REGRESYON ANALİZİ

ELİF KARTAL

Eskişehir Teknik Üniversitesi
Fen Fakültesi

a) Doğrusal regresyon denklemini kurunuz.

Regresyon Denklemi

y = -117.7 + 0.145 x1 + 0.001 x2 + 0.226 x3 + 0.669 x4 + 0.572 x5

b) Katsayıları yorumlayınız

Regresyon doğrusu ekseni -117.7 noktasında keser.

X1 bir birim arttırıldığı ve diğer xler sabit tutulduğu zaman y'de 0.145 lik artış görülür.

X2 bir birim arttırıldığı ve diğer xler sabit tutulduğu zaman y'de 0.001 lik artış görülür.

X3 bir birim arttırıldığı ve diğer xler sabit tutulduğu zaman y'de 0.226 lık artış görülür.

X4 bir birim arttırıldığı ve diğer xler sabit tutulduğu zaman y'de 0.669 luk artış görülür.

X5 bir birim arttırıldığı ve diğer xler sabit tutulduğu zaman y'de 0.572 lik artış görülür.

Cinsiyetlere Göre Regresyon Denklemi

x6			
0	y	=	-101.0 + 0.056 x1 + 0.013 x2 + 0.231 x3 + 0.637 x4 + 0.574 x5
1	y	=	-98.2 + 0.056 x1 + 0.013 x2 + 0.231 x3 + 0.637 x4 + 0.574 x5

Kadınlar İçin:

Regresyon doğrusu ekseni -101.0 noktasında keser.

X1 bir birim arttırıldığı ve diğer xler sabit tutulduğu zaman y'de 0.056 lık artış görülür.

X2 bir birim arttırıldığı ve diğer xler sabit tutulduğu zaman y'de 0.013 lük artış görülür.

X3 bir birim arttırıldığı ve diğer xler sabit tutulduğu zaman y'de 0.231 lik artış görülür.

X4 bir birim arttırıldığı ve diğer xler sabit tutulduğu zaman y'de 0.637 lik artış görülür.

X5 bir birim arttırıldığı ve diğer xler sabit tutulduğu zaman y'de 0.574 lük artış görülür.

Erkekler İçin:

Regresyon doğrusu ekseni -98.2 noktasında keser.

X1 bir birim arttırıldığı ve diğer xler sabit tutulduğu zaman y'de 0.056 lık artış görülür.

X2 bir birim arttırıldığı ve diğer xler sabit tutulduğu zaman y'de 0.013 lük artış görülür.

X3 bir birim arttırıldığı ve diğer xler sabit tutulduğu zaman y'de 0.231 lik artış görülür.

X4 bir birim arttırıldığı ve diğer xler sabit tutulduğu zaman y'de 0.637 lik artış görülür.

X5 bir birim arttırıldığı ve diğer xler sabit tutulduğu zaman y'de 0.574 lük artış görülür.

c) EKK tahmin edicilerinin standart hatalarını bulunuz.

Standart hatalar, ilgili katsayıların tahminlerinin ne kadar değişkenlik gösterebileceğini ölçer. EKK tahmin edicilerinin standart hatalarını hesaplatırken R programını kullandım ve aşağıdaki sonıçlara ulaştım:

	Std. Error
(Intercept)	28.94249
x1	0.22370
x2	0.20277
x3	0.15296
x4	0.13171
x5	0.15449
x6	3.68661

d) Örneklem regresyon doğrusu için R^2 bulunuz ve yorumlayınız.

Modelin Özetlenmesi

S	R-sq	R-sq(adj)	R-sq(pred)
5.65783	86.76%	84.69%	80.55%

Örneklem regresyon doğrusu için R^2, bağımsız değişkenlerin bağımlı değişken üzerindeki varyansı açıklama oranını temsil eder. Bağımsız değişkenler bağımlı değişkeni %84.69 açıklar

e) Korelasyon matrisini oluşturunuz.

Korelasyonlar

	x1	x2	x3	x4	x5
x2	0.375				
x3	0.749	0.091			
x4	0.684	0.288	0.388		
x5	0.483	0.475	0.147	0.594	
x6	0.837	0.287	0.591	0.718	0.460

