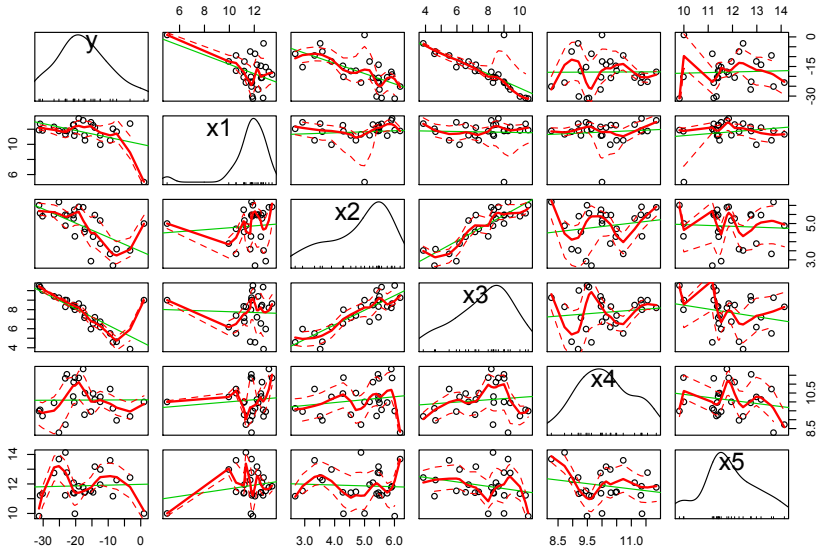
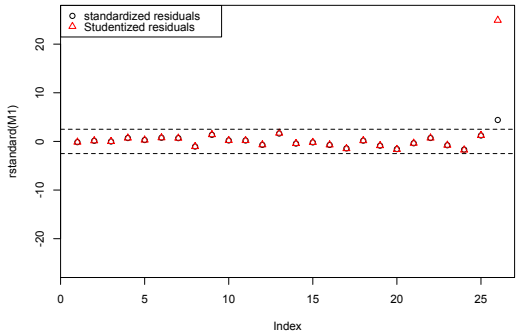
```
rm(list=ls())
n=25
x11<-rnorm(n,12,1)
x22<-rnorm(n,5,1)
x33<-1.5*x22+rnorm(25,.4)
x44<-rnorm(n,10,1)
x55<-rnorm(n,12,1)
Err<-rnorm(n,0,.5)
y1=2*x22+(-5)*x33+x44+Err
x1<-c(x11,5)
x2<-c(x22,5)
x3<-c(x33,9)
x4<-c(x44,10)
x5<-c(x55,10)
y<-c(y1,1)
##### data
data<-cbind(y=y,x1=x1,x2=x2,x3=x3,x4=x4,x5=x5)
data
> library(car)
> scatterplotMatrix(data)
```

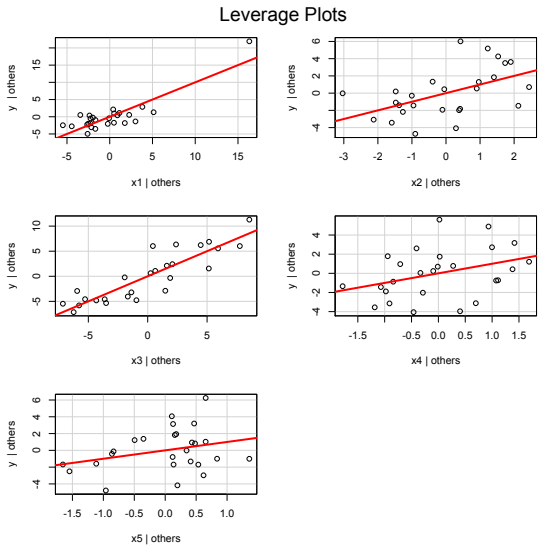



```
> M1<-lm(y~x1+x2+x3+x4+x5)
> summary(M1) # summary of estimated model
Call:
lm(formula = y ~ x1 + x2 + x3 + x4 + x5)
Residuals:
    Min      1Q  Median      3Q     Max
-4.3670 -1.8050  0.1526  1.7052  5.5881
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  38.4609    10.3249   3.725  0.00134 **
x1           -2.7390     0.3524  -7.772 1.82e-07 ***
x2            2.9988     1.1185   2.681  0.01436 *
x3           -5.2430     0.6318  -8.298 6.59e-08 ***
x4            1.0550     0.6073   1.737  0.09771 .
x5           -0.7712     0.5506  -1.401  0.17665
```

