پروژه ی مدل

تحلیل رگرسیون چندگانه

بررسی عوامل موثر بر مرگ و میر در شهرهای کوچک آمریکا

منیره معنوی- زهرا سیفی

بهار ۹۷

تعریف متغیر ها

Y = میزان مرگ و میر در هر ۱۰۰۰ نفر
X1 = در دسترس بودن پزشک در هر ۱۰۰،۰۰۰ ساکنان
X2 = دسترسی به بیمارستان در هر ۱۰۰،۰۰۰ ساکنان
X3 = درآمد سالیانه سرانه در هزار دلار است
X4 = تراکم جمعیت در هر مایل مربع

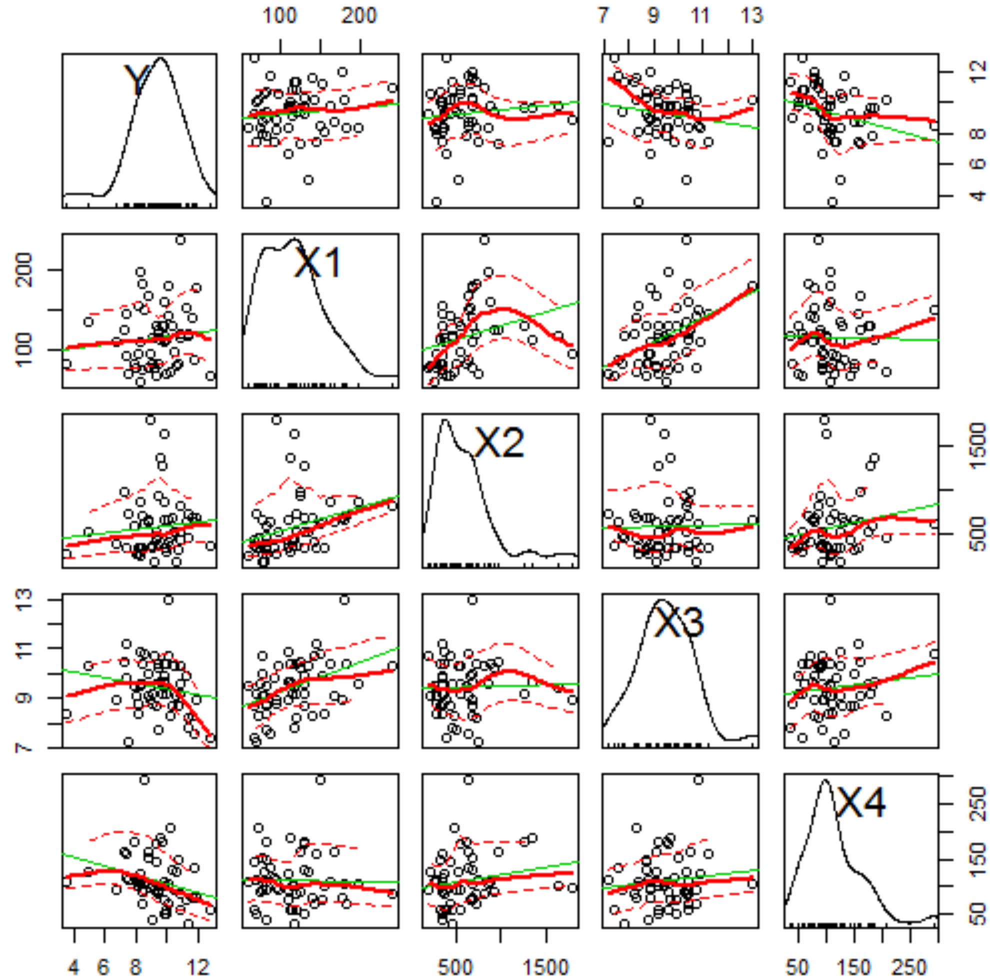
تعداد مشاهدات ۵۳ می باشد.

فراخوانی داده

```
> rm(list=ls())
> A<-file.choose()
> data<-read.table(A,header=T)
> hea(data)
     Y  X1   X2   X3  X4
1   8.0  78  284  9.1 109
2   9.3  68  433  8.7 144
3   7.5  70  739  7.2 113
4   8.9  96 1792  8.9  97
5  10.2  74  477  8.3 206
6   8.3 111  362 10.9 124
```

نمودار پراکنش

```
> library(car)
> scatterplotMatrix(cbind(Y,X1,X2,X3,X4))
```



با توجه به دو نمودار بالا در می یابیم که بین متغیر Y و متغیر های مستقل ارتباط خطی وجود ندارد . و ممکن است بین متغیر X1 و X3 هم خطی وجود داشته باشد که باید بررسی گردد . و به نظر تمامی متغیرها چولگی در توزیع خود دارند. که در ادامه به بررسی آن می پردازیم .

مدل رگرسیونی خطی چندگانه

```
> M<-lm(Y~X1+X2+X3+X4)

> M

Call:

lm(formula = Y ~ X1 + X2 + X3 + X4)
```

Coefficients:

(Intercept)	X1	X2	X3
12.2662557	0.0073916	0.0005837	-0.3302303

X4
-0.0094629

خلاصه ی مدل

Call:

lm(formula = Y ~ X1 + X2 + X3 + X4)

Residuals:

↓ مینیمم مانده ها   ↓میانه باقی مانده ها   ماکزیمم باقی ماند ها↓

Min	1Q	Median	3Q	Max
-5.6404	-0.7904	0.3053	0.9164	2.7906

↑چارک اول مانده ها   ↑چارک سوم باقی مانده ها

Coefficients:

	↓ $\widehat{\beta_\iota}$	↓ $\sqrt{var(\widehat{\beta_\iota})}$	↓آماره	↓ $p - value$
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	12.2662557	2.0201466	6.072	1.95e-07 ***
X1	0.0073916	0.0069336	1.066	0.2917
X2	0.0005837	0.0007219	0.809	0.4228
X3	-0.3302303	0.2345517	-1.408	0.1656
X4	-0.0094629	0.0048868	-1.936	0.0587 .

Signif. codes:

0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.601 on 48 degrees of freedom

Multiple R-squared: 0.1437,   Adjusted R-squared: 0.07235

$\uparrow R^2$ $\uparrow R^2_{adj}$

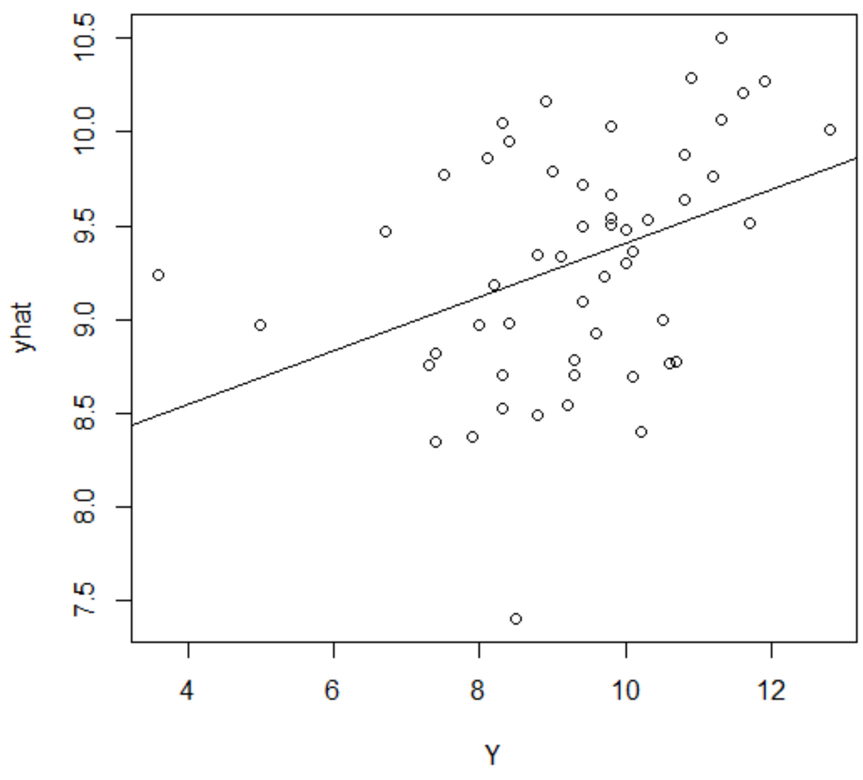
F-statistic: 2.014 on 4 and 48 DF, p-value: 0.1075

پی مقدار مربوط به متغیرها به جز متغیر x4 بسیار زیاد می باشد . مقدار R^2 و R^2_{adj} بسیار کم می باشد .
بنابراین این مدل ، مدل مناسبی نیست.