X1 ile X2 arasındaki kolerasyon pozitif yönde 0.375 büyüklüğündedir

X1 ile X2 arasındaki kolerasyon pozitif yönde 0.749 büyüklüğündedir

X1 ile X4 arasındaki kolerasyon pozitif yönde 0.684 büyüklüğündedir

X1 ile X6 arasındaki kolerasyon pozitif yönde 0.837 büyüklüğündedir

X2 ile X3 arasındaki kolerasyon pozitif yönde 0.091 büyüklüğündedir

X2 ile X4 arasındaki kolerasyon pozitif yönde 0.288 büyüklüğündedir

X2 ile X5 arasındaki kolerasyon pozitif yönde 0.475 büyüklüğündedir

X2 ile X6 arasındaki kolerasyon pozitif yönde 0.287 büyüklüğündedir

X3 ile X4 arasındaki kolerasyon pozitif yönde 0.388 büyüklüğündedir

X3 ile X5 arasındaki kolerasyon pozitif yönde 0.147 büyüklüğündedir

X3 ile X6 arasındaki kolerasyon pozitif yönde 0.591 büyüklüğündedir

X4 ile X5 arasındaki kolerasyon pozitif yönde 0.594 büyüklüğündedir

X4 ie X6 arasındaki kolerasyon pozitif yönde 0.718 büyüklüğündedir

X5 ile X6 arasındaki kolerasyon pozitif yönde 0.460 büyüklüğündedir

f) Parametreler için güven aralıklarını oluşturunuz. (α=0.05)

	2.5 %	97.5 %
(Intercept)	-160.02628263	-41.9691040
x1	-0.40066673	0.5118180
x2	-0.40102341	0.4260810
x3	-0.08142353	0.5425006
x4	0.36801577	0.9052444
x5	0.25915732	0.8893136
x6	-4.75773035	10.2800323

Parametreler %95 olasılıkla yukarıda belirtilen aralıklarda bulunur.

g)Yorumlayınızin sınaması (F-testi), katsayılar için anlamlılık sınaması (t-testi) yaparak, sonuçları yorumlayınız. (α=0.05)

ANOVA

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	5	6713.36	1342.67	41.94	0.000
x1	1	18.96	18.96	0.59	0.447
x2	1	0.00	0.00	0.00	0.995
x3	1	71.22	71.22	2.22	0.146
x4	1	942.66	942.66	29.45	0.000
x5	1	445.25	445.25	13.91	0.001
Error	32	1024.35	32.01		
Total	37	7737.71			

H0: Model anlamlı değildir

H1: Model anlamlıdır.

F0>F H0 reddedilir model anlamlıdır.

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
x1	0.145	0.188	0.77	0.447	4.74
x2	0.001	0.201	0.01	0.995	1.43
x3	0.226	0.152	1.49	0.146	2.81
x4	0.669	0.123	5.43	0.000	2.35
x5	0.572	0.153	3.73	0.001	1.91

H0: katsayı anlamlı değildir.

H1: katsayı anlamlıdır.

X1: X1'in t-değeri 0.77 ve p-değeri 0.447, bu da X1'in istatistiksel olarak anlamlı bir etkisi olmadığını gösterir. H0 kabul edilir.

X2: X2'nin t-değeri 0.01 ve p-değeri 0.995, bu da X2'nin istatistiksel olarak anlamlı bir etkisi olmadığını gösterir. H0 kabul edilir.

X3: X3'ün t-değeri 149 ve p-değeri 0.146, bu da X3'ün istatistiksel olarak anlamlı bir etkisi olmadığını gösterir. H0 kabul edilir.

X4: X4'ün t-değeri 5.43 ve p-değeri 0.00, bu da X4'ün istatistiksel olarak anlamlı bir etkisi olduğunu gösterir. H reddedilir.

X5: X5'in t-değeri 3.73 ve p-değeri 0.001, bu da X5'in istatistiksel olarak anlamlı bir etkisi olduğunu gösterir.

h) Bir x_0 vektörüne karşı ortalama kestirim için güven aralığı hesaplayınız (α=0.05).

```
data:  reg$x1
t = 103.11, df = 37, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 176.6434 183.7250
sample estimates:
mean of x
 180.1842
```

ı) Açıklayıcı değişkenler arasında çoklu bağlantı (multi-collinearity) olup olmadığını tespit ediniz.

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
x1	0.145	0.188	0.77	0.447	4.74
x2	0.001	0.201	0.01	0.995	1.43
x3	0.226	0.152	1.49	0.146	2.81
x4	0.669	0.123	5.43	0.000	2.35
x5	0.572	0.153	3.73	0.001	1.91

VIF değerleri, her bir bağımsız değişkenin diğer bağımsız değişkenlerle olan ilişkisini ölçen bir ölçüdür. Genel olarak, VIF değeri 10'dan büyükse, çoklu bağlantı sorunu olduğu düşünülür. İncelediğiniz VIF değerleri aşağıdaki gibidir:

X1: 4.74

X2: 1.43

X3: 2.81

X4: 2.35

X5: 1.91

Bu değerlere göre, çoklu bağlantı sorunu belirgin değildir. Ancak, genel olarak kabul edilen bir sınıra sahip olmayan bir ölçü olduğu için değerlendirmenin bağlam ve analiz bağlamında yapılması önemlidir.

i) Hata terimleri arasında otokorelayon olup olmadığını Durbin-Watson ile test ediniz.