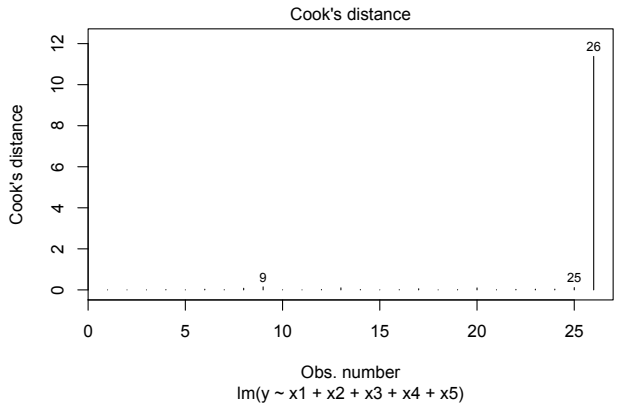
```
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 2.696 on 20 degrees of freedom
Multiple R-squared:  0.9129,     Adjusted R-squared:  0.8912
F-statistic: 41.94 on 5 and 20 DF,  p-value: 6.276e-10
-----------------------------------------------------------------
> anova(M1)
Analysis of Variance Table
Response: y
          Df Sum Sq Mean Sq F value    Pr(>F)
x1         1 361.02  361.02 49.6599 7.804e-07 ***
x2         1 596.81  596.81 82.0942 1.618e-08 ***
x3         1 518.96  518.96 71.3846 4.964e-08 ***
x4         1  33.47   33.47  4.6045   0.04434 *
x5         1  14.26   14.26  1.9618   0.17665
Residuals 20 145.40    7.27
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
## plot Standardized residuals,Studentized residuals
> aa=abs(max(rstudent(M1)))
> plot(rstandard(M1),ylim=c(-aa-1,aa+1))
> points(rstudent(M1),pch=2,col=2)
> legend("topleft",legend=c("standardized residuals","Studentized
 residuals"),col=1:2,pch=1:2)
> abline(h=c(-2.5,2.5),lty=2)
```




```
> leveragePlots(M1) # leverage plots
```

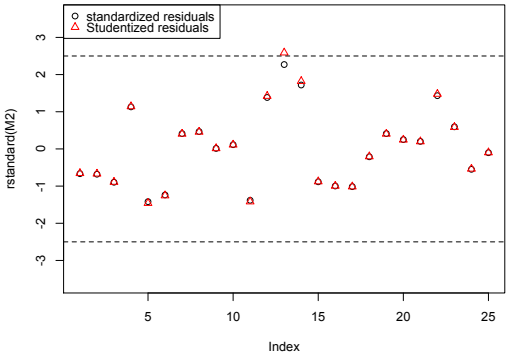


```
# Cook's D plot# identify D values > 4/(n-k-1)
> cutoff <- 4/((nrow(mtcars)-length(M1$coefficients)-2))
> plot(M1, which=4, cook.levels=cutoff)
```




```
> outlierTest(M1) # Bonferonni p-value for most extreme obs
rstudent unadjusted p-value Bonferonni p
26 24.90604         5.7104e-16   1.4847e-14

## without outlier observation that is 26 observation
> M2<-lm(y~x1+x2+x3+x4+x5,subset=-26)
> summary(M2) # summary of estimated model
Call:
lm(formula = y ~ x1 + x2 + x3 + x4 + x5, subset = -26)
Residuals:
     Min       1Q   Median       3Q      Max
-0.62597 -0.34919  0.00501  0.18394  0.93783
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.7075     2.4100  -0.294    0.772
x1           -0.1677     0.1206  -1.391    0.180
x2            2.3821     0.1994  11.948 2.79e-10 ***
x3           -5.1275     0.1118 -45.843  < 2e-16 ***
x4            1.0320     0.1074   9.607 9.99e-09 ***
x5            0.1218     0.1038   1.174    0.255
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 0.4769 on 19 degrees of freedom
Multiple R-squared:  0.9967,     Adjusted R-squared:  0.9958
F-statistic:  1141 on 5 and 19 DF,  p-value: < 2.2e-16
---------------------------------------------------------------------
> anova(M2)
Analysis of Variance Table
Response: y
          Df Sum Sq Mean Sq   F value    Pr(>F)
x1         1  25.96   25.96  114.1249 1.797e-09 ***
x2         1 678.64  678.64 2983.9654 < 2.2e-16 ***
x3         1 571.52  571.52 2512.9541 < 2.2e-16 ***
x4         1  20.92   20.92   91.9945 1.026e-08 ***
x5         1   0.31    0.31    1.3772    0.2551
Residuals 19   4.32    0.23
---
Signif. codes:
```

```
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
-----------------------------------------------------------------
## plot standardized residuals,Studentized residuals
> aa=abs(max(rstudent(M2)))
> plot(rstandard(M2),ylim=c(-aa-1,aa+1))
> points(rstudent(M2),pch=2,col=2)
> legend("topleft",legend=c("standardized residuals","Studentized
 residuals"),col=1:2,pch=1:2)
> abline(h=c(-2.5,2.5),lty=2)
```




```
-----------------------------------------------------------------
> outlierTest(M2) # Bonferonni p-value for most extreme obs
```