نمودار Y درمقابل yhat

یکی از راه های شهودی بررسی نیکویی برازش مدل M رسم این نمودار می باشد .

```
yhat<-fitted(M)
> plot(Y,yhat)
> abline(lsfit(Y,yhat))
```



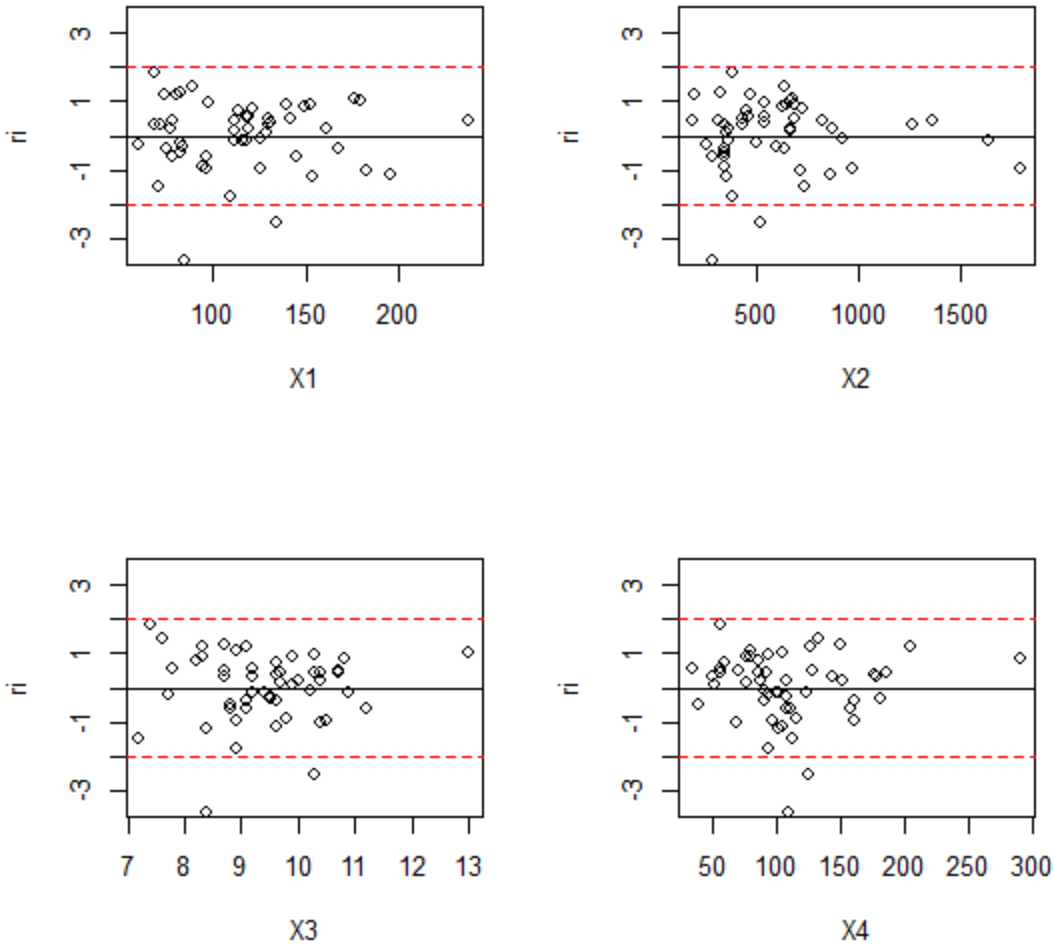
با توجه به شکل به وضوح می توان فهمید که مدل M مدل خوبی نمی باشد چرا که نقاط روی خط برازش داده شده قرار ندارند .

بررسی مناسبت مدل

نمودار باقی مانده های استاندارد شده

مانده ها برخلاف خطاها هم واریانس نیستند ولی مانده های استاندارد شده هم واریانس می باشند.

```
> ei<-resid(M)

> ri<-rstandard(M)

par(mfrow=c(2,2))

plot(X1,ri)

abline(0,0)

abline(h=c(-2,2),col=2,lty=2)

plot(X2,ri)

abline(0,0)

abline(h=c(-2,2),col=2,lty=2)

plot(X3,ri)

abline(0,0)

abline(h=c(-2,2),col=2,lty=2)

plot(X4,ri)

abline(0,0)

abline(h=c(-2,2),col=2,lty=2)
```



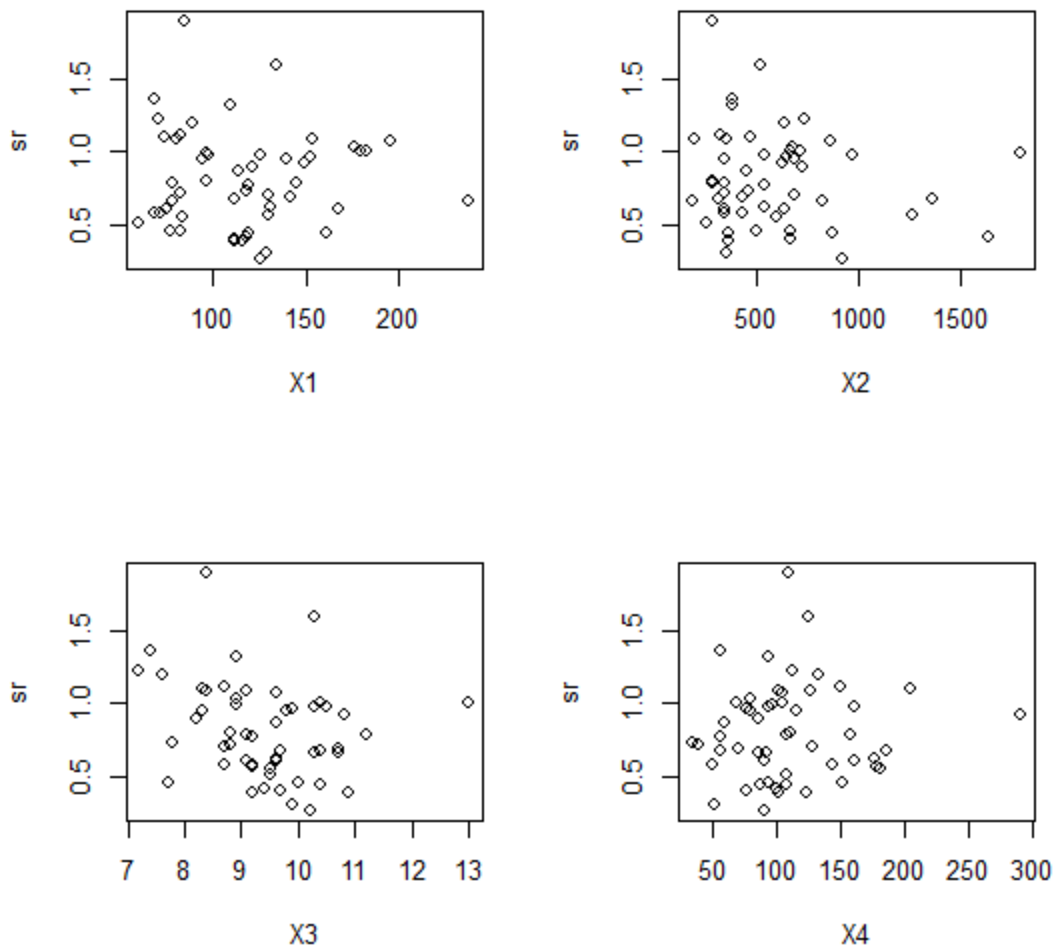
روند غیر تصادفی در نمودار های بالا مشاهده نمی شود. البته در هرکدام از نمودارهای بالا دو نقطه وجود دارند ، که $|r_i|>2$ است . که ممکن است به دلیل مناسب نبودن مدل و سیگنال هایی نشان از بدی مدل برازش شده باشند **شاید** با بهبود مدل بهبود یابند .

نمودار جذر قدر مطلق باقی مانده های استاندارد شده

به دلیل این که مانده های استاندارد شده دارای علامت + و − می باشند و ممکن است یکدیگر را خنثی کنند و تاثیر آن ها به طور کامل بررسی نشود، به منظور بررسی دقیق تر از قدر مطلق آن ها استفاده می کنیم . و توان 1/2 به دلیل کاهش مقدار مطلق چولگی داده هاست .

```
> sr<-sqrt(abs(ri))
par(mfrow=c(2,2))
plot(X1,sr)
```

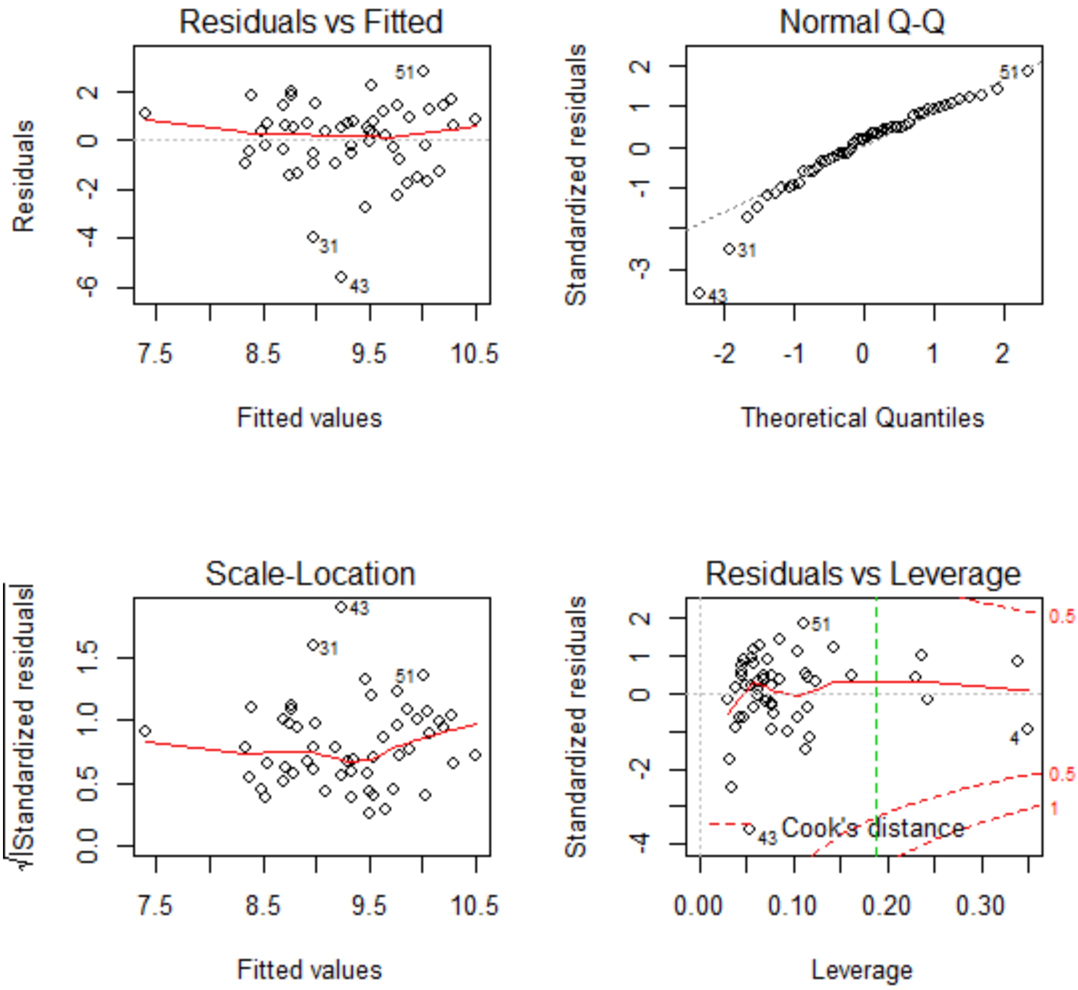
```
plot(X2,sr)

plot(X3,sr)

plot(X4,sr)
```



روند غیر تصادفی در نمودار های بالا مشاهده نمی شود. و به نظر ثبات واریانس وجود دارد.