DW = 2: Otokorelasyon yok (hata terimleri arasında bağlantı yok).

DW < 2: Pozitif otokorelasyon (hata terimleri arasında pozitif bir ilişki var).

DW > 2: Negatif otokorelasyon (hata terimleri arasında negatif bir ilişki var).

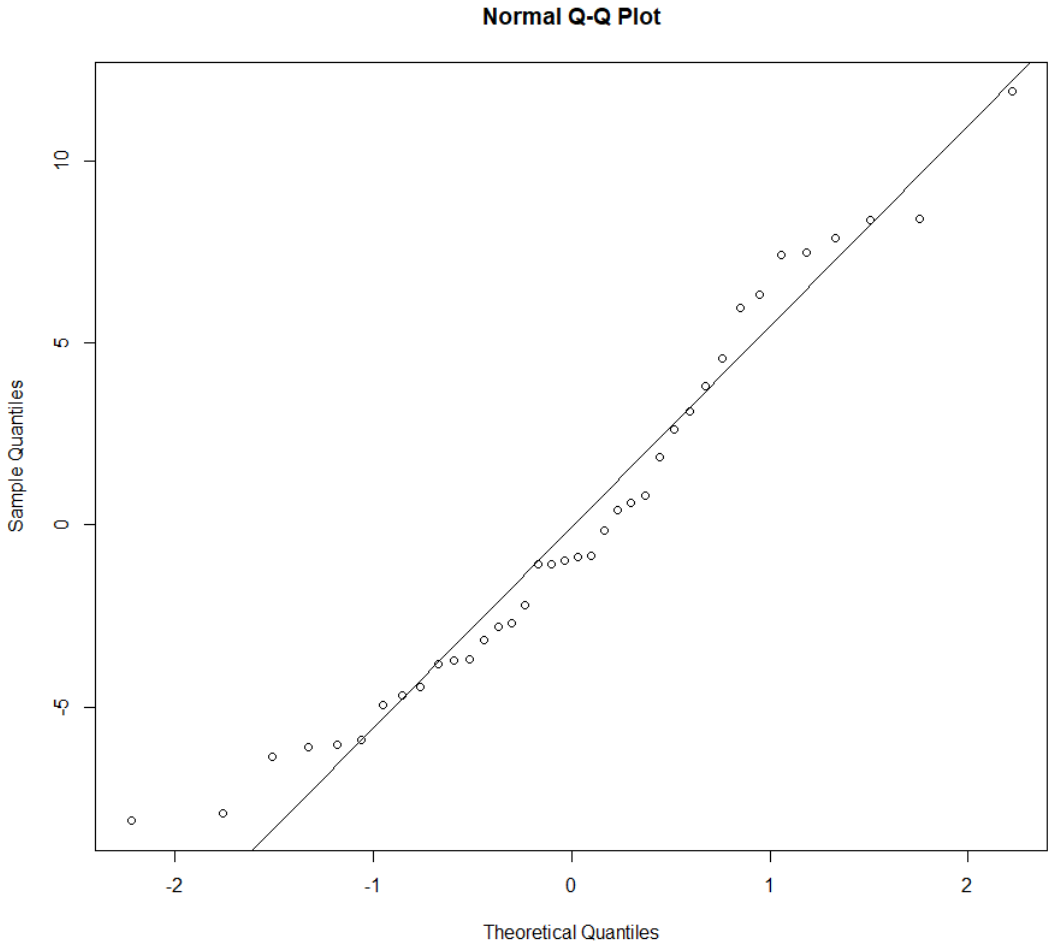
```
lag Autocorrelation D-W Statistic p-value
  1       0.1885779      1.481744   0.068
 Alternative hypothesis: rho != 0
```

Regresyon modeli üzerinde Durbin-Watson testi uygulandı ve elde edilen sonuçlar gözlemlendi. Durbin-Watson istatistiği yaklaşık olarak 2'ye yaklaşıyorsa, bu durumda otokorelasyon olmadığı yorumu yapılabilir. Ancak, Durbin-Watson istatistiği belirgin bir şekilde 2'den küçük veya 2'den büyükse, otokorelasyon olasılığını değerlendirmek önemlidir. Yapılan test sonucunda Durbin-Watson istatistiğinin 2'den küçük olduğu görüldü.

Ayrıca, p-değerine göz atıldığında, bu değerin 0.05'ten küçük olduğu belirlendi. Bu durumda, alternatif hipotezi kabul ediyoruz ve bu, gerçek otokorelasyonun sıfırdan büyük olduğu anlamına gelir. Bu bulgu, hata terimleri arasında pozitif bir otokorelasyon olabileceği ihtimalini ortaya koyar.

