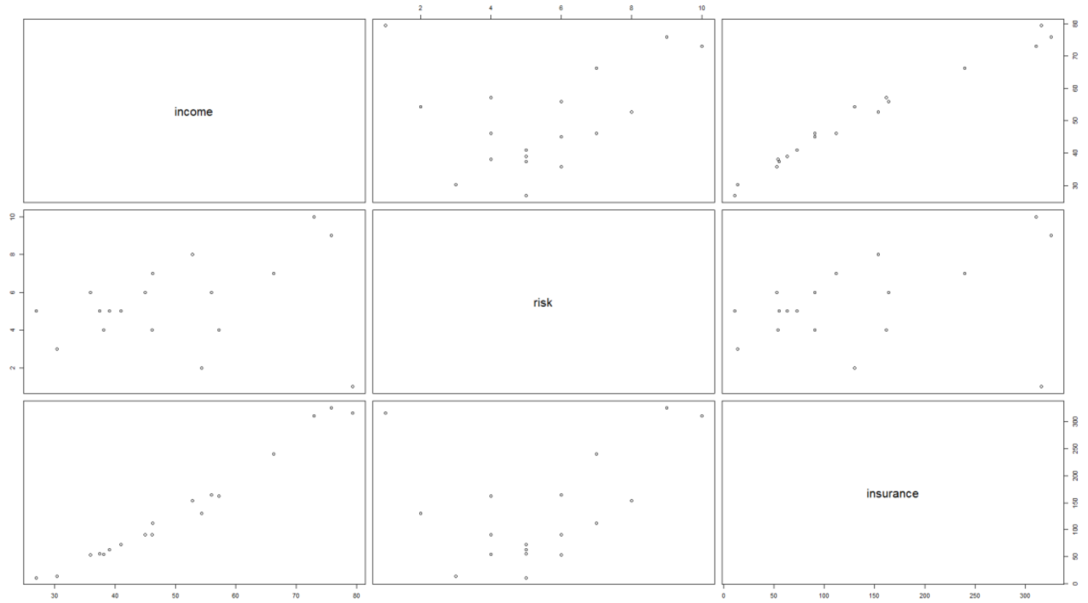
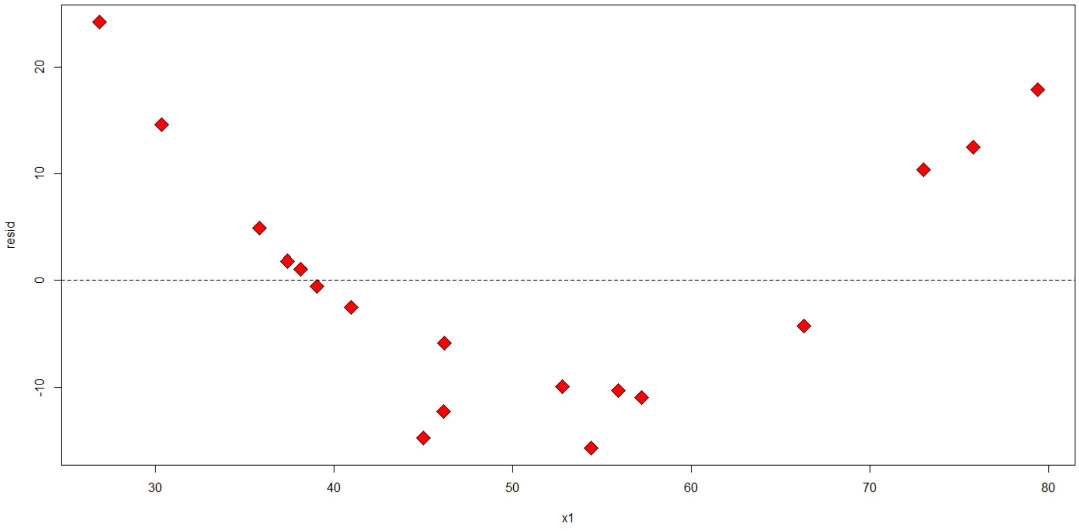
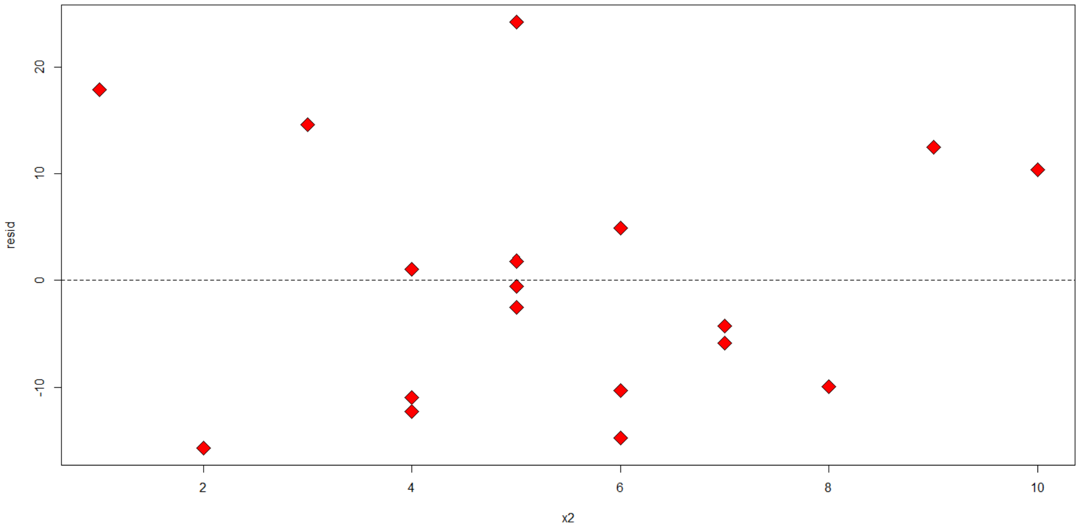
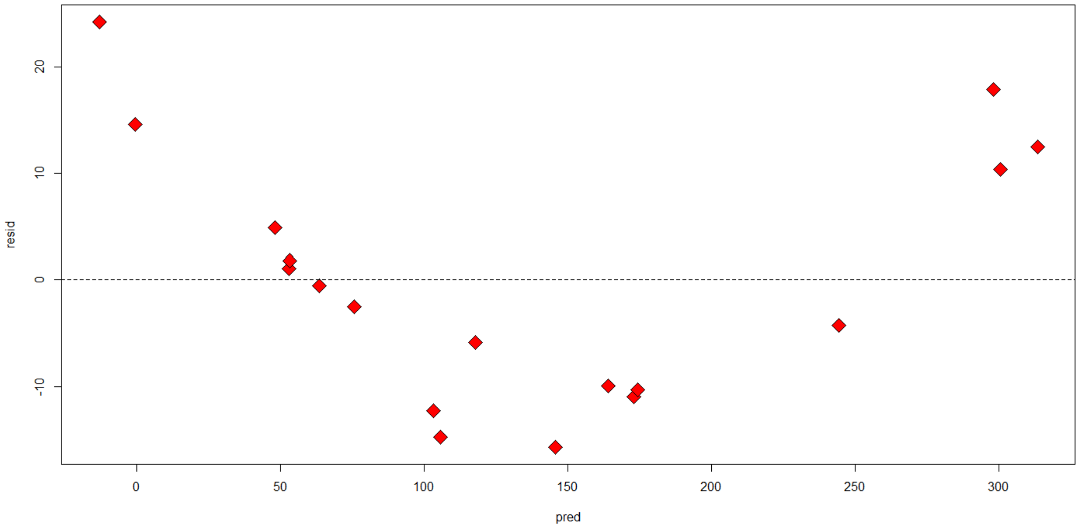
```
##############################################################
################  Handout #9 for ST440/540  ################
############### outliers and influential data#################
##############################################################
##Life Insurance Example: a portion of the data on average annual income of
##managers during the past
##two years (x1), a score measuring each manger's risk aversion (x2), and the
##amount of life insurance carried (y) for a sample of 18 managers in the 30-39
##age group. Risk aversion was measured by a standard questionnaire administered
##to each manager: the higher the score, the greater the degree of risk aversion.