نمودارهای مربوط به عیب شناسی رگرسیون

```
par(mfrow=c(2,2))

plot(M)

abline(v=2*(5/53),lty=2,col=3)
```



نمودار اول (بالا- سمت چپ) مانده ها را در برابر مقادیر برازش شده رسم نموده است .که نیکویی برازش را تعیین می کند که در این جا نقاط ۳۱ و۴۳ و۵۱ به نظر طبیعی می آیند و باید بررسی شوند .

نمودار دوم (بالا - سمت راست) مانده های استاندارد شده را در برابر چندک های تجربی رسم نموده است . که به نظر در انتهای نمودار مقداری چولگی مشاهده می شود . و نقاط ۴۳ و ۵۱ غیر طبیعی به نظر می رسند .

نمودار سوم (پایین - سمت چپ) جذر قدر مطلق مانده های استاندارد شده را در برابر مقادیر برازش شده نشان می دهد که در ابتدا مقدار بسیار کمی نزول و در انتها صعود داشته است ولی به نظر این صعود و نزول شیب بسیار کمی را داراست و ثبات واریانس وجود دارد .

نمودار چهارم (پایین – سمت راست) مانده های استاندارد شده را در مقابل leverage نشان می دهد. نقاطی که بعد از خط عمودی در شکل قرار گرفتند نقاط اهرمی می باشند . و طبق نمودار پنج نقطه ی اهرمی در شکل وجود دارد.

بررسی میزان همخطی

براساس عامل تورم واریانس

```
> vif(M)
      X1       X2       X3       X4
1.399501 1.169338 1.290415 1.078055
```

نتیجه گیری کلی : مدل رگرسیونی خطی ساده مناسب نمی باشد با استفاده ازتبدیل باکس- کاکس ادامه می دهیم و بعد از تبدیل مجددا مدل بندی می کنیم .

تبدیل باکس – کاکس توزیع متغیر ها را به سمت نرمال می برد .

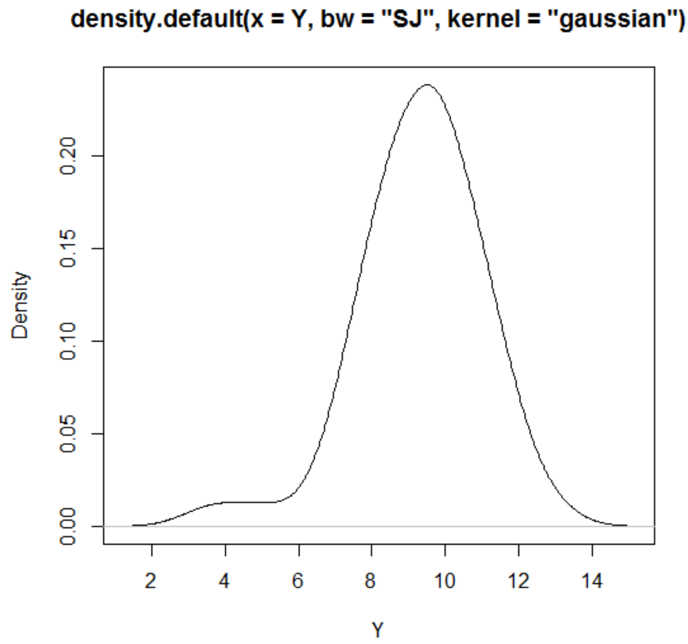
تبدیلات

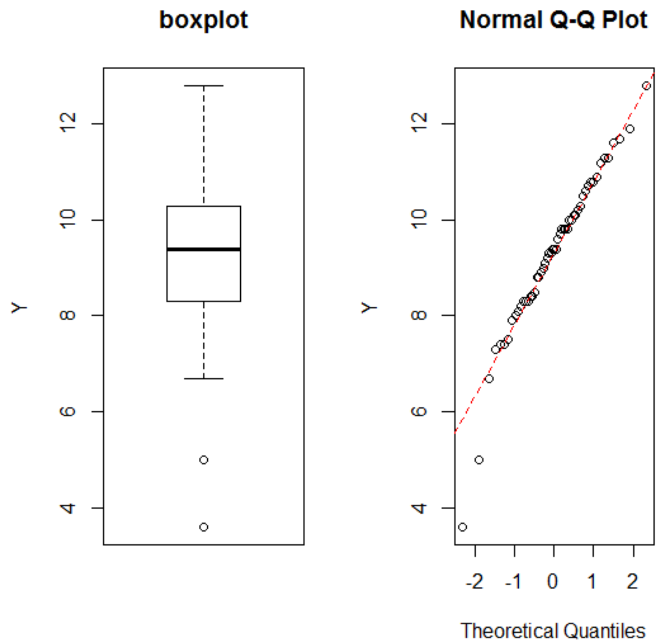
با توجه بررسی هایی که انجام شد **عدم** مناسبت مدل برازش شده تایید شد ، و به دلیل **عدم** وجود رابطه ی خطی بین متغیر وابسته و متغیرهای مستقل نیاز به انجام تبدیل داریم .

ابتدا باید بررسی کنیم که آیا همه ی متغیر ها به تبدیل نیاز دارند یا نه ؟

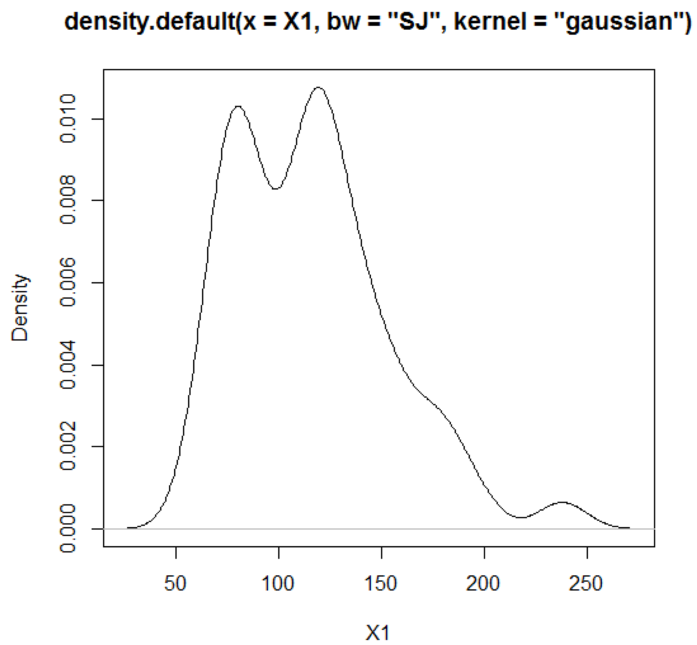
بدین منظور برآورد تابع چگالی براساس روش رگرسیون ناپارامتری کرنل و نمودار جعبه ای و نمودار qqnorm را برای تمامی متغیر ها رسم می کنیم .

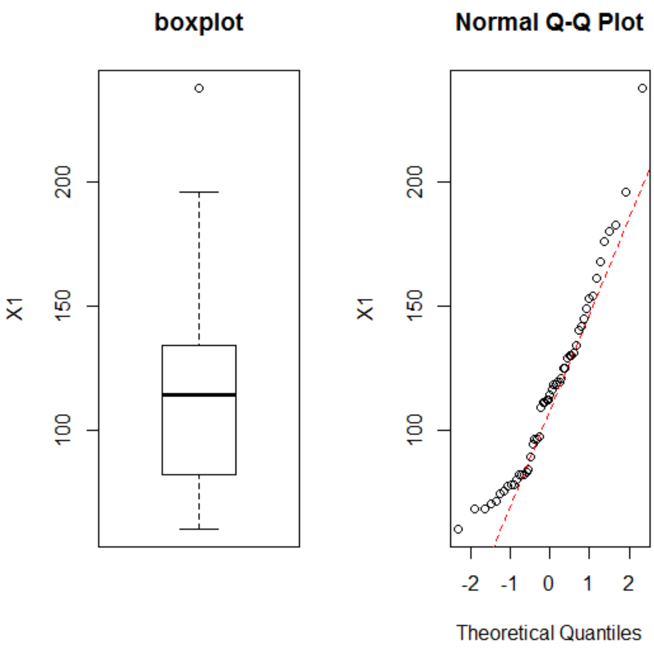
Y:


density.default(x = Y, bw = "SJ", kernel = "gaussian")
Density
0.00
0.05
0.10
0.15
0.20
2
4
6
8
10
12
14
Y


boxplot
Y
4
6
8
10
12
Normal Q-Q Plot
Y
4
6
8
10
12
-2
-1
0
1
2
Theoretical Quantiles