Bu sonuçlar, otokorelasyonun istatistiksel olarak anlamlı olduğunu göstermektedir. Bu durum, regresyon modelinin varsayımlarından biri olan hata terimleri arasındaki otokorelasyonun dikkate alınması gerektiğini işaret etmektedir.

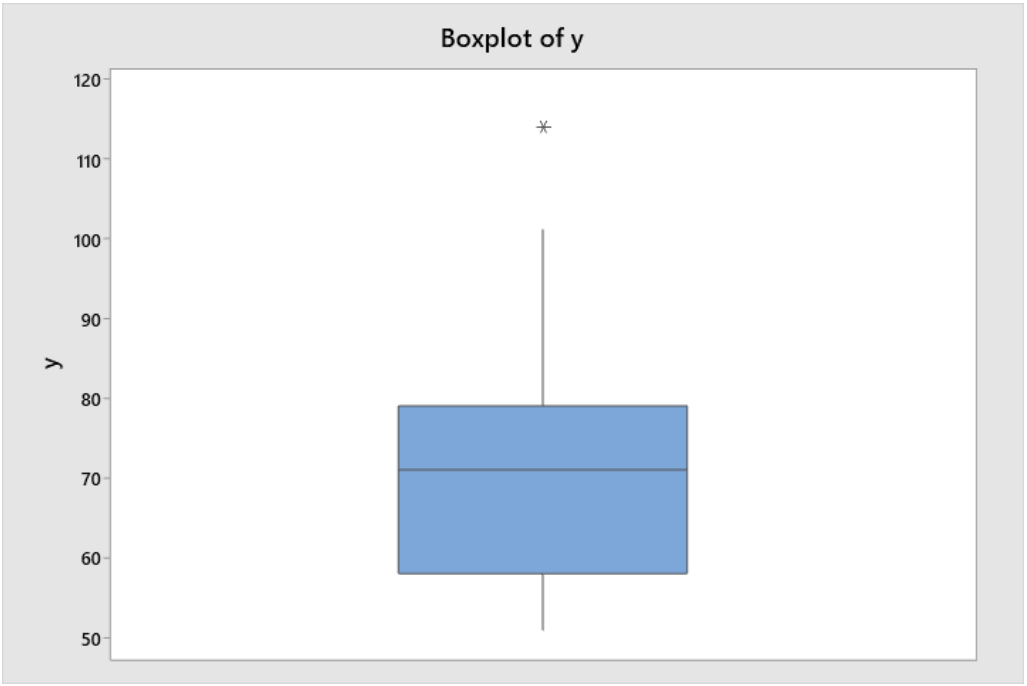
j) Regresyon artıkları normal dağılıma sahip midir? (q-q plot grafiği ve bir tane normalik testi ile araştırınız)

