İlişki Sorunu

```
library(lmtest)
dwtest(sonuc)

##
##  Durbin-Watson test
##
## data:  sonuc
## DW = 1.0031, p-value = 1.758e-10
## alternative hypothesis: true autocorrelation is greater than 0
```

Model özeti tablosundan **"d=1.0031"** bulunmuştur. **"d=2(1-r)"** formülü dikkate alındığında ve tablodan d_{lower} ve d_{upper} değerleri alınıp bir tablo oluşturulduğunda *(k=4 ve n=147 için)*:

Variable	Value
d_{lower}	*1.68*
d_{upper}	*1.79*

Tablodan yola çıkarak **[1.68 ; 1.79]** aralığı bizim için kararsızlık bölgesidir. Ancak elimizdeki Durbin-Watson değerimiz DW = 1.0031 bu kararsızlık bölgesinde yer almadığı ve **[0;2]** aralığında yer aldığı için öz ilişki vardır deriz. Buna ek olarak kuracağımız seçenek hipotezimizde ise pozitif yönlü kurmalıyız:

H_o = Öz ilişki yoktur.
H_s = Pozitif öz ilişki vardır.

Durbin-Watson test sonucunda elde ettiğimiz p-value = 1.758e-10 **<** α=0.05 değerleri için söyleyebiliriz ki, %95 güvenle pozitif öz ilişki vardır.

Part 11. Çoklu Bağlantı Sorunu

```
library("Hmisc")
res2 <- rcorr(as.matrix(hwData[2:5]))
res2$r

##          X1       X2        X3        X4
## X1 1.0000000 0.7391816 0.48124728  0.10858837
## X2 0.7391816 1.0000000  0.68568350  0.09581460
## X3 0.4812473 0.6856835  1.00000000 -0.01248072
## X4 0.1085884 0.0958146 -0.01248072  1.00000000
```

Değişkenlerimizin arasındaki ilişki incelendiğinde, X1 ve X2 arasında yüksek bir ilişki olduğu görülmekte. Bu durumda çoklu bağlantı olabilir. Bir diğer korelasyon inceleme yöntemi olarak:

```
cor(hwData[2:4], method = "pearson")

##          X1       X2       X3
## X1 1.0000000 0.7391816 0.4812473
## X2 0.7391816 1.0000000 0.6856835
## X3 0.4812473 0.6856835 1.0000000
```

Bu şekilde de X1 ve X2 arasında yüksek bir ilişki olduğunu söyleyebiliriz fakat bakalım bu ilişki düzeyi bizi etkiliyor mu?

```
Library(DAAG)
sonuc <- lm(hwData$y~hwData$x1 + hwData$x2 + hwData$x3 + hwData$x4)
vif(sonuc)

##
## hwData$X1  hwData$X2  hwData$X3 hwData$X42 hwData$X43
##    2.2334       3.2991       1.9156       1.3426       1.3514
```

vif() ile çoklu bağlantı var mı diye baktığımda o kadar da büyük vif değerleri elde edemedim, çoklu bağlantı sorunu olabilir diyebilmek için.

Son olarak bu sorulara cevap verebilmek, çoklu bağlantı sorunu vardır veya yoktur diyebilmek için colldiag() ve ols_eigen_cindex() fonksiyonlarını kullandım. İki yöntemi de göstermek istedim:

```
library(perturb)
colldiag(model.matrix(sonuc),add.intercept=FALSE)

## Condition
## Index    Variance Decomposition Proportions
##          (Intercept) hwData$X1 hwData$X2 hwData$X3 hwData$X42 hwData$X43
## 1  1.000 0.001       0.000     0.003     0.010     0.009       0.011
## 2  2.052 0.000       0.000     0.000     0.000     0.287       0.205
## 3  2.973 0. 001       0.000      0.007      0.271      0.144       0.266
## 4  4.505 0.012       0.006     0.006     0.246     0.521       0.518
## 5  7.807 0.030       0.002     0.614     0.468     0.037       0.000
## 6 29.164 0.956       0.991     0.370     0.004     0.002       0.000
```

```
library(olsrr)
ols_eigen_cindex(sonuc)

##   Eigenvalue Condition Index    intercept    hwData$X1    hwData$X2
## 1 4.23248799       1.000000 6.492089e-04 4.458218e-04 0.0032267962
## 2 1.00563003       2.051534 2.744917e-05 7.598666e-06 0.0001785427
## 3 0.47893160       2.972769 9.040309e-04 2.021343e-04 0.0068424461
## 4 0.20852317       4.505269 1.182631e-02 5.987210e-03 0.0055019709
## 5 0.06945111       7.806534 3.049607e-02 2.388053e-03 0.6143028529
## 6 0.00497610      29.164417 9.560969e-01 9.909692e-01 0.3699473912
##     hwData$X3  hwData$X42   hwData$X43
## 1 0.0101307562 0.008963669 1.101448e-02
## 2 0.0004497329 0.287115178 2.051181e-01
## 3 0.2711077016 0.143526265 2.660812e-01
## 4 0.2463663319 0.521320353 5.177619e-01
## 5 0.4683967105 0.037050125 1.944119e-05
## 6 0.0035487668 0.002024409 4.919151e-06
```

