Behaviour of the routines *fitlm* and *stepwiselm*

Batch de démonstration pour le cours de plans d'expériences. Le batch a pour but d'illustrer les méthodes reliées aux fonctions *fitlm* et *stepwiselm*. Cela se fait au travers de Analyse de mesures publiées par Hydro Quebec pour un brevet d'acier inoxydable austénitique au cobalt ultra résistant à la cavitation érosive.

Batch demonstrating the behaviour of the least square fit routines for the course of design of experiments. This is done analysing some public data on austenic steel. Austenitic cobalt stainless steel is ultra resistant to erosive cavitation

Table of Contents

Load of the data from an external file

The data is load in a table called Steel_tbl

```
Steel_tbl = readtable('dataDOE2019master.xlsx',...
    'Sheet','acier');
```

Here is the data:

```
disp(Steel_tbl)
```

C	Mn	Si	Cr	Co	Mo	N	Durete	Erosion
0.46	5.3	3.3	16.6	11.6	0.02	0.05	29	0.78
1.1	1	0.5	28	60	0	0.05	40	0.8
0.41	5	3.5	16.4	8	0	0.042	26	0.86
0.4	5	3	16	8	0	0.02	30	0.9
0.78	0.1	1.4	18.4	8	0	0.09	34	1
0.9	0.1	1.1	26	8	0	0.01	33	1.06
0.45	4.9	1.7	18.1	10	0.02	0.05	26	1.06
0.6	2	1	18	10	0	0.02	27	1.06
0.39	5.2	3.5	15.8	10	0.02	0.05	23	1.08
1.1	0.1	1.3	25	8	0	0.017	31	1.11
0.46	4.9	3.1	16.9	11.6	0.02	0.05	26	1.11
0.27	10.2	1.36	12.9	10.4	2.6	0.078	26	1.12
0.5	2	1	18	10	0	0.02	26	1.12
0.39	4.4	1.6	18.6	9.1	0.02	0.05	25	1.17
0.37	5	0.76	19	11	0.02	0.05	26	1.18
0.34	11.4	1.25	13.5	9.9	2.5	0.05	23	1.26
0.26	10.26	1.1	13.2	11.6	0	0.02	12	1.3
0.4	2	1	18	10	0	0.02	26	1.32
0.41	5.5	0	15.1	12	0	0.04	21	1.34
0.26	10.2	1.43	13	9.7	2.7	0.047	21	1.35
0.27	10.4	1.16	12.5	7.3	0	0.02	15	1.36
0.31	10.7	1.23	14.2	9	2.5	0.05	26	1.36
0.27	10.2	1.11	13.5	9.96	0	0.02	15	1.36
0.4	2	3	16	8	0	0.02	30	1.37
0.11	15.8	0.55	12.8	9.1	3.3	0.057	26	1.4
0.32	10.9	1.18	14.5	9.3	2.4	0.05	26	1.4
0.34	3.31	0.99	18.5	11.6	0.02	0.05	26	1.4
0.25	1	0.5	28	60	5	0.05	30	1.4

Analysis of the hardness (dureté)

All possible factors

Linear regression of the hardness of the steel in function of the content in cupper (C), cobalt (Co), Manganese (Mn), Scilicium (Si), Chromium (Cr), Molibdene (Mo) and Nitrogen (N)

```
mdl_hardness_1 = fitlm(Steel_tbl,'Durete~C+Mn+Si+Cr+Co+Mo+N');
disp(mdl_hardness_1)
```

```
Linear regression model:
    Durete ~ 1 + C + Mn + Si + Cr + Co + Mo + N

Estimated Coefficients:
                   Estimate       SE        tStat       pValue  
                   _________    ________    _______    ________

    (Intercept)      11.806      8.4109      1.4037     0.17576
    C                 12.32      4.9058      2.5113    0.020727
    Mn             -0.40825     0.34226     -1.1928     0.24691
    Si               1.2178     0.73573      1.6553     0.11348
    Cr              0.28777     0.45302     0.63522     0.53249
    Co            -0.019538    0.080835     -0.2417     0.81147
    Mo               1.4938     0.77759       1.921    0.069097
    N                77.839       37.96      2.0505    0.053647
```

```
Number of observations: 28, Error degrees of freedom: 20
Root Mean Squared Error: 3.31
R-squared: 0.756,  Adjusted R-Squared: 0.671
F-statistic vs. constant model: 8.87, p-value = 5.62e-05
```