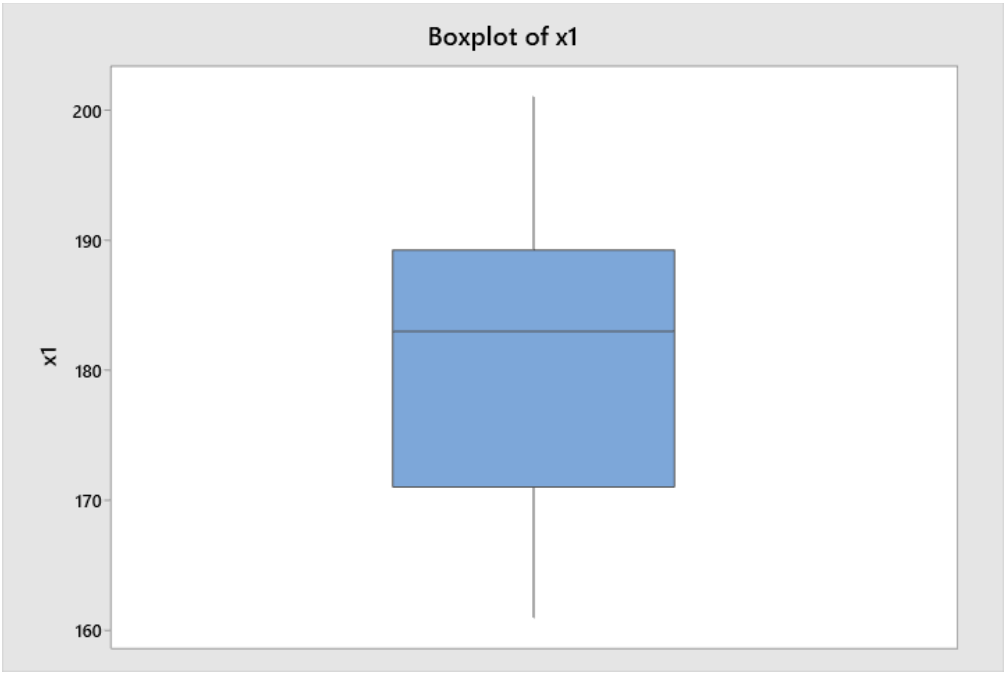


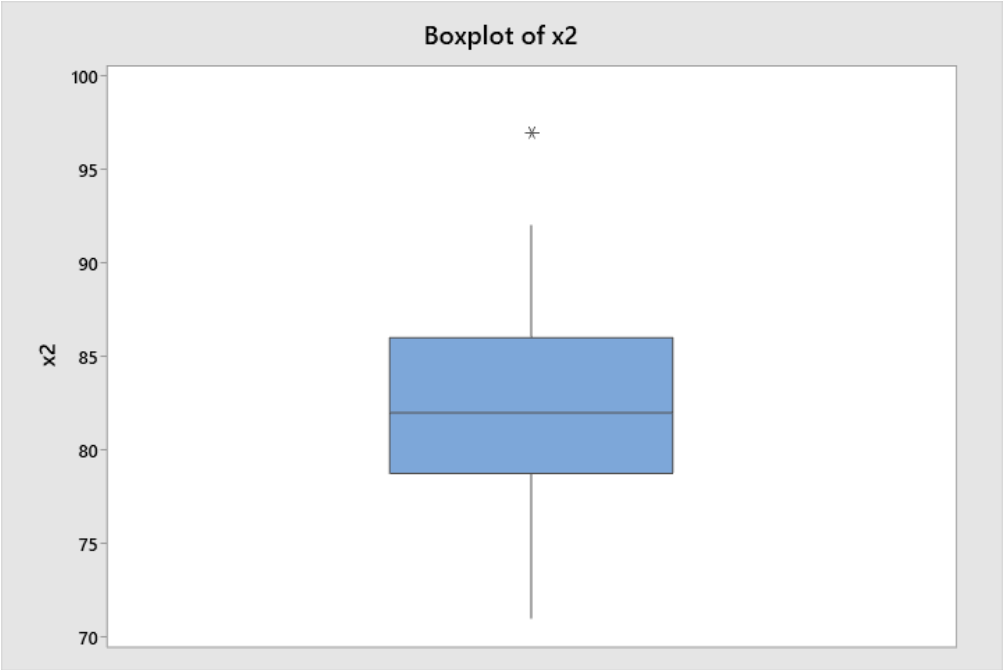

```
        Shapiro-Wilk normality test

data:  residuals
W = 0.95345, p-value = 0.1157
```

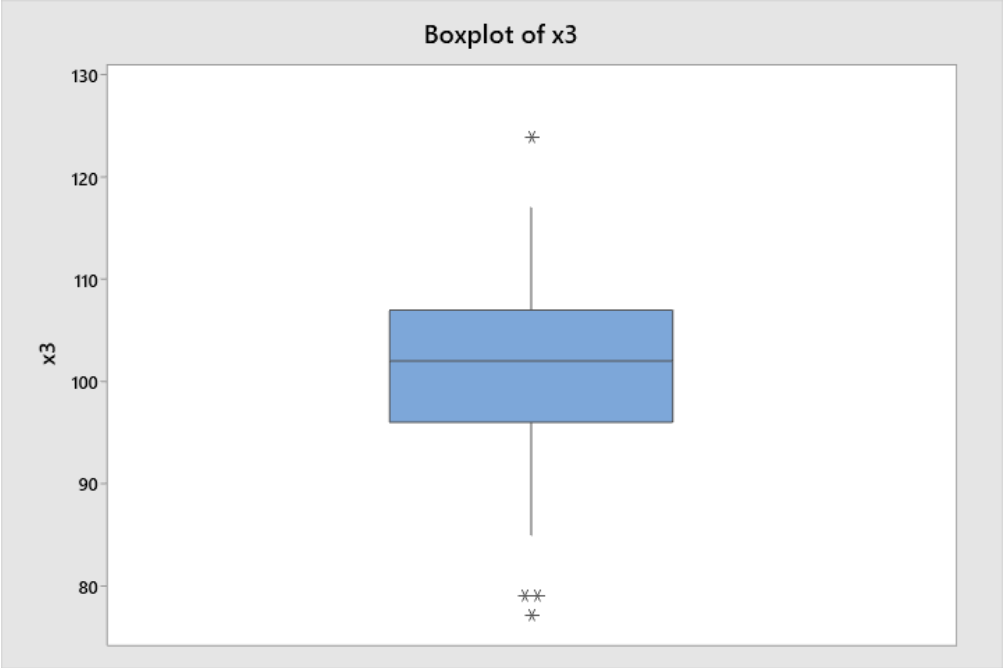
Artıklar normal dağılıma sahiptir.