Koşul sayısı 30'dan büyük olduğu durumda çoklu bağlantıdan etkilenilmektedir. Fakat bizim yüksek derecede korelasyona sahip bağımsız değişkenlerimiz olsa da çoklu bağlantı sorunu var diyebileceğimiz değişkenler değilmiş. Çünkü Koşul başlığı altındaki değerlerin hiçbiri 30'dan büyük değil.

Part 12. Uyum Kestirimi

Veri kümesindeki x_i değerlerine karşılık gelen y_i kestirimi, *uyum kestirimi* adını alır.
Bu tanıma ilişkin y_i kestirimini yapmak üzere "*X1=7.035412 , X2=2.29930316 , X3=1.417905782 , X42=0 , X43=0*" değerlerini aldım.

```
yi <- 5.4312 + 1.0369*7.035412 + 2.3206*2.29930316 - 1.1229*1.417905782
yi-1.965

## [1] 14.50482

yi+1.965

## [1] 18.43482
```

Kestirim denklemi ile hesaplanarak %95 güvenle **[14.50482;18.43482]** sonucu bulunur.
Yani, araç satışı ve kiralamadan gelen gelir 7.035412 BinTL, hisse senetleri 2.29930316TL ve enflasyon %1.417905782 arttığında LPG'li bir aracın aylık olarak bize getirisi ortalama 14.50482 MilyonTL ile 18.43482 MilyonTL arasında olacaktır.

Part 13. Ön Kestirim

Veri kümesinde bulunmayan yeni $\tilde{x}_i$ gözlemleri için $\tilde{y}$ değerinin kestirimi, önkestirim adını alır. Önkestirimini yapmak üzere "*X1=11 , X2=-2 , X3=-2.5 , X42=1 , X43=0"* değerlerini aldım.

```
yi <- 5.4312 + 1.0369*11 - 2.3206*2 - 1.1229*2.5
yi-1.965

## [1] 13.03815

yi+1.965

## [1] 16.96815
```

%95 güvenle **[13.03815;16.96815]** sonucu bulunur.
Yani, araç satışı ve kiralamadan gelen gelir 11 BinTL arttığında, hisse senetleri 2TL değer kaybettiğinde ve enflasyon %2.5 azaldığında benzinle çalışan bir aracın aylık olarak bize getirisi ortalama 13.038,15TL ile 16.968,15TL arasında olacaktır.

Part 14. Uyum ve Ön Kestirime İlişkin Beklenen Değerlerin Güven Aralıkları

Uyum kestirimi için:

$$P(\hat{y}_i - t_{\frac{\alpha}{2},n-k-1} - S_{\hat{y}_i} \leq E(\hat{y}_i) \leq \hat{y}_i + t_{\frac{\alpha}{2},n-k-1} - S_{\hat{y}_i}) = 1 - \alpha$$

$$S\hat{y} = \hat{\sigma} \sqrt{x_i (X'X)'^{-1} x'_i}$$

```
predict(sonuc ,interval="confidence", level=0.95)[13,]

##       fit       lwr       upr
## 13.35064 12.57582 14.12547
```

predict() kodunu kullanmak için linear model olan "sonuc" değişkenimi içine attım, uyum kestirimine ilişkin güven aralığı elde etmek için interval="confidence" olarak belirttim ve son olarak güven aralığını ` level=0.95` olarak belirttim. Uyum kestirimi değerleri de veri setimin 12. satırı olduğu için 13. satırı belirttim.

Sonuç olarak "*X1=6.473778 , X2=0.50627316, X3=-0.028457406 , X42=0 , X43=0*" değerlerinde uyum kestirimine ilişkin güven aralığını %95 güvenle [12.57582;14.12547] olarak buldum.