ANALYSIS

1. The linear coefficients for Mn (p=24%) Si (p=11%), Cr (p=53%), Co (p=81%) are clearly over the usual limit of 5%
2. The Mo is close of the limit with $p \approx 7\%$
3. Equivalently the corresponding values of *tStat* are lower than the criterion *t(5%/2,N-P=20)=2.08*
4. The F statistic of 8.87 is higher than the criterion F(5%,1,N-P=20)=4.35 indicating that the model is globaly acceptable
5. It makes sens to test a simpler model without the questionable terms

---

Use of ANOVA

First try with the routine with default parameters:

```
anova(mdl_hardness_1)
```

ans = 8×5 table

	SumSq	DF	MeanSq	F	pValue
1 C	69.0703	1	69.0703	6.3067	0.0207
2 Mn	15.5822	1	15.5822	1.4228	0.2469
3 Si	30.0078	1	30.0078	2.7399	0.1135
4 Cr	4.4191	1	4.4191	0.4035	0.5325
5 Co	0.6398	1	0.6398	0.0584	0.8115
6 Mo	40.4169	1	40.4169	3.6904	0.0691
7 N	46.0500	1	46.0500	4.2047	0.0536
8 Error	219.0395	20	10.9520	1.0000	0.5000

1. The results are equivalent to the fitlm() output, exept that the test is now based on the Fisher distribution.The linear coefficients for Mn
2. (p=24%) Si (p=11%), Cr (p=53%), Co (p=81%) and Mo (p=7%) are over the limit of 5%
3. The p-values are slightly different, but the figure is strickly the same

In the next step, the parameter of the routine let us precise that we want all the component and the sum of squares of type I (sequential)

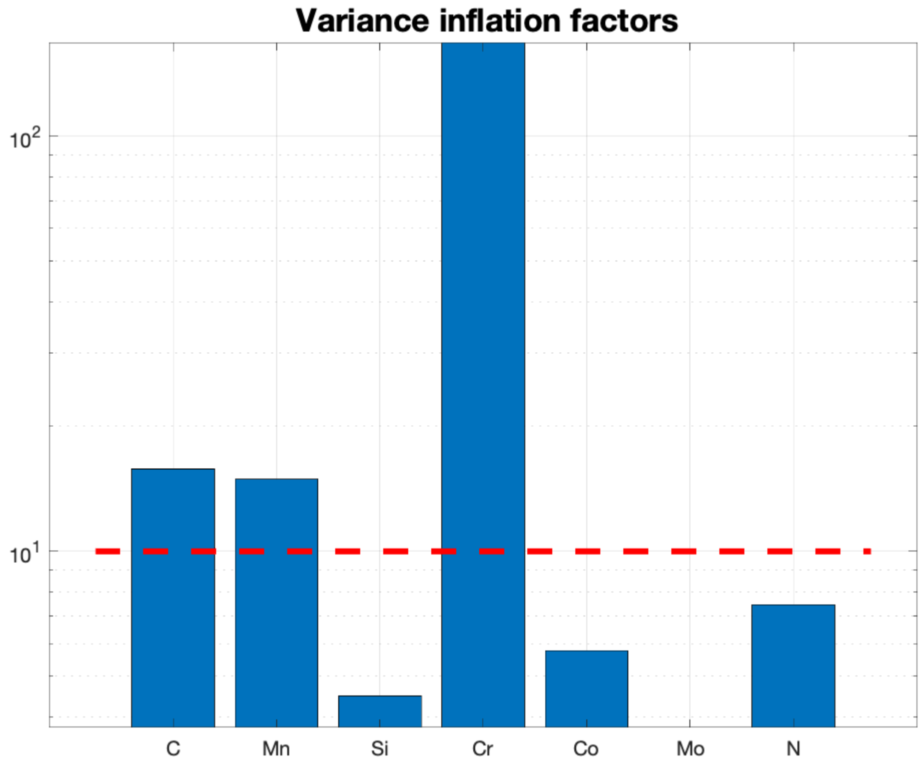
```
anova(mdl_hardness_1,'component',1)
```

ans = 8×5 table

	SumSq	DF	MeanSq	F	pValue
1 C	407.0547	1	407.0547	37.1672	0.0000
2 Mn	44.6765	1	44.6765	4.0793	0.0570
3 Si	3.7614	1	3.7614	0.3434	0.5644
4 Cr	76.7341	1	76.7341	7.0064	0.0155
5 Co	5.4695	1	5.4695	0.4994	0.4879
6 Mo	95.8928	1	95.8928	8.7558	0.0078
7 N	46.0500	1	46.0500	4.2047	0.0536
8 Error	219.0395	20	10.9520	1.0000	0.5000