##Income and risk aversion are mildly correlated, the coefficient of correlation
##is r12 =.254.

> ex.data
   income risk insurance
1  45.010    6       91
2  57.204    4      162
3  26.852    5       11
4  66.290    7      240
5  40.964    5       73
6  72.996   10      311
7  79.380    1      316
8  52.766    8      154
9  55.916    6      164
10 38.122    4       54
11 35.840    6       53
12 75.796    9      326
13 37.408    5       55
14 54.376    2      130
15 46.186    7      112
16 46.130    4       91
17 30.366    3       14
18 39.060    5       63

> myfit<-lm(y~x1+x2, data=ex.data)
> summary(myfit)

Call:
lm(formula = y ~ x1 + x2, data = ex.data)


Coefficients:
              Estimate     Std. Error  t value    Pr(>|t|)
(Intercept) -205.7187    11.3927   -18.057   1.38e-11 ***
x1            6.2880       0.2041    30.801    5.63e-15 ***
x2            4.7376       1.3781    3.438     0.00366 **
---
Residual standard error: 12.66 on 15 degrees of freedom
Multiple R-squared:  0.9864,    Adjusted R-squared:  0.9845
```

```
F-statistic: 542.3 on 2 and 15 DF,  p-value: 1.026e-14

> anova(myfit)
Analysis of Variance Table

Response: y
          Df    Sum Sq   Mean Sq      F value      Pr(>F)
x1         1   172024   172024      1072.851    2.268e-15 ***
x2         1   1895       1895         11.819      0.003662 **
Residuals 15   2405     160
---
```



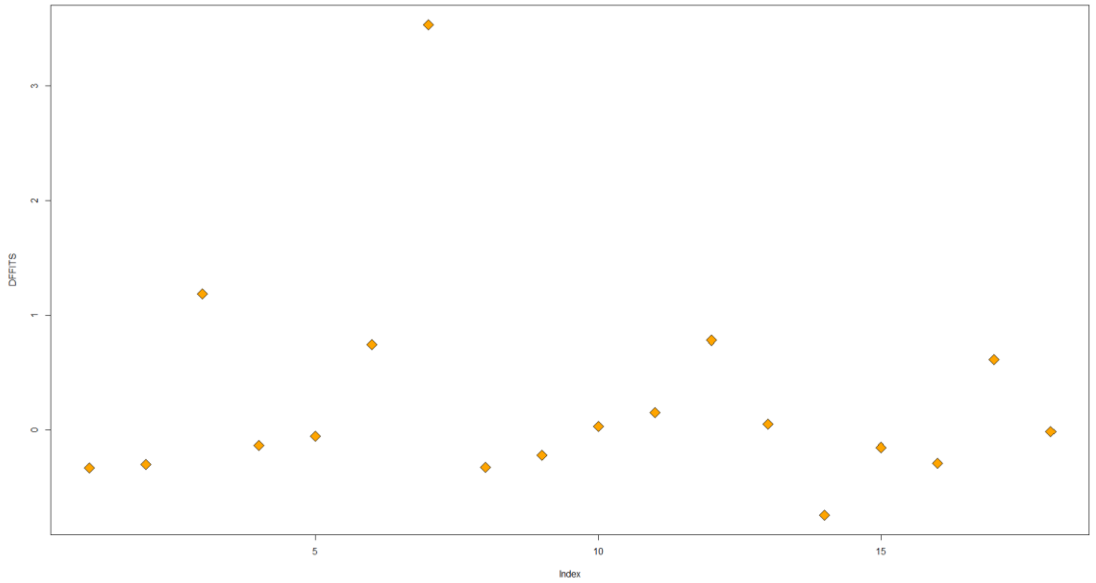






```
> rstudent(myfit)  ##gives rstudent values
          1           2           3           4           5           6 
-1.22592579 -0.90484533  2.44867347 -0.35178460 -0.20281761  1.01382844 
          7           8           9          10          11          12 
 2.74826933 -0.83709929 -0.83362782  0.08497349  0.40331472  1.19332347 
         13          14          15          16          17          18 
 0.   14506769 -1.44149247 -0.47418536 -1.01204637  1.30041597 -0.04624043 

> library(car)
> outlierTest(myfit)  ##R

No Studentized residuals with Bonferonni p < 0.05
Largest |rstudent|:
    rstudent     unadjusted p-value Bonferonni p  #use Bonferonni p-value
7   2.748269           0.015698            0.28257
```

```
> ##leverage, x outliers
> aa<-lm.influence(myfit)   ##hat: a vector containing the diagonal of the "hat" matrix.