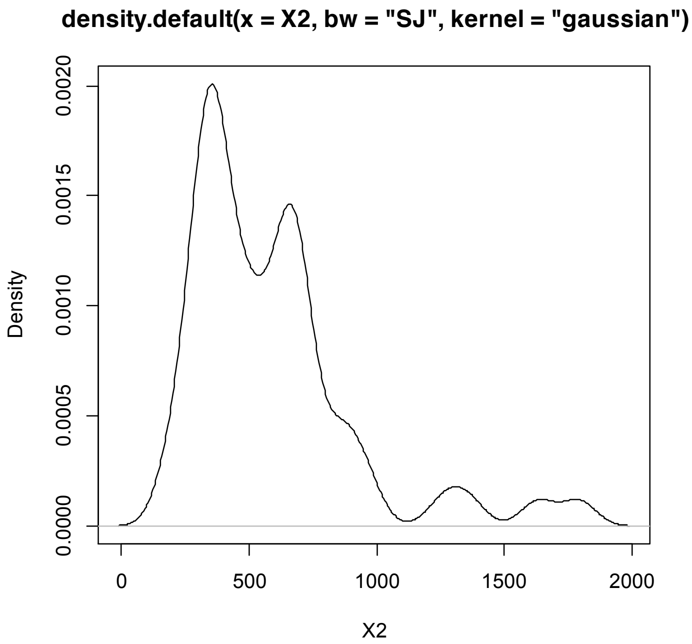
X1:

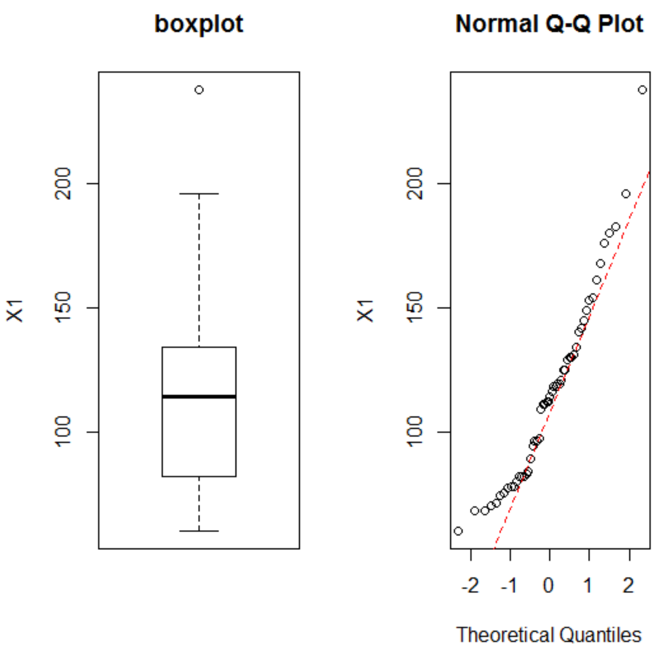




با توجه به نمودار های بالا به وضوح مشخص است که متغیر X1 نرمال نیست .

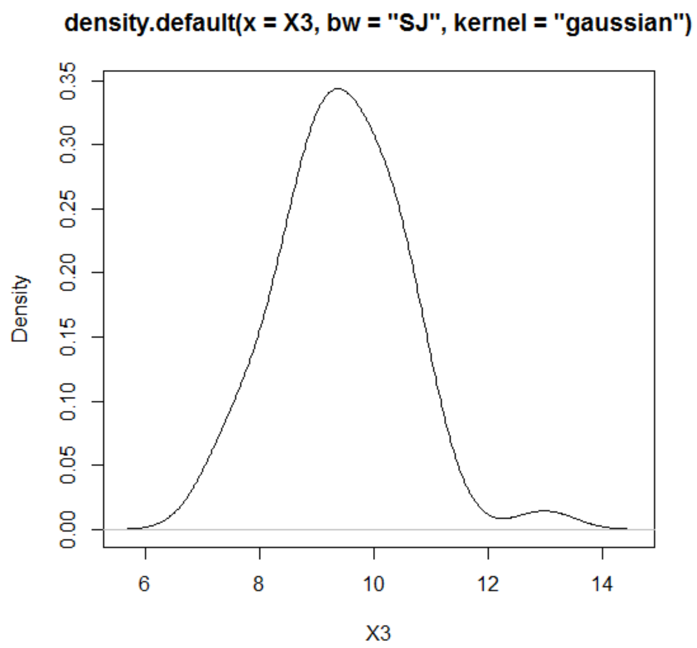
X2:

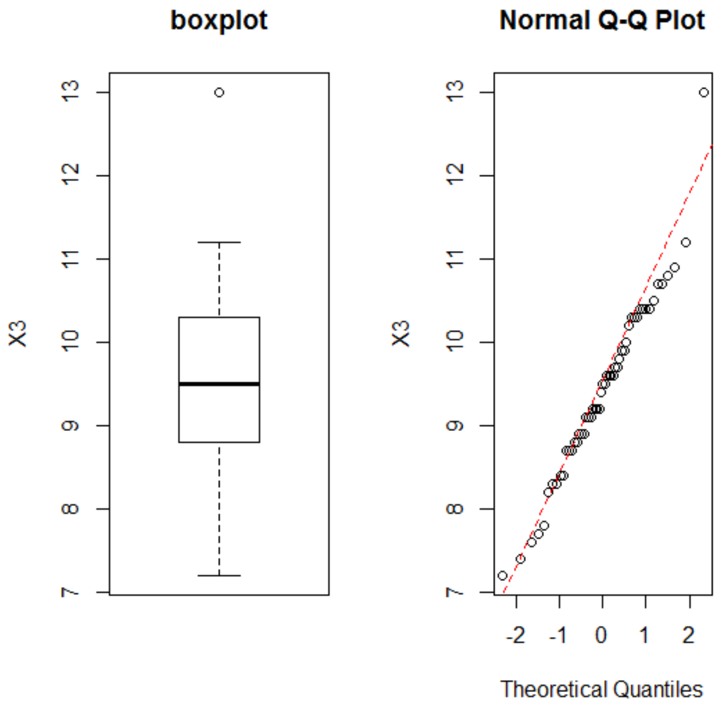




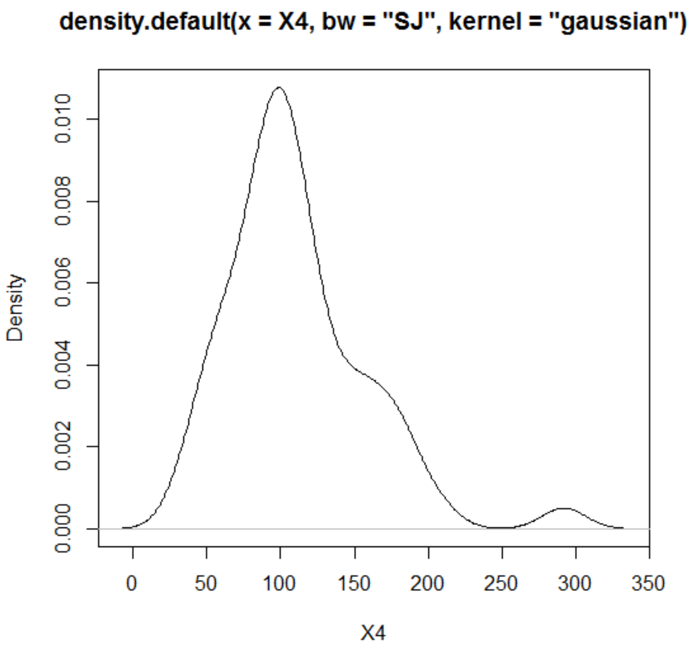
با توجه به نمودار های بالا به وضوح مشخص است که متغیر X1 نرمال نیست .

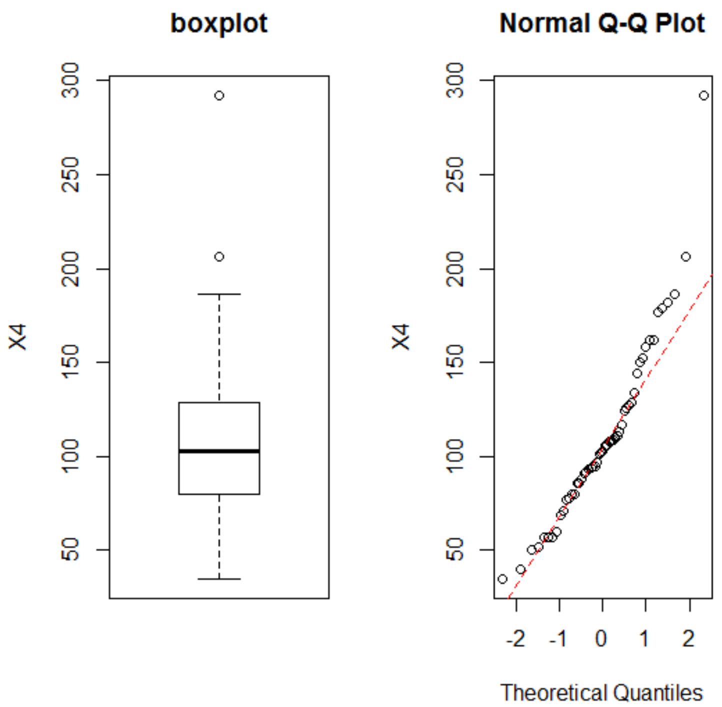
X3 :


density.default(x = X3, bw = "SJ", kernel = "gaussian")
Density
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
6
8
10
12
14
X3


boxplot
X3
7
8
9
10
11
12
13
Normal Q-Q Plot
X3
-2
-1
0
1
2
Theoretical Quantiles

X4:



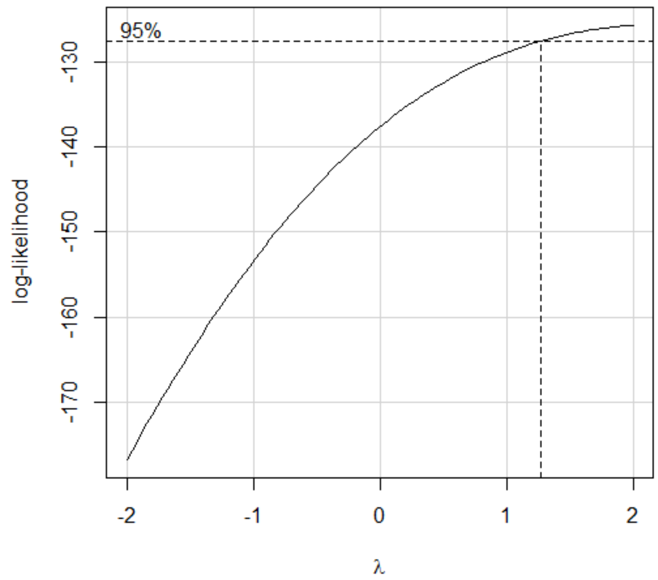


با توجه به نمودار های بالا به وضوح مشخص است که متغیر X4 چولگی دارد و متقارن نیست .