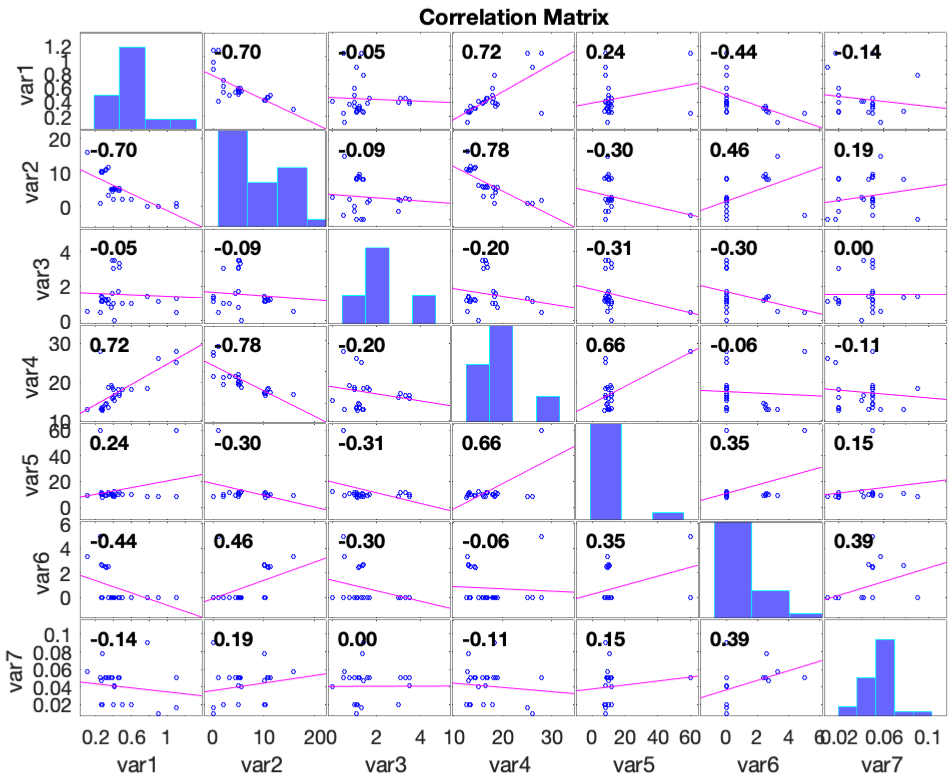
In this case, we observe strong differences in the sums of squares and the p-values. If we compute the VIF, it is evident that the data is not orthogonal with several VIF coefficients that are higher than 10. This is a situation that requires SS of type II due to the non orthoganality of the regressors.

```
X=x2fx(Steel_tbl{:,1:7});
D=inv(X'*X);
C=corrcov(D);
VIF=diag(inv(C));
bar(VIF(2:8))
xticklabels({'C','Mn','Si','Cr','Co','Mo','N'})
set(gca,"YScale","log")
grid on
title(' Variance inflation factors',"FontSize",16)
hold on
plot([0.25,7.75],[10,10],'r--',"LineWidth",3)
hold off
```



The correlation plot unfortunately does not help to see precisely why the estimation of the coefficient for the molibden is sensitive to the type of sum of squares. Nevertheless it is evident that the data is not sufficient to make a safe estimation of the linear effects of the different elements.

```
corrplot(Steel_tbl{:,1:7})
```



The anova computed with the parameter *'summary'* confirm that the model is globaly ok with a p-value of 56 ppm

```
anova(mdl_hardness_1,'summary')
```

ans = 3×5 table

	SumSq	DF	MeanSq	F	pValue
1 Total	898.6786	27	33.2844	NaN	NaN
2 Model	679.6391	7	97.0913	8.8652	5.6178e-05
3 Residual	219.0395	20	10.9520	NaN	NaN