No Studentized residuals with Bonferonni p < 0.05
Largest |rstudent|:
  rstudent unadjusted p-value Bonferonni p
13 2.588012   0.018561   0.46403

```
-----------------------------------------------------------------
### final model without outlier data
> FM<-lm(y~x1+x2+x3+x4+x5,subset=-c(26,13))
> summary(FM) # summary of estimated model
Call:
lm(formula = y ~ x1 + x2 + x3 + x4 + x5, subset = -c(26, 13))
Residuals:
     Min       1Q   Median       3Q      Max
-0.53168 -0.30678 -0.00818  0.18533  0.65545
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.26962    2.24763   0.565   0.5791
x1          -0.26325    0.11203  -2.350   0.0304 *
x2           2.31063    0.17703  13.052 1.29e-10 ***
x3          -5.09939    0.09870 -51.664  < 2e-16 ***
x4           0.97654    0.09662  10.107 7.58e-09 ***
x5           0.10514    0.09124   1.152   0.2643
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 0.4183 on 18 degrees of freedom
Multiple R-squared:  0.9976,     Adjusted R-squared:  0.9969
F-statistic:  1484 on 5 and 18 DF,  p-value: < 2.2e-16
-----------------------------------------------------------------
> anova(FM)
Analysis of Variance Table
Response: y
          Df Sum Sq Mean Sq   F value    Pr(>F)
```

```
x1         1  27.94   27.94  159.6732 2.187e-10 ***
x2         1 707.89  707.89 4045.9770 < 2.2e-16 ***
x3         1 544.29  544.29 3110.9064 < 2.2e-16 ***
x4         1  17.92   17.92  102.4462 7.413e-09 ***
x5         1   0.23    0.23    1.3277    0.2643
Residuals 18   3.15    0.17
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

---

best model by Cp

```
> X<-model.matrix(FM)[,-1]
> library(leaps)
> outs <- leaps(X, y[c(-13,-26)], int = FALSE,
 method="Cp",strictly.compatible = FALSE)
```

```
> cbind(outs$which,Cp=outs$Cp,size=outs$size)
   x1 x2 x3 x4 x5     Cp     size
1  0  0  1  0  0 1388.688607    1
1  0  1  0  0  0 3729.742899    1
1  1  0  0  0  0 7531.032567    1
1  0  0  0  1  0 8181.693743    1
1  0  0  0  0  1 8902.539839    1
2  0  0  1  1  0  215.908843    2
2  0  0  1  0  1  378.529929    2
2  1  0  1  0  0  449.517047    2
2  0  1  1  0  0  571.860477    2
2  0  1  0  0  1 3083.627508    2
2  0  1  0  1  0 3238.707146    2
2  1  1  0  0  0 3427.468487    2
2  1  0  0  0  1 7392.407276    2
2  1  0  0  1  0 7531.256562    2
2  0  0  0  1  1 8169.861448    2
3  0  1  1  1  0    9.617147    3
3  0  1  1  0  1  187.412214    3
3  0  0  1  1  1  190.773240    3
3  1  1  1  0  0  191.394075    3
3  1  0  1  1  0  217.841041    3
3  1  0  1  0  1  350.919329    3
3  1  1  0  0  1 3032.833843    3
3  0  1  0  1  1 3074.894690    3
3  1  1  0  1  0 3229.536567    3
3  1  0  0  1  1 7390.563141    3
4  1  1  1  1  0    8.366348    4
4  0  1  1  1  1   10.549473    4
4  1  1  1  0  1  160.030621    4
4  1  0  1  1  1  183.779540    4
4  1  1  0  1  1 2946.742870    4
5  1  1  1  1  1    5.000000    5
```