پس تبدیل روی متغیر های مستقل به تنهایی انجام می شود.

library(MASS)

boxCox(M,plotit=T)



> MF<-powerTransform(cbind(X1,X2,X4))

> MF

Estimated transformation parameters

```
        X1         X2         X4
-0.2490971 -0.3755732  0.1584006
```

> summary(MF)

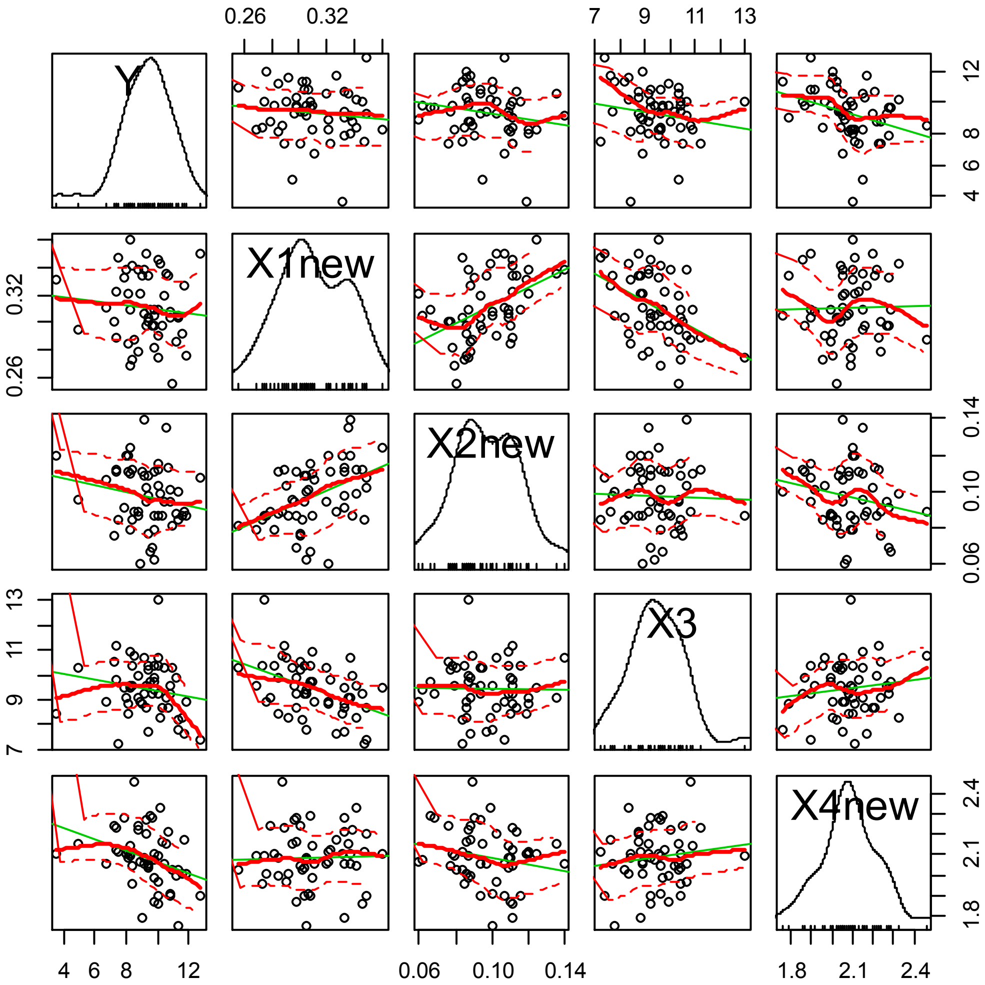
bcPower Transformations to Multinormality

	Est.Power	Std.Err.	Wald Lower Bound	Wald Upper Bound
X1	-0.2491	0.4086	-1.0499	0.5517
X2	-0.3756	0.2277	-0.8220	0.0708
X4	0.1584	0.2575	-0.3463	0.6631

Likelihood ratio tests about transformation parameters

	LRT	df	pval
LR test, lambda = (0 0 0)	3.480189	3	3.233407e-01
LR test, lambda = (1 1 1)	56.940814	3	2.645772e-12

```
> X1new<-X1^-0.2491
> X2new<-X2^-0.3756
> X4new<-X4^0.1584
> scatterplotMatrix(cbind(Y,X1new,X2new,X3,X4new))
```

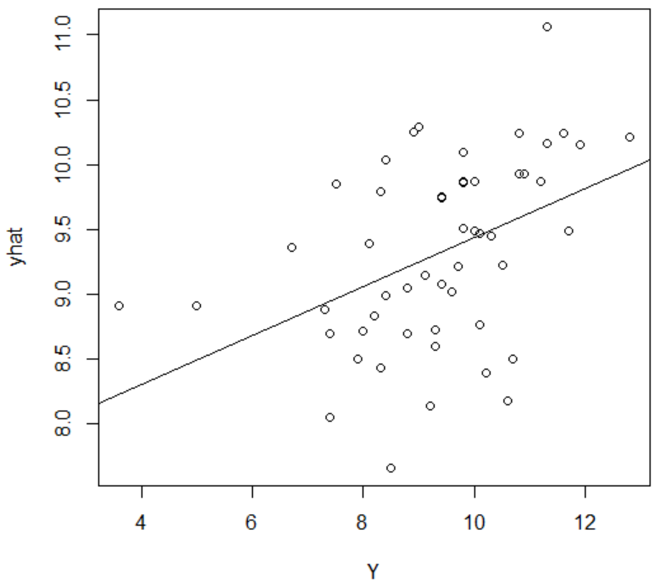


همان طور که ملاحظه می شود تبدیل به خوبی عمل نکرده است . چرا که توزیع متغیر های X1new و X2new نزدیک به نرمال نیست . و ارتباط بین متغیر وابسته و متغیر های مستقل خطی نشده است .

```
M1<-lm(Y~X1new+X2new+X3+X4new)
```

نمودار Y درمقابل yhat

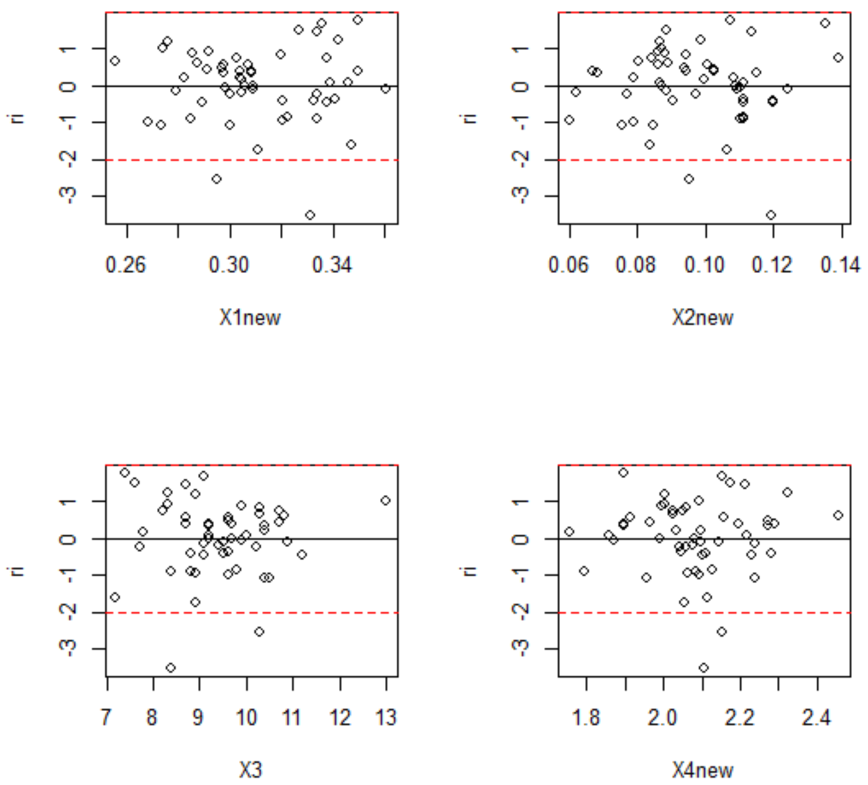
```
> yhat<-fitted(M1)
plot(Y,yhat)
> abline(lsfit(Y,yhat))
```



با توجه به شکل به وضوح می توان فهمید که مدل M1 مدل خوبی نمی باشد چرا که اکثر نقاط روی خط برازش داده شده قرار ندارند .

نمودار باقی مانده های استاندارد شده

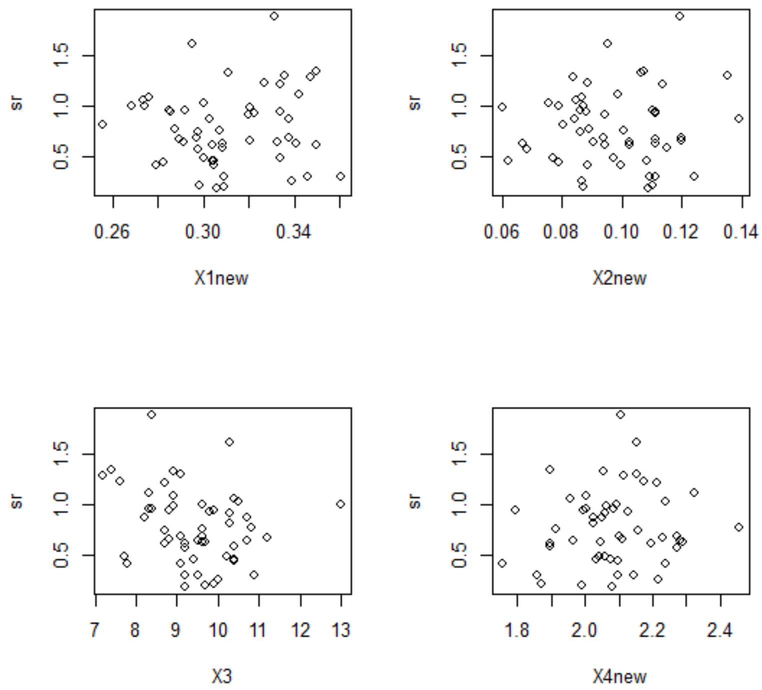
```
> ei<-resid(M1)
> ri<-rstandard(M1)
```



روند غیر تصادفی در نمودار های بالا مشاهده نمی شود. البته در هرکدام از نمودارهای بالا دو نقطه وجود دارند ، که $|r_i|>2$ است . که ممکن است به دلیل مناسب نبودن مدل و سیگنال هایی نشان از بدی مدل برازش شده باشند که شاید با بهبود مدل بهبود یابند .