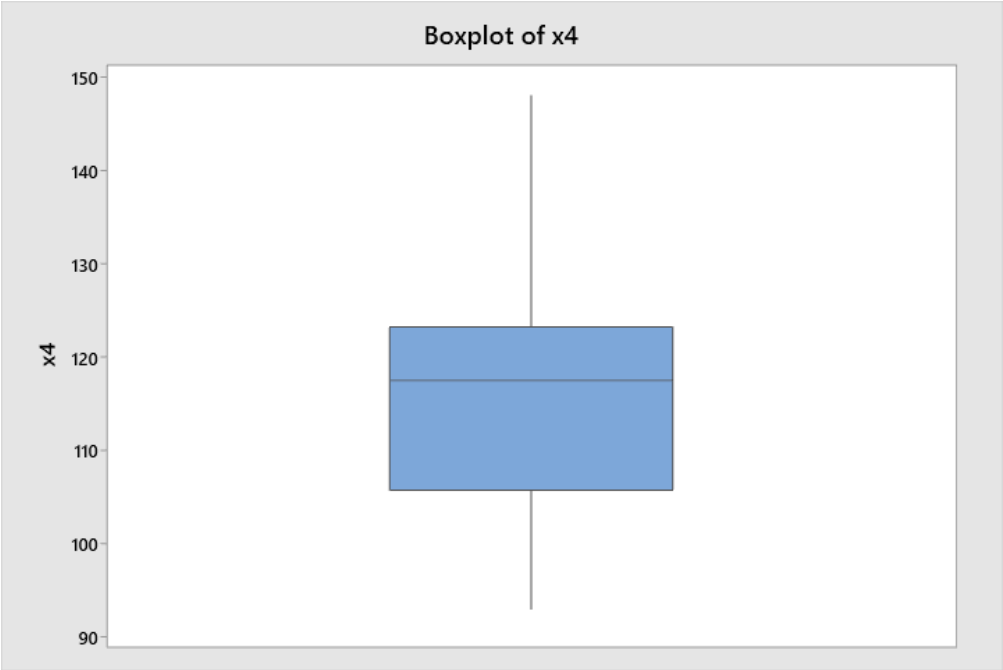
k) Verilerde aykırı değer var mıdır?

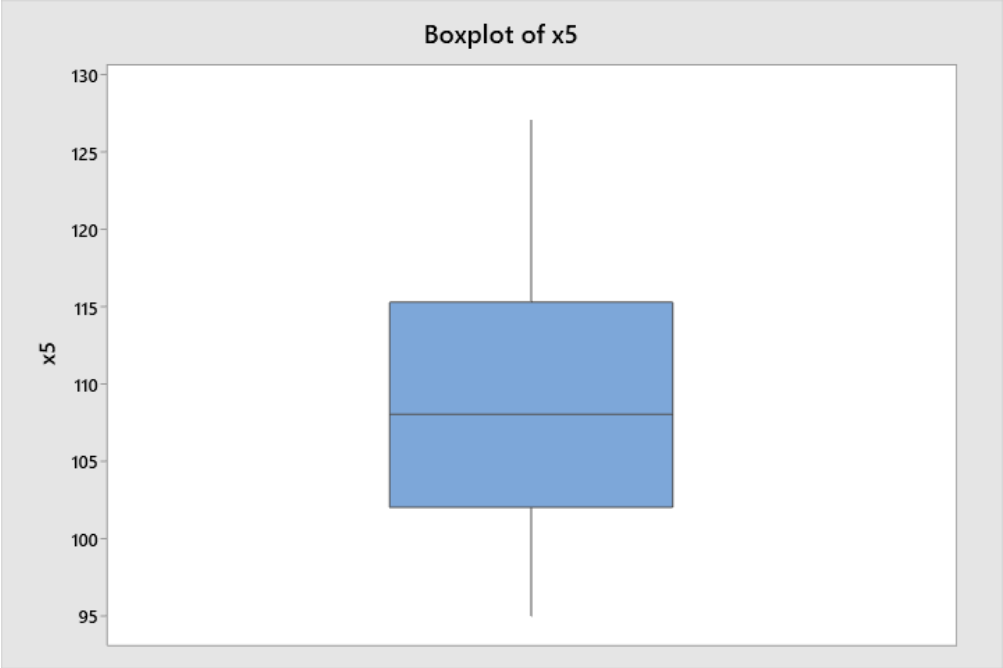





Boxplot of x2
x2
100
95
90
85
80
75
70


Boxplot of x3
x3
130
120
110
100
90
80





Yukarıdaki grafiklere baktığımızda Y, X3 ve X2 verileri dışındaki verilerde aykırı değer olmadığını görebiliriz.