Use of coefCI()

Let us now compute the confidence intervals of the coefficients with the routine *coefCI()*

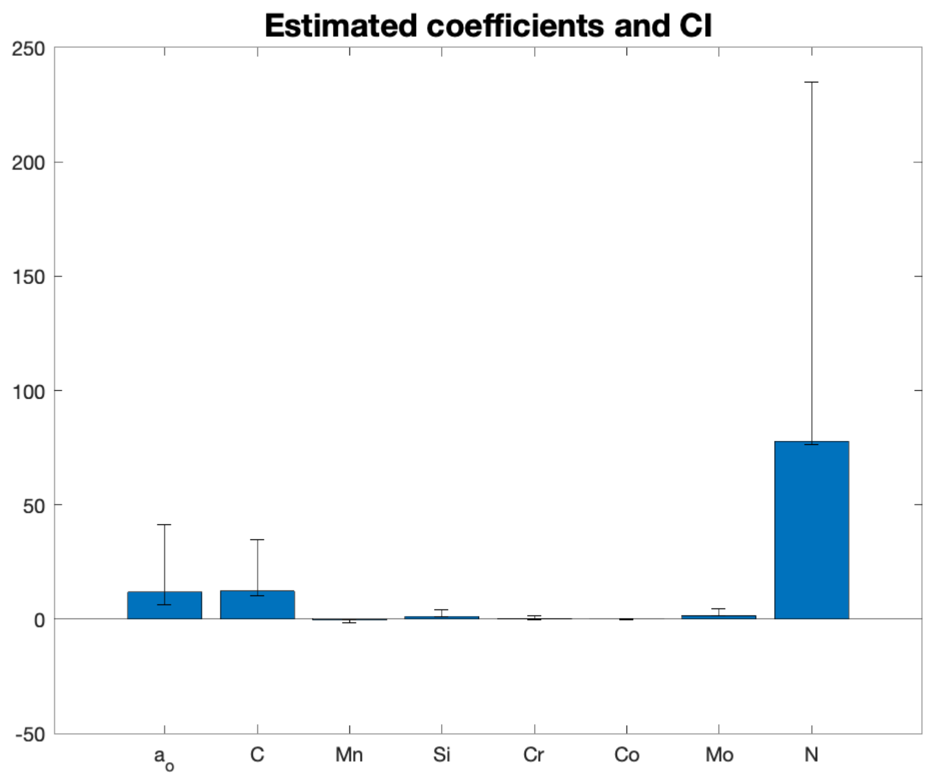
```
CI_1=coefCI(mdl_hardness_1);
disp(array2table(CI_1))
```

```
     CI_11       CI_12
    ________    _______

     -5.7389     29.351
      2.0867     22.553
     -1.1222     0.3057
    -0.31687     2.7525
    -0.65722     1.2328
    -0.18816    0.14908
```

```
    -0.12825      3.1158
     -1.3447      157.02
```

Let us produce a graphical representation of this data:

```
coef=mdl_hardness_1.Coefficients.Estimate;
bar(1:8,coef)
title("Estimated coefficients and CI","FontSize",16)
xticklabels({'a_o','C','Mn','Si','Cr','Co','Mo','N'})
hold on
H=errorbar(1:8,coef,CI_1(:,1),CI_1(:,2));
H.Color=[0,0,0];
H.LineStyle="none";
hold off
```



The bad quality of the estimate is evident in this bar chart

Use of coefTest()

This routine gives the possibility to test the values of the coefficients. We must be carefull between the difference of the ANOVA results giving the probability to have a null model and the test of the values of the coefficient.

The routine is not straight forward... be care full to what is really computed

```
H=eye(8);
[p,F,r]=coefTest(mdl_hardness_1,H)
```

```
p = 8.1296e-18
F = 222.0143
r = 8
```

```
H=eye(8);
[p,F,r]=coefTest(mdl_hardness_1,H,coef)
```

```
p = 1
F = 0
r = 8
```

```
H=eye(8);
C=[11.8,12,0,0,0,0,1.5,78];
[p,F,r]=coefTest(mdl_hardness_1,H,C)
```

```
p = 3.2720e-05
F = 9.0796
r = 8
```