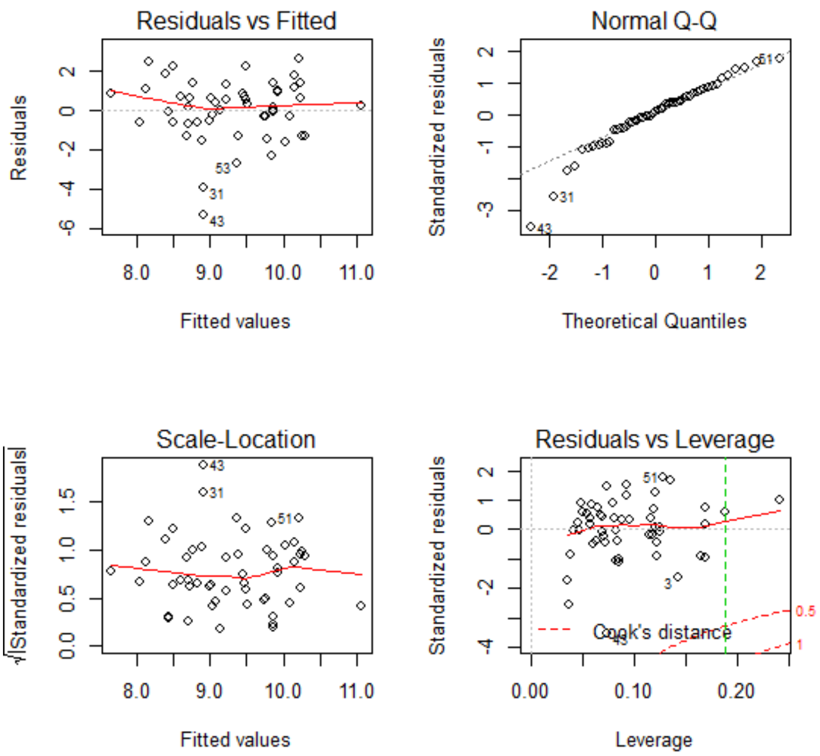
نمودار جذر قدر مطلق باقی مانده های استاندارد شده

```
> sr<-sqrt(abs(ri))
```



روند غیر تصادفی در نمودار های بالا مشاهده نمی شود. و به نظر ثبات واریانس وجود دارد.

نمودارهای مربوط به عیب شناسی رگرسیون



تعداد نقاط اهرمی کمتر شده است .

یافت بهترین مدل از لحاظ معیار های AIC و BIC و AIC_C

```
> om1<-lm(Y~X4new)

> om2<-lm(Y~X4new+X2new)

> om3<-lm(Y~X4new+X2new+X3)

> om4<-lm(Y~X4new+X2new+X3+X1new)

> om5<-M1


> #Subset size=1

> n<-length(om1$residuals)

> n.p<-length(om1$coefficients) +1

> #Calculate AIC

> extractAIC(om1,k=2)

[1]  2.00000    50.75652

> #Calculate AICc

> extractAIC(om1,k=2)+2*n.p*(n.p+1)/(n-n.p-1)

[1]  2.489796    51.246317

> #Calculate BIC

> extractAIC(om1,k=log(n))

[1]  2.00000     54.69711


> #Subset size=2

> n<-length(om2$residuals)

> n.p <- length(om2$coefficients) +1
```

> #Calculate AIC

> extractAIC(om2,k=2)

[1] 3.00000   **49.02499**

> #Calculate AICc

> extractAIC(om2,k=2)+2*n.p*(n.p+1)/(n-n.p-1)

[1] 3.833333   **49.858322**

> #Calculate BIC

> extractAIC(om2,k=log(n))

[1] 3.00000   **54.93586**

> #Subset size=3

> n<-length(om3$residuals)

> n.p <- length(om3$coefficients) +1

> #Calculate AIC

> extractAIC(om3,k=2)

[1] 4.00000 **49.96393**

> #Calculate AICc

> extractAIC(om3,k=2)+2*n.p*(n.p+1)/(n-n.p-1)

[1] 5.276596 **51.240525**

> #Calculate BIC

> extractAIC(om3,k=log(n))

[1] 4.0000   **57.8451**

> **#Subset size=4**

> n<-length(om4$residuals)

```
> n.p <- length(om4$coefficients) +1

> #Calculate AIC

> extractAIC(om4,k=2)

[1]  5.00000    51.81504

> #Calculate AICc

> extractAIC(om4,k=2)+2*n.p*(n.p+1)/(n-n.p-1)

[1]  6.826087    53.641126

> #Calculate BIC

> extractAIC(om4,k=log(n))

[1]  5.0000    61.6665
```

از لحاظ معیار AIC ، om2 بهتر است چرا که کمترین مقدار AIC را داراست . ①

از لحاظ معیار BIC ، om2 بهتر است چرا که کمترین مقدار BIC را داراست . ②

از لحاظ معیار AIC_C ، om1 بهتر است چرا که کمترین مقدار AIC_C را داراست .

```
> backAIC<-step(M1,direction="backward", data=x)
Start:  AIC=51.82
Y ~ X1new + X2new + X3 + X4new

       Df Sum of Sq    RSS    AIC
- X1new  1    0.3282 116.99 49.964
- X3     1    2.6041 119.26 50.985
<none>              116.66 51.815
- X2new  1    4.7286 121.39 51.921
- X4new  1   15.9438 132.60 56.605

Step:  AIC=49.96
Y ~ X2new + X3 + X4new

       Df Sum of Sq    RSS    AIC
- X3     1    2.3657 119.35 49.025
```

```
<none>              116.99 49.964
- X2new  1    8.6958 125.68 51.764
- X4new  1   17.8853 134.87 55.504

Step:  AIC=49.02
Y ~ X2new + X4new

        Df Sum of Sq    RSS    AIC
<none>              119.35 49.025
- X2new  1    8.7061 128.06 50.757
- X4new  1   20.0030 139.36 55.237
```

بهترین مدل از لحاظ AIC با روش پسرو Y ~ X2new + X4new می باشد چرا که کمترین مقدار AIC را دارست . ③

```
> backBIC<-step(M1,direction="backward", data=x, k=log(n))
Start:  AIC=61.67
Y ~ X1new + X2new + X3 + X4new

        Df Sum of Sq    RSS    AIC
- X1new  1    0.3282 116.99 57.845
- X3     1    2.6041 119.26 58.866
- X2new  1    4.7286 121.39 59.802
<none>              116.66 61.666
- X4new  1   15.9438 132.60 64.486

Step:  AIC=57.85
Y ~ X2new + X3 + X4new

        Df Sum of Sq    RSS    AIC
- X3     1    2.3657 119.35 54.936
- X2new  1    8.6958 125.68 57.675
<none>              116.99 57.845
- X4new  1   17.8853 134.87 61.415

Step:  AIC=54.94
Y ~ X2new + X4new
```

```
       Df Sum of Sq    RSS    AIC
- X2new  1    8.7061 128.06 54.697
<none>              119.35 54.936
- X4new  1   20.0030 139.36 59.178

Step:  AIC=54.7
Y ~ X4new

       Df Sum of Sq    RSS    AIC
<none>              128.06 54.697
- X4new  1    15.669 143.73 56.845
```

بهترین مدل از لحاظ BIC با روش پسرو Y ~ X4new می باشد چرا که کمترین مقدار BIC را دارست .

```
> forwardAIC <- step(M1,scope=list(lower=~1,
upper=~X1new+X2new+X3+X4new),direction="forward",data=x)
Start:  AIC=51.82
Y ~ X1new + X2new + X3 + X4new

> forwardBIC <- step(M1,scope=list(lower=~1,
upper=~X1new+X2new+X3+X4new),direction="forward",data=x,k=log(n))
Start:  AIC=61.67
Y ~ X1new + X2new + X3 + X4new
```

یافت بهترین مدل با استفاده از تابع leaps

معیار Cp :

```
> library(leaps)
> data1<-cbind(Y,X1new,X2new,X3,X4new)
> A<-data1[,-1]
> B<-data1[,1]
> leaps(A,B,method="Cp")
$which
     1     2     3     4
1 FALSE FALSE FALSE  TRUE
1 FALSE  TRUE FALSE FALSE
1 FALSE FALSE  TRUE FALSE
1  TRUE FALSE FALSE FALSE
2 FALSE  TRUE FALSE  TRUE
```

2 FALSE FALSE TRUE TRUE
2 TRUE FALSE FALSE TRUE
2 TRUE FALSE TRUE FALSE
2 FALSE TRUE TRUE FALSE
2 TRUE TRUE FALSE FALSE
3 FALSE TRUE TRUE TRUE
3 TRUE TRUE FALSE TRUE
3 TRUE FALSE TRUE TRUE
3 TRUE TRUE TRUE FALSE
4 TRUE TRUE TRUE TRUE

$label
[1] "(Intercept)" "1"   "2"
[4] "3"   "4"

$size
 [1] 2 2 2 2 3 3 3 3 3 3 4 4 4 4 5

$Cp
 [1] 3.690571 8.338742 8.388450 9.519627 2.108407
 [6] 4.712985 5.140294 8.040685 8.494032 10.306278
[11] 3.135034 **4.071486** 4.945625 9.560176 5.000000

> 4-3.690571
[1] 0.309429
> 4-4.071486
[1] -0.071486

دراین معیار اگر مقدار cp به p نزدیک باشد مدل ، مدل مناسبی خواهد بود . که با توجه به این معیار مدل دوازدهم یعنی $Y = X1new + X2new + X4new$ با مقدار cp ، 4.071486 بهترین مدل خواهد بود .