"%95 güvenle söylenebilir ki araç satışı ve kiralamadan gelen gelir 6.473778 BinTL arttığında, hisse senetleri 0.50627316 TL değerlendiğinde ve enflasyon %0.028457406 azaldığında LPG'yle çalışan bir aracın aylık olarak bize getirisi nin beklenen değeri ortalama 12.57582 MilyonTL ile 14.12547 MilyonTL arasında olacaktır."

Ön kestirim için:

$$P(\tilde{y} - t_{\frac{\alpha}{2},n-k-1} - S_{\tilde{y}} \leq E(\tilde{y}) \leq \tilde{y} + t_{\frac{\alpha}{2},n-k-1} - S_{\tilde{y}}$$

$$S\tilde{y} = \hat{\sigma} \sqrt{1 + \tilde{x}(X'X)^{-1}\tilde{x}'}$$

```
predict(sonuc,data.frame(x1=11, x2=-2, x3=-2.5),interval="prediction", level=0.95)

##       fit       lwr       upr
##13.350642 9.389428 17.311855
```

Ön kestirimi yapmak üzre "*X1=11 , X2=-2, X3=-2.5 , X42=1 , X43=0*" değerlerini aldım ve sonucunda %95 güvenle elde ettiğim güven aralığı [9.389428;17.311855] şeklinde oldu.

Buradan hareketle yorumumuzu "%95 güvenle söylenebilir ki araç satışı ve kiralamadan gelen gelir 11 BinTL arttığında, hisse senetleri 2TL değer kaybettiğinde ve enflasyon %2.5 azaldığında benzinle çalışan bir aracın aylık olarak bize getirisi nin beklenen değeri ortalama 9.389428 MilyonTL ile 17.311855 MilyonTL arasında olacaktır."

Part 15. En İyi Model

İleriye doğru seçim yöntemi:

Bu yöntem, hiç bir bağımsız değişkenin bulunmadığı regresyon denklemine (y = $\beta 0 + \epsilon$) değişkenlerin her adımda tek tek eklenmesiyle uygulanır.

```
library(stats)
attach(hwData)
null_model <- lm(y ~ 1)
ileri <- step(null_model,y~x1+x2+x3+x4,  direction = "forward")

## Start:  AIC=406.95
## y ~ 1
##
##       Df Sum of Sq    RSS    AIC
## + x4   2    791.80 1518.7 349.27
## + x2   1    691.76 1618.8 356.65
## + x1   1    573.39 1737.1 367.02
## + x3   1    166.81 2143.7 397.94
## <none>             2310.5 406.95
##
## Step:  AIC=349.27
## y ~ x4
##
##       Df Sum of Sq     RSS    AIC
## + x2   1    802.46  716.26 240.79
## + x1   1    716.11  802.61 257.52
## + x3   1    145.37 1373.35 336.48
## <none>             1518.72 349.27
##
## Step:  AIC=240.79
## y ~ x4 + x2
##
##       Df Sum of Sq    RSS    AIC
## + x3   1   104.903 611.36 219.51
## + x1   1    75.211 641.05 226.48
## <none>             716.26 240.79
##
## Step:  AIC=219.51
## y ~ x4 + x2 + x3
##
##       Df Sum of Sq    RSS    AIC
## + x1   1    66.918 544.44 204.47
## <none>             611.36 219.51
##
## Step:  AIC=204.47
## y ~ x4 + x2 + x3 + x1
```

```
ileri

##
## Call:
## lm(formula = y ~ x4 + x2 + x3 + x1)
##
## Coefficients:
## (Intercept)         x42         x43          x2          x3          x1
##       5.431      -3.761      -6.305       2.321      -1.123       1.037
```

Bağımlı değişken y olduğu durumda birinci adımda x_4 modele değişkeni girmiştir. Daha sonra ikinci adımda x_2 değişkeni modele girmiştir. 3. Adımda x_3 değişkeni, 4. adımda ise x_1 değişkeni modele girmiştir.
Demek ki tüm değişkenler modelde anlamlı olduğu için tüm değişkenler modelimizde yer almıştır.

```
summary(ileri)

## Call:
## lm(formula = y ~ x4 + x2 + x3 + x1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.1262 -1.1206  0.2143  1.2867  3.9992
##
## Coefficients:
##             Estimate   Std. Error    t value      Pr(>|t|)
## (Intercept)  5.4312     1.4467   3.754  0.000254 ***
## x42         -3.7608     0.4027  -9.340   < 2e-16 ***
## x43         -6.3051     0.3977 -15.855   < 2e-16 ***
## x2           2.3206     0.2844   8.160   1.67e-13 ***
## x3          -1.1229     0.2245  -5.002   1.66e-06 ***
## x1           1.0369     0.2491   4.163   5.44e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.965 on 141 degrees of freedom
## Multiple R-squared:  0.7644, Adjusted R-squared:  0.756
## F-statistic: 91.48 on 5 and 141 DF,  p-value: < 2.2e-16
```