m) Modelde sabit terim olmasaydı (β_0), modelin veriye uyumu nasıl değişirdi? 7

Sabit terimin (β_0) bulunmaması durumu, regresyon modelinin verilere uydurulmasını ve tahminlerini etkiler. Sabit terim, bağımsız değişkenlerin değerleri sıfır olduğunda bağımlı değişkenin beklenen değerini temsil eder. Sabit terimin olmaması durumunda modelin formülü şu şekilde değişir:

$$Y = B_1 \cdot X_1 + B_2 \cdot X_2 + \ldots + B_n \cdot X_n + E$$

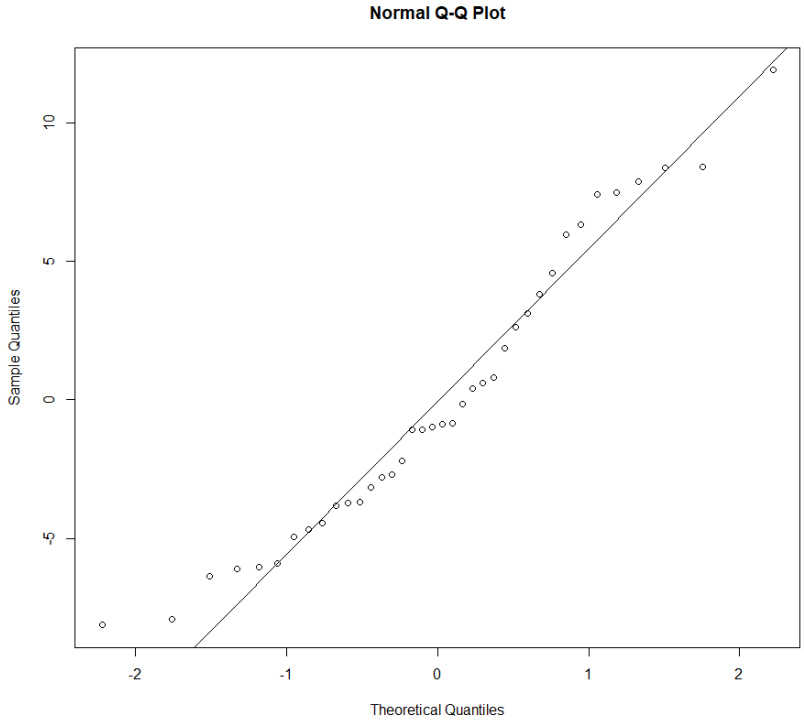
Burada:

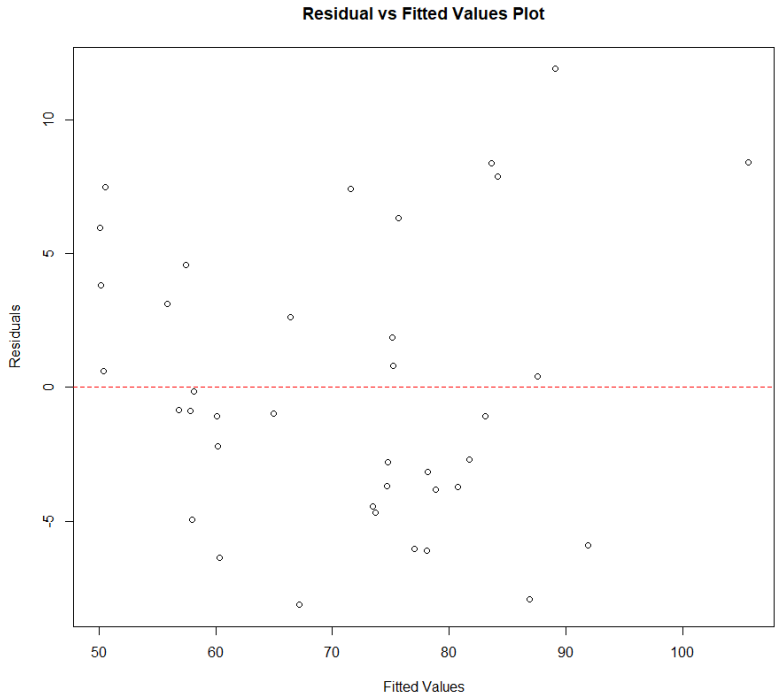
- Y bağımlı değişken,
- $X_1, X_2, \ldots, X_n$ bağımsız değişkenler,
- $B_1, B_2, \ldots, B_n$ katsayılar,
- E hata terimini temsil eder.