```
## what is best Model? Stepwise method
> library(MASS)
> step<-stepAIC(FM,direction="both")
Start:  AIC=-36.74
```

```
y ~ x1 + x2 + x3 + x4 + x5
       Df Sum of Sq    RSS     AIC
- x5    1      0.23   3.38 -37.033
<none>                3.15 -36.741
- x1    1      0.97   4.12 -32.320
- x4    1     17.87  21.02   6.821
- x2    1     29.81  32.95  17.610
- x3    1    466.99 470.14  81.400

Step:  AIC=-37.03
y ~ x1 + x2 + x3 + x4

       Df Sum of Sq    RSS     AIC
<none>                3.38 -37.033
+ x5    1      0.23   3.15 -36.741
- x1    1      1.28   4.66 -31.347
- x4    1     17.92  21.31   5.142
- x2    1     35.06  38.45  19.309
- x3    1    543.43 546.81  83.025

> step$anova # display results
Initial Model:
y ~ x1 + x2 + x3 + x4 + x5
Final Model:
y ~ x1 + x2 + x3 + x4
```

```
> vif(FM)
x1       x2       x3       x4       x5
1.084438 4.263762 4.359252 1.077980 1.275930
> sqrt(vif(FM))> 2 # problem?
   x1   x2   x3   x4   x5
FALSE  TRUE  TRUE FALSE FALSE
```

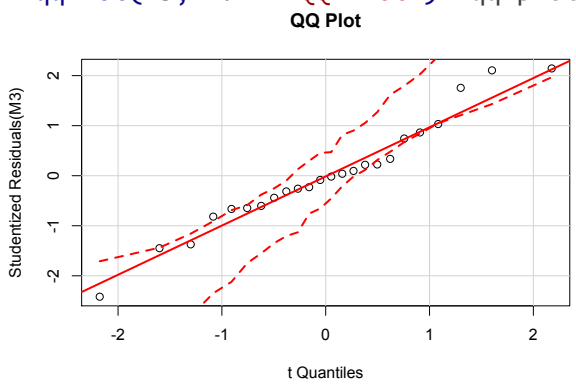
```
> x<-cbind(x1,x2,x3,x4,x5)
> rcorr(x, type="pearson")
      x1    x2    x3    x4    x5
x1  1.00  0.08 -0.04  0.10  0.18
x2  0.08  1.00  0.85  0.18 -0.05
x3 -0.04  0.85  1.00  0.13 -0.25
x4  0.10  0.18  0.13  1.00 -0.20
x5  0.18 -0.05 -0.25 -0.20  1.00

P
   x1     x2     x3     x4     x5
x1        0.6801 0.8527 0.6227 0.3732
x2 0.6801        0.0000 0.3920 0.7994
x3 0.8527 0.0000        0.5397 0.2242
x4 0.6227 0.3920 0.5397        0.3195
x5 0.3732 0.7994 0.2242 0.3195
```

```
## best model
M3<-lm(y~-1+x1+x3+x4)
summary(M3) # summary of estimated model
Call:
```

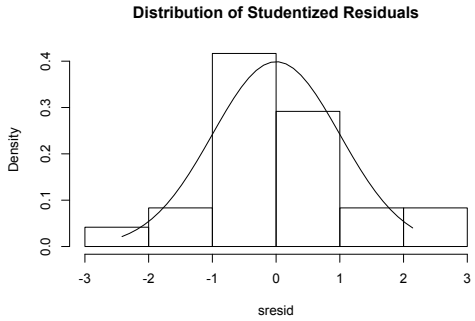
```
lm(formula = y ~ -1 + x1 + x3 + x4, subset = -c(26, 13))