معیار R^2_{adj} :

> leaps(A,B,method="adjr2")
$which
    1   2   3   4
1 FALSE FALSE FALSE TRUE
1 FALSE TRUE FALSE FALSE
1 FALSE FALSE TRUE FALSE
1 TRUE FALSE FALSE FALSE
2 FALSE TRUE FALSE TRUE

```
2 FALSE FALSE  TRUE  TRUE
2  TRUE FALSE FALSE  TRUE
2  TRUE FALSE  TRUE FALSE
2 FALSE  TRUE  TRUE FALSE
2  TRUE  TRUE FALSE FALSE
3 FALSE  TRUE  TRUE  TRUE
3  TRUE  TRUE FALSE  TRUE
3  TRUE FALSE  TRUE  TRUE
3  TRUE  TRUE  TRUE FALSE
4  TRUE  TRUE  TRUE  TRUE

$label
[1] "(Intercept)" "1"          "2"
[4] "3"          "4"

$size
 [1] 2 2 2 2 3 3 3 3 3 3 4 4 4 4 5

$adjr2
 [1]  0.091550710  0.011410601  0.010553574 -0.008949283
 [5]  0.136377794  0.090573587  0.083058915  0.032052535
 [9]  0.024079962 -0.007790261  0.136219968  0.119415390
[13]  0.103729021  0.020921211  0.120698201
```

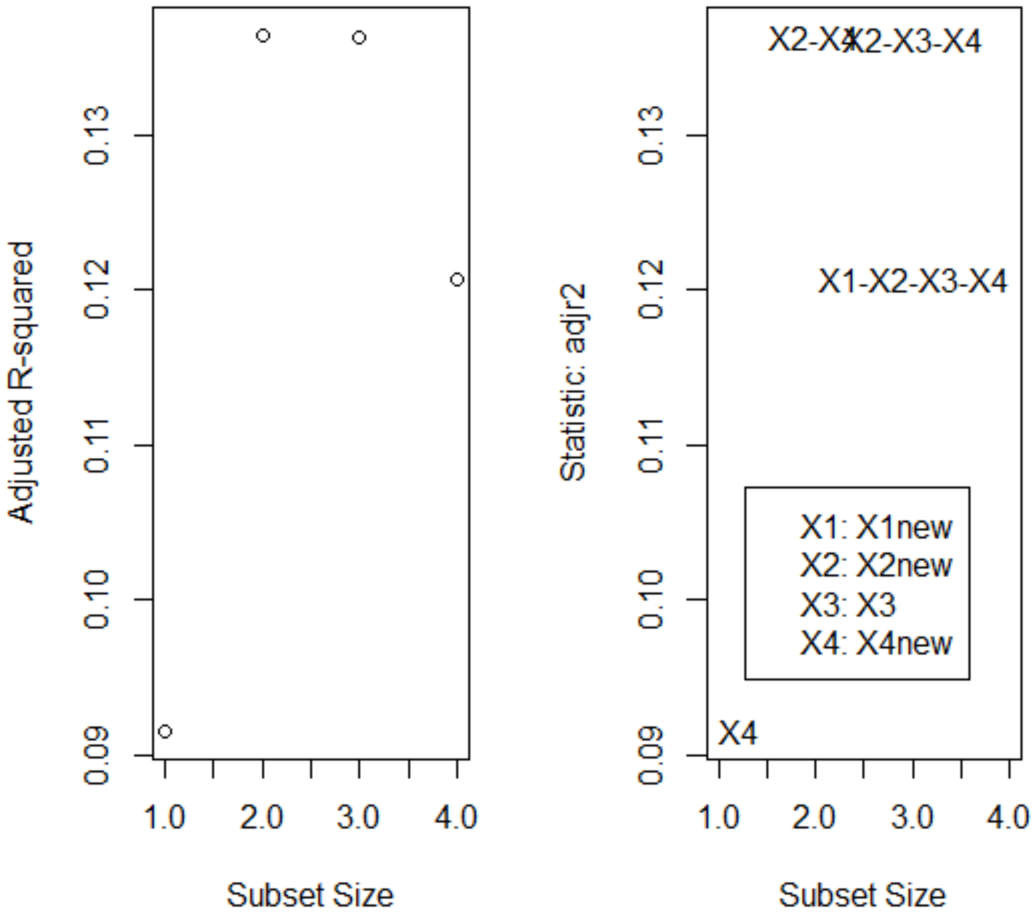
دراین معیار هرچه مقدار R^2_{adj} بیشتر باشد مدل ، مدل بهتری خواهد بود . که با توجه به این معیار مدل پنجم یعنی $Y = X2new + X4new$ با مقدار R^2_{adj} ، 0.13637779 بهترین مدل خواهد بود . ④

نمودار R^2_{adj}

```
library(leaps)

b <- regsubsets(as.matrix(A),B)

rs <- summary(b)

par(mfrow=c(1,2))

plot(1:4,rs$adjr2,xlab="Subset Size",ylab="Adjusted R-squared")
```

library(car)

subsets(b,statistic=c("adjr2"))



با توجه به نمودار بالا در می یابیم که بالا ترین میزان R^2_{adj} مربوط به مدل $Y = X2new + X4new$ می باشد.

عامل تورم واریانس

```
> library(car)

> VIF<-vif(M1)

  X1new    X2new      X3    X4new

1.769074 1.469807 1.384714 1.117273
```

همان طور که ملاحظه می شود متغیر های مستقل هم خطی ندارند .

با چهار مورد از سه معیار از معیار های بالا مدل $Y = X2new + X4new$ مدل مناسب انتخاب شده است . بنابراین مدل نهایی خواهد بود .

مدل بندی

> modelnahayii<-lm(Y~X2new+X4new)

> modelnahayii

Call:

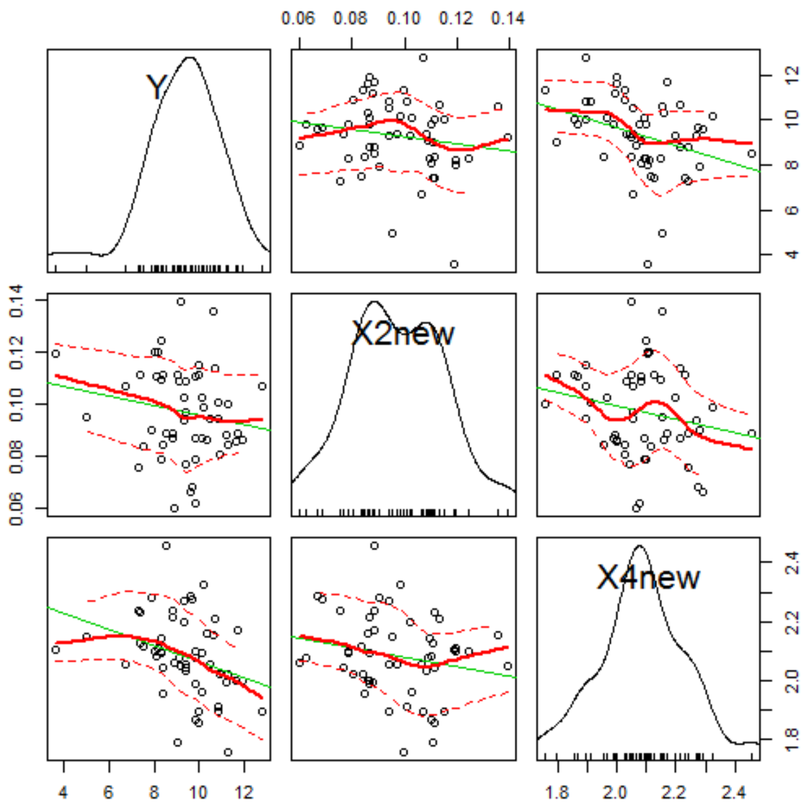
lm(formula = Y ~ X2new + X4new)

Coefficients:

(Intercept)   X2new   X4new

21.157   -23.895   -4.575

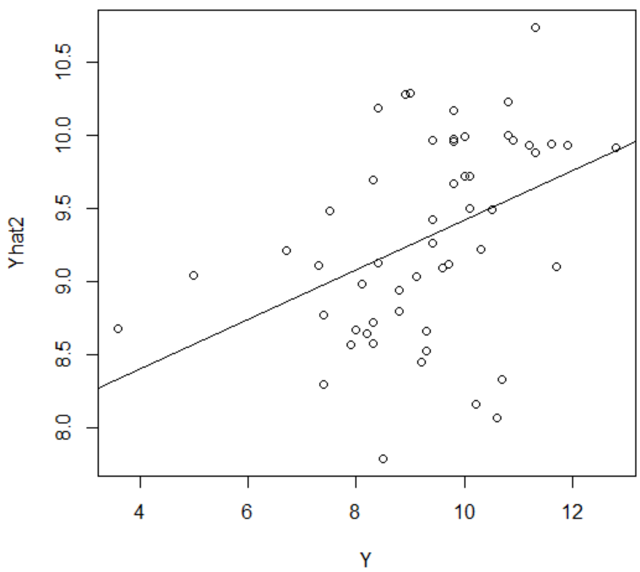
نمودار پراکنش



نمودار Y درمقابل yhat

یکی از راه های شهودی بررسی نیکویی برازش مدل نهایی رسم این نمودار می باشد .