Sabit terim olmaması durumunda, modelin başlangıç noktası sıfırdır. Yani, tüm bağımsız değişkenler sıfır olduğunda beklenen bağımlı değişken değeri sıfırdır. Sabit terimin olmaması durumunda model, orijinden geçmek zorunda olduğu için daha sınırlı bir esnekliğe sahip olabilir. Modelin başlangıç noktası belirgin bir etkisi olmadan sıfırdır. Sabit terim eklemek, modelin daha genel bir başlangıç noktasına sahip olmasını ve veriye daha iyi uymasını sağlar.

Sabit terimin olmaması durumunda, modelin başlangıç noktasının sıfır olması, modelin yorumlanmasını ve genelleştirilmesini zorlaştırabilir. Ayrıca, sabit terim olmaması durumunda modelin, özellikle orijinden geçmesi gerektiği için bazı durumlarda uyumsuzluğa neden olabilir.

n) Sabit varyans veya normal olma varsayımlarını artık grafikleri yardımıyla inceleyiniz.





Veri normal dağılır ve değişen varyans sorunu yoktur.

o) Kadın ve Erkek öğrenciler arasında ağırlık bakımından anlamlı bir farklılık var mıdır?

Method

μ_1: mean of y when x6 = 0

μ_2: mean of y when x6 = 1

Descriptive Statistics: y

x6	N	Mean	StDev	SE Mean
0	16	58.69	5.49	1.4
1	22	80.3	11.9	2.5

Estimation for Difference

Difference	95% CI for Difference
-21.59	(-27.46, -15.71)

Test

Null hypothesis	$H_0: \mu_1 - \mu_2 = 0$
Alternative hypothesis	$H_1: \mu_1 - \mu_2 \neq 0$

T-Value	DF	P-Value
-7.49	31	0.000

Yukarıdaki sonuçlara baktığımızda H0 reddedilir. İki örneklem arasında farklılık olduğunu kabul ederiz.

ö) Kurduğunuz modelin performansı (R^2, F veya başla bir kriter açısından) arttırılabilir mi? Nasıl arttırabilirsiniz işlem ve yorumla anlatınız.

Regresyon modelinin performansını artırmak için birkaç strateji bulunmaktadır. Performans artırma çabaları genellikle modelin doğruluğunu, genelleme yeteneğini ve açıklama gücünü geliştirmeye odaklanır. Regresyon modelinin performansını artırmak için bazı yaygın stratejiler şunlardır:

Daha Fazla Veri Toplama:

Daha fazla veri toplamak, modelin genelleme yeteneğini artırabilir ve daha güvenilir tahminler yapmasına olanak tanır.

Değişken Seçimi:

Modeldeki değişkenleri dikkatlice seçmek ve yeni değişkenler türetmek, modelin açıklama gücünü ve tahmin yeteneğini artırabilir.

Eksik Veri İle Başa Çıkma:

Eksik verilerle başa çıkma stratejileri kullanarak, eksik veri sorunlarını ele almak ve daha güvenilir sonuçlar elde etmek mümkündür.

Outlier'ları İncelenme ve İşleme Alma:

Modelin performansını etkileyebilecek aykırı değerleri inceleyip işlemek, modelin daha dengeli ve güvenilir olmasını sağlar.

Model Kompleksliğini Artırma veya Azaltma:

Modelin karmaşıklığını artırmak (polinom eklemek gibi) veya azaltmak (değişkenleri çıkarmak veya düşük dereceli terimleri kullanmak gibi), modelin uyum ve genelleme yeteneğini etkileyebilir.

Duyarlılık Analizi:

Değişkenlerin model üzerindeki etkilerini inceleyerek, önemli değişkenlere odaklanmak ve modelin duyarlılık analizini yapmak, modelin anlaşılabilirliğini artırabilir.

Regresyon Modelleri Karşılaştırma:

Farklı regresyon modellerini karşılaştırarak, en iyi performansı sağlayan modeli seçmek ve kullanmak.

Cross-Validation Kullanma:
Veriyi eğitim ve test setlerine ayırmak yerine, cross-validation yöntemleri kullanarak modelin daha güvenilir bir performans ölçümü elde etmek.

Bu stratejiler, modelinizi geliştirmek ve daha iyi sonuçlar elde etmek için kullanılabilecek genel yaklaşımlardır. Ancak, her durum benzersiz olduğu için, spesifik duruma bağlı olarak değişiklikler yapılması gerekebilir. Model performansını artırmak için yapılan değişikliklerin etkisini değerlendirmek için kapsamlı bir model değerlendirmesi yapmak önemlidir.