> xoutliers <- which(aa$hat > .333)  #0.33= 2*3/18
> xoutliers
6 7
6 7
> x1[xoutliers]
[1] 72.996 79.380
> x2[xoutliers]
[1] 10  1

#observation 7 with salary 79.380*1000, but risk score is only 1

> y[xoutliers]
[1] 311 316
> ex.data[6:7,]
  income risk insurance
6 72.996   10       311
7 79.380    1       316

#dffits
> dffits(myfit)
          1           2           3           4           5           6 
-0.33449702 -0.30269393  1.18214133 -0.13693096 -0.05799765  0.74371046 
          7           8           9          10          11          12 
 3.52921562 -0.32626776 -0.22115900  0.02840685  0.14901409  0.78010158 
         13          14          15          16          17          18 
 0.04684108 -0.74231979 -0.15425222 -0.29338760  0.61289495 -0.01408022 
>
```

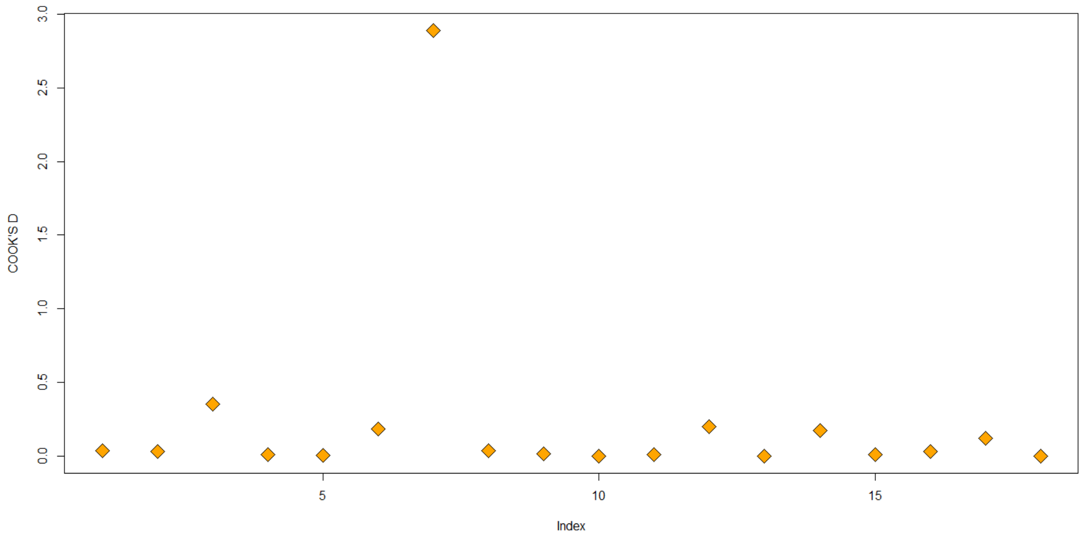


```
> ex.data[which(dffits(myfit) > 1),]
   income   risk  insurance
```

```
3  26.852    5       11
7  79.380     1       316
>
```

#It seems that observation No.7 had a high dffits 3.52921562>1.