Residuals:
     Min       1Q   Median       3Q      Max
-2.72946 -0.78800 -0.06554  0.57456  2.65762
Coefficients:
   Estimate Std. Error t value Pr(>|t|)
x1  0.01758    0.22630   0.078  0.93880
x3 -4.01504    0.15732 -25.522  < 2e-16 ***
x4  1.17739    0.25812   4.561  0.00017 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.376 on 21 degrees of freedom
Multiple R-squared:  0.9958,     Adjusted R-squared:  0.9952
F-statistic:  1675 on 3 and 21 DF,  p-value: < 2.2e-16
```

```
> anova(M3)
Analysis of Variance Table
Response: y
          Df Sum Sq Mean Sq  F value    Pr(>F)
x1         1 8285.3  8285.3 4373.407 < 2.2e-16 ***
x3         1 1194.9  1194.9  630.738 < 2.2e-16 ***
x4         1   39.4    39.4   20.807 0.0001698 ***
Residuals 21   39.8     1.9
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
> qqPlot(M3, main="QQ Plot") #qq plot for studentized resid
```




```
library(MASS)
sresid <- studres(M3)
hist(sresid, freq=FALSE, main="Distribution of Studentized Residuals")
xfit<-seq(min(sresid),max(sresid),length=40)
yfit<-dnorm(xfit)
lines(xfit, yfit)
```

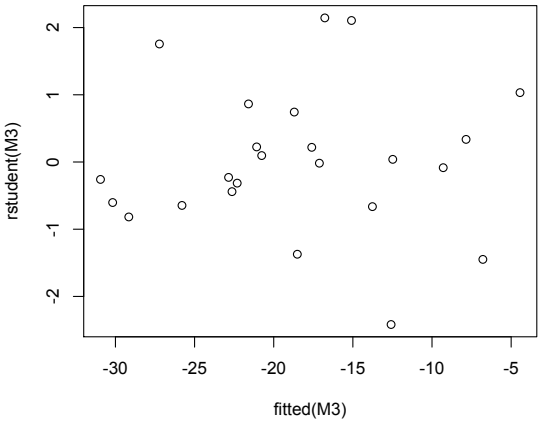


```
# Test for Autocorrelated Errors
> durbinWatsonTest(M3)
lag Autocorrelation D-W Statistic p-value
   1      -0.3492845      2.660145    0.07
 Alternative hypothesis: rho != 0
```

```
# Evaluate homoscedasticity# non-constant error variance test
> ncvTest(M3)
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 0.7987605    Df = 1     p = 0.3714642
```

```
>data.frame( residuals(M3),rstudent(M3),rstandard(M3),fitted(M3))
residuals.M3. rstudent.M3. rstandard.M3. fitted.M3.
1     0.43402059   0.33637721    0.34371317  -7.848626
2     0.30795988   0.22448693    0.22974139 -21.078954
3    -0.30878596  -0.22899397   -0.23434199 -22.852084
4     2.65761928   2.14385314    1.98094080 -16.772598
5    -0.86218310  -0.64609895   -0.65525146 -25.796000
6     0.99616112   0.74301785    0.75107110 -18.701802
7     0.13120319   0.09569846    0.09803930 -20.759353
8    -0.33150793  -0.25918282   -0.26513846 -30.944118
9    -2.72945880  -2.41802931   -2.17955215 -12.585265
10    2.06349864   1.75587735    1.67477592 -27.221513
11   -0.89107094  -0.66401830   -0.67303783 -13.764677
12    2.61647930   2.10526467    1.95180129 -15.087302
14   -0.10768300  -0.08462184   -0.08669607  -9.291322
15   -0.40955431  -0.31433918   -0.32130908 -22.307798
16   -0.58142958  -0.44100772   -0.44971711 -22.635488
17    1.15952870   0.86404390    0.86930521 -21.594025
18   -1.76768386  -1.44898155   -1.41247514  -6.787483
19    0.30145556   0.21818630    0.22330882 -17.597325
20    0.05104943   0.03938999    0.04036116 -12.480244
21   -1.04356514  -0.81750712   -0.82404063 -29.156483
22   -0.76326813  -0.60305144   -0.61240111 -30.175812
23   -1.74335672  -1.37216750   -1.34420312 -18.509019
24   -0.02339832  -0.01768099   -0.01811749 -17.115264
25    1.15770549   1.03217731    1.03057405  -4.439831
```

```
> plot(fitted(M3),rstudent(M3))
```


rstudent(M3)
2
1
0
-1
-2
-30
-25
-20
-15
-10
-5
fitted(M3)