Yukarıdaki çıktımızdaki p değerlerinden de görüldüğü üzere tüm değişkenlerimizin modele katkısı %99 güven düzeyinde bile anlamlıdır.
Ayrıca model de anlamlıdır (p-value: < 2.2e-16 < $\alpha = 0.05$)

En iyi model: format($y_i = \beta 0_{S_{\beta 0}} + \beta 1 x 1_{S_{\beta 1}}$)

$$y_i = \underset{(1.4467)}{5.4312} - \underset{(0.4027)}{3.7608X42} - \underset{(0.3977)}{6.3051X43} + \underset{(0.2844)}{2.3206X2} - \underset{(0.2245)}{1.1229X3} + \underset{(0.2491)}{1.0369X1} \pm 1.965$$

Geriye doğru seçim yöntemi:

Bu yöntem, tüm değişkenlerin bulunduğu regresyon denkleminden, değişkenlerin her adımda tek tek çıkartılmasıyla uygulanır.

```
geri <- step(sonuc, direction = "backward")

## Start:  AIC=204.47
## hwData$Y ~ hwData$X1 + hwData$X2 + hwData$X3 + hwData$X4
##
##           Df Sum of Sq     RSS    AIC
## <none>                  544.44 204.47
## - hwData$X1  1     66.92  611.36 219.51
## - hwData$X3  1     96.61  641.05 226.48
## - hwData$X2  1    257.14  801.58 259.33
## - hwData$X4  2    984.81 1529.25 352.29
```

Geriye doğru seçim yönteminin özelliğinden tüm değişkenler modelde olarak başlıyor. İlk model ve ayrıca son model tüm bağımsız değişkenlerin modelde bulunduğu durumdur ki o şekilde de kalmıştır.

```
summary(geri)
## Call:
## lm(formula = hwData$Y ~ hwData$X1 + hwData$X2 + hwData$X3 + hwData$X4)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.1262 -1.1206  0.2143  1.2867  3.9992
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   5.4312     1.4467   3.754 0.000254 ***
## hwData$X1     1.0369     0.2491   4.163 5.44e-05 ***
## hwData$X2     2.3206     0.2844   8.160 1.67e-13 ***
## hwData$X3    -1.1229     0.2245  -5.002 1.66e-06 ***
## hwData$X42   -3.7608     0.4027  -9.340  < 2e-16 ***
## hwData$X43   -6.3051     0.3977 -15.855  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.965 on 141 degrees of freedom
## Multiple R-squared:  0.7644, Adjusted R-squared:  0.756
## F-statistic: 91.48 on 5 and 141 DF,  p-value: < 2.2e-16
```

Yukarıdaki çıktımızdan anlaşılan o ki değişkenlerimizin hepsi anlamlı değişkenler.
(p-value değerlerinin hepsi $\alpha = 0.05$ 'ten küçük)

Geriye doğru seçim yöntemi de ileriye doğru seçim yöntemi ile aynı sonucu verdi.
En iyi modelde tüm değişkenlerimiz var ayrıca modelimiz de anlamlı.

Adımsal Seçim Yöntemi::

Adımsal regresyon yöntemi, ileriye doğru seçim ve geriye doğru çıkarma yöntemlerinin aynı anda kullanılmasıyla uygulanır.