```
>> cooks.distance(myfit)
           1            2            3            4            5 
3.608625e-02 3.091477e-02 3.494344e-01 6.637786e-03 1.197812e-03 
           6            7            8            9           10 
1.840268e-01 2.889475e+00 3.620589e-02 1.664223e-02 2.880476e-04 
          11           12           13           14           15 
7.839345e-03 1.972762e-01 7.824262e-04 1.713652e-01 8.363443e-03 
          16           17           18 
2.864581e-02 1.196986e-01 7.079365e-05 
> max(cooks.distance(myfit))
[1] 2.889475
> order(cooks.distance(myfit))[18]
[1] 7
> par(mfrow=c(1,1))
plot(cooks.distance(myfit))
```



```
➢
> smyfit<-summary(myfit)
> highcook <- which((cooks.distance(myfit)) > qf(0.5,smyfit$df[1],smyfit$df[2]))
> cooks.distance(myfit)[highcook]
       7 
2.889475 
>
>> dfbetas(myfit)
   (Intercept)           x1            x2
1  -0.11791502  0.124491661 -0.1107217037
2  -0.03945312 -0.146953233  0.1722774459
3   0.95935296 -0.987078887  0.1435731540
4   0.07701539 -0.082073331 -0.0410156333
5  -0.03935568  0.028583776  0.0010754435
```

```
6  -0.52978181  0.304838003  0.5125354924
7  -0.36492941  2.659822663 -2.6750533100
8   0.08157574  0.025440338 -0.2452456420
9   0.03078321 -0.067151914 -0.0365559869
10  0.02384654 -0.013764209 -0.0091627889
11  0.08634720 -0.105688246  0.0536400695
12 -0.58199873  0.449491490  0.4096139916
13  0.03482702 -0.029395861  0.0014469428
14 -0.27058334 -0.265611499  0.6268600751
15 -0.01637040  0.053207315 -0.0953091071
16 -0.18104226  0.025836093  0.1423819102
17  0.58027432 -0.360800840 -0.2577287527
18 -0.01010224  0.008033481 -0.0001311733
>
```

```
> influence.measures(myfit)
Influence measures of
         lm(formula = y ~ x1 + x2, data = ex.data) :

     dfb.1_   dfb.x1    dfb.x2     dffit   cov.r   cook.d     hat  inf
1  -0.1179   0.12449  -0.110722  -0.3345  0.973   3.61e-02  0.0693
2  -0.0395  -0.14695   0.172277  -0.3027  1.153   3.09e-02  0.1006
3   0.9594  -0.98708   0.143573   1.1821  0.521   3.49e-01  0.1890
4   0.0770  -0.08207  -0.041016  -0.1369  1.379   6.64e-03  0.1316
5  -0.0394   0.02858   0.001075  -0.0580  1.319   1.20e-03  0.0756
6  -0.5298   0.30484   0.512535   0.7437  1.530   1.84e-01  0.3499
7  -0.3649   2.65982  -2.675053   3.5292  0.893   2.89e+00  0.6225   *
8   0.0816   0.02544  -0.245246  -0.3263  1.224   3.62e-02  0.1319
9   0.0308  -0.06715  -0.036556  -0.2212  1.138   1.66e-02  0.0658
10  0.0238  -0.01376  -0.009163   0.0284  1.365   2.88e-04  0.1005
11  0.0863  -0.10569   0.053640   0.1490  1.350   7.84e-03  0.1201
12 -0.5820   0.44949   0.409614   0.7801  1.313   1.97e-01  0.2994
13  0.0348  -0.02940   0.001447   0.0468  1.352   7.82e-04  0.0944
14 -0.2706  -0.26561   0.626860  -0.7423  1.027   1.71e-01  0.2096
15 -0.0164   0.05321  -0.095309  -0.1543  1.297   8.36e-03  0.0957
16 -0.1810   0.02584   0.142382  -0.2934  1.079   2.86e-02  0.0775
17  0.5803  -0.36080  -0.257729   0.6129  1.068   1.20e-01  0.1818
18 -0.0101   0.00803  -0.000131  -0.0141  1.343   7.08e-05  0.0849
```

```
> # Evaluate Collinearity
> vif(myfit) # variance inflation factors
      x1       x2
1.069249 1.069249
>
```