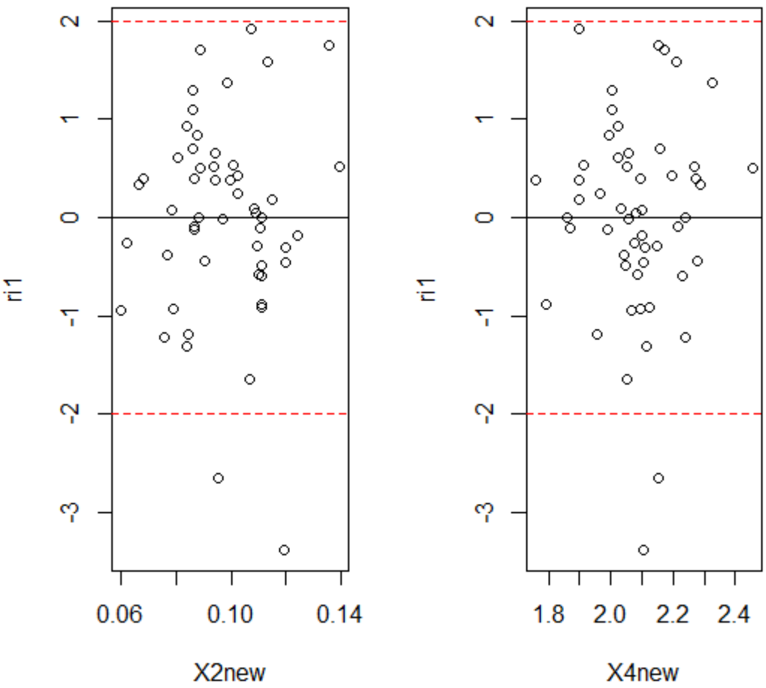
```
>Yhat2<-fitted(modelnahayii)
> plot(Y,Yhat2)
> abline(lsfit(Y,Y hat2))
```



با توجه به شکل به وضوح می توان فهمید که مدل نهایی نیز مدل خوبی نمی باشد چرا که اکثر نقاط روی خط برازش داده شده قرار ندارند .

نمودار باقی مانده های استاندارد شده

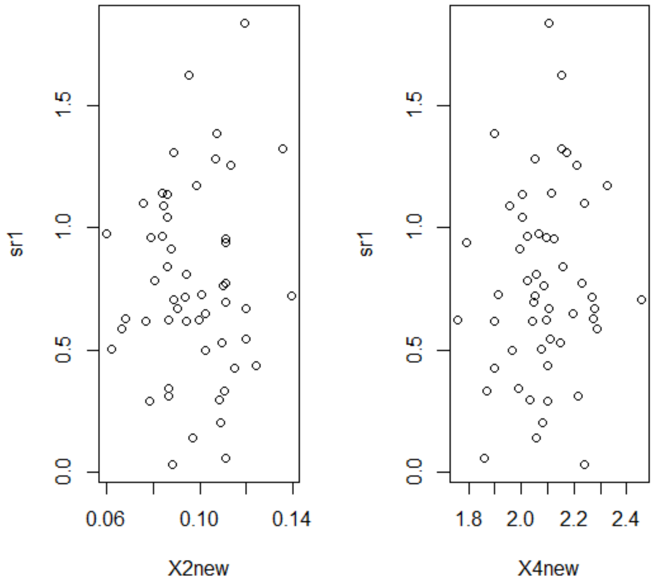
مانده ها برخلاف خطاها هم واریانس نیستند ولی مانده های استاندارد شده هم واریانس می باشند.



روند غیر تصادفی در نمودار های بالا مشاهده **نمی شود**. البته در هرکدام از نمودارهای بالا دو نقطه وجود دارند ، که $|r_i|>2$ است . که ممکن است به دلیل مناسب نبودن مدل و سیگنال هایی نشان از بدی مدل برازش شده باشند **شاید** با بهبود مدل بهبود یابند .

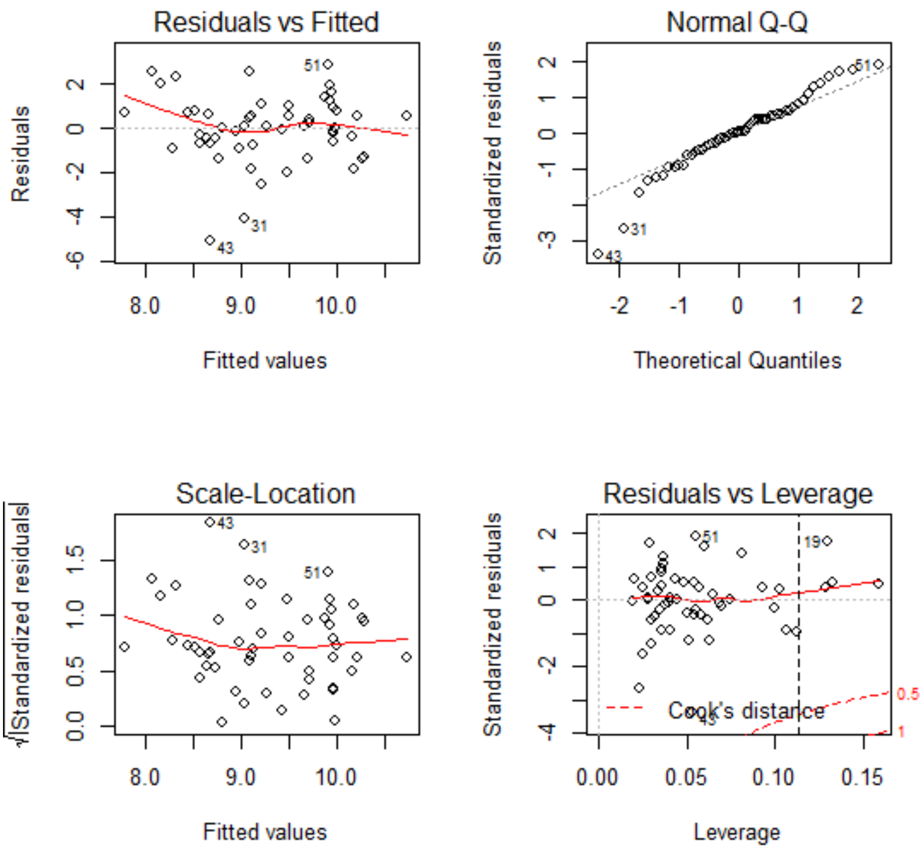
نمودار جذر قدر مطلق باقی مانده های استاندارد شده

به دلیل این که مانده های استاندارد شده دارای علامت + و — می باشند و ممکن است یکدیگر را خنثی کنند و تاثیر آن ها به طور کامل بررسی نشود، به منظور بررسی دقیق تر از قدر مطلق آن ها استفاده می کنیم . و توان 1/2 به دلیل کاهش مقدار مطلق چولگی داده هاست .



روند غیر تصادفی در نمودار های بالا مشاهده نمی شود. و ثبات واریانس وجود دارد.

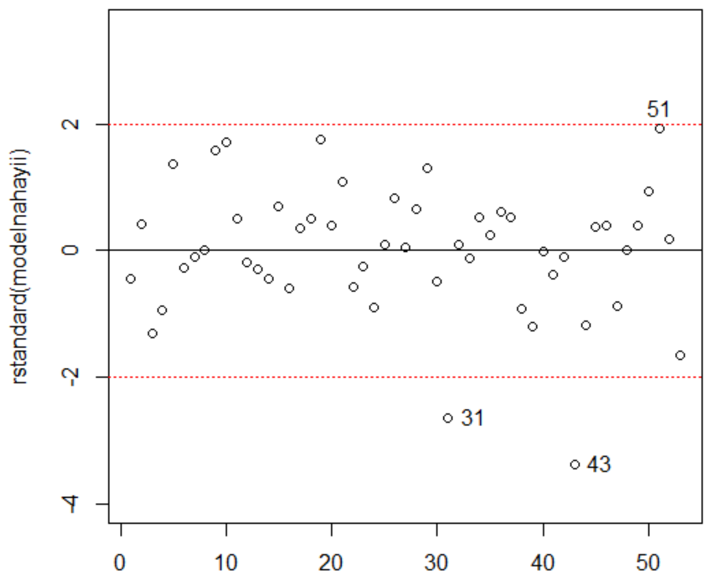
نمودارهای مربوط به عیب شناسی رگرسیون



با توجه به شکل سوم در می یابیم که ثبات واریانس داریم وبا توجه به شکل آخر در می یابیم که چهار نقطه ی اهرمی داریم .

بررسی نقاط اهرمی و پرت

مقدار بحرانی برای **h_{ii}** به صورت $h_{ii} > \frac{2\times(p+1)}{n} = 0.1132075$ می باشد . که طبق آخرین نمودار از نمودار های عیب شناسی رگرسیون چهار نقطه ی اهرمی داشتیم یعنی مقدار **h_{ii}** آن ها بزرگتر از ۰.۱۱۳۲ بوده است . که با توجه به شکل زیر می توانیم نقطه ی اهرمی بد و داده های پرت با تاثیر زیاد را روی مدل تشخیص دهیم .



داده های ۳۱ و۴۳ و ۵۱ یا داده ی پرت ویا اهرمی بد هستند چرا که برای آن ها $|r_i|>2$ است. مدل را بدون حضور این داده بررسی می کنیم . البته حذف کردن داده ها کار درستی نیست و در ابتدا باید دلیل وجود داده های پر تاثیر درمدل بررسی شود . ولی در اینجا داده ها را حذف کرده ایم تا نشان دهیم که با حذف این داده ها نیز مدل مناسب نخواهد بود و شاید این داده ها در مدل مناسب دیگری اصلا پرت نباشند.

نتیجه گیری

تبدیل روی این داده به خوبی عمل نکرد ، و ارتباط بین متغیر وابسته ومتغیر های مستقل خطی نشد ، بنیابراین چاره ای جز استفاده از روش های رگرسیون ناپارامتری وجود ندارد .