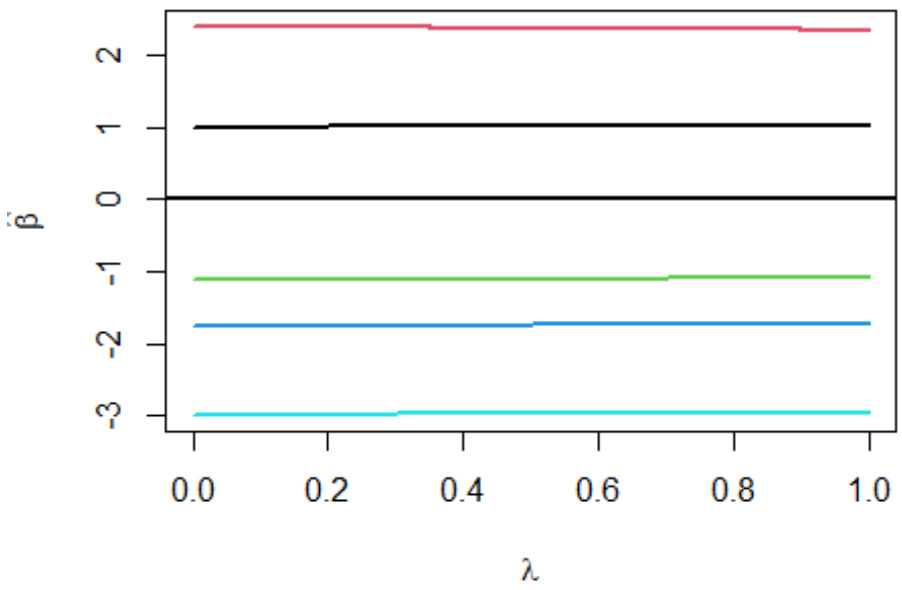
```
library(MASS)
wise_model <- stepAIC(sonuc, direction = "both" ,trace="FALSE")
summary(wise_model)

##
## Call:
## lm(formula = hwData$Y ~ hwData$X1 + hwData$X2 + hwData$X3 + hwData$X4)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.1262 -1.1206  0.2143  1.2867  3.9992
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   5.4312     1.4467   3.754 0.000254 ***
## hwData$X1     1.0369     0.2491   4.163 5.44e-05 ***
## hwData$X2     2.3206     0.2844   8.160 1.67e-13 ***
## hwData$X3    -1.1229     0.2245  -5.002 1.66e-06 ***
## hwData$X42   -3.7608     0.4027  -9.340  < 2e-16 ***
## hwData$X43   -6.3051     0.3977 -15.855  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.965 on 141 degrees of freedom
## Multiple R-squared:  0.7644, Adjusted R-squared:  0.756
## F-statistic: 91.48 on 5 and 141 DF,  p-value: < 2.2e-16

detach(hwData)
```

Adımsal seçim yöntemimiz de diğer iki yöntemimizden farksız olarak tüm değişkenlerimizi anlamlı buldu ve en iyi model seçimimize ekledi.
En iyi modelimiz tüm değişkenlerimizin bulunduğu önceki modelimizle aynıdır. Ayrıca modelimiz de anlamlıdır.

Part 16. Ridge Regresyon

Ridge Regresyon çoklu bağlantı sorunlarının oluşmasında çalınan bir kapıdır. Benim çoklu bağlantı sorunum yok ama hocalarımız yine de bu kapıyı çalmamızı istiyor. Genelde cevabını bildiğim sorular sormam konusunda kızıyorlar ama…
Çalalım bakalım cevabını biliyor olsak da…

```
attach(hwData)
ridge <- lm.ridge(y~x1+x2+x3+x4 , lambda = seq(0,1,0.05))

matplot(ridge$lambda,
      t(ridge$coef),
      type="s",
      lwd=2,
      lty=1,
      xlab=expression(lambda),
      ylab=expression(hat(beta)))
abline(h=0,lwd=2)
```



Hızlı artış ya da azalış gösteren katsayılara karşılık gelen değişkenler çoklu bağlantılı değişkenlerdir.
Ayrıca sıfır eksenin civarında seyreden değişkenler de modelde önemsiz değişkeni göstermektedir.

Ama görüldüğü üzere mis gibi günahsız alkolsüz sim suyuyla yıkanmış gibi bir grafiğimiz var ne "çoklu bağlantı sorunu var ya bunda" dedirtir, ne de "şu değişken önemsiz gibi mi duruyor" gibi şüpheye düşürtür.

Analizimizin önceki adımlarında da yaptığımız üzre söyleyebiliriz ki çoklu bağlantı sorunu yoktur ve hiç bir değişkenimiz 0 ekseni civarında da seyretmiyor ki önemsiz sayılsın.
Bu da sayısal açıklaması:

```
ridge$coef[,ridge$lam == 0.4]
```

```
##       x1       x2       x3      x42      x43 
## 1.012164 2.380995 -1.105855 -1.742963 -2.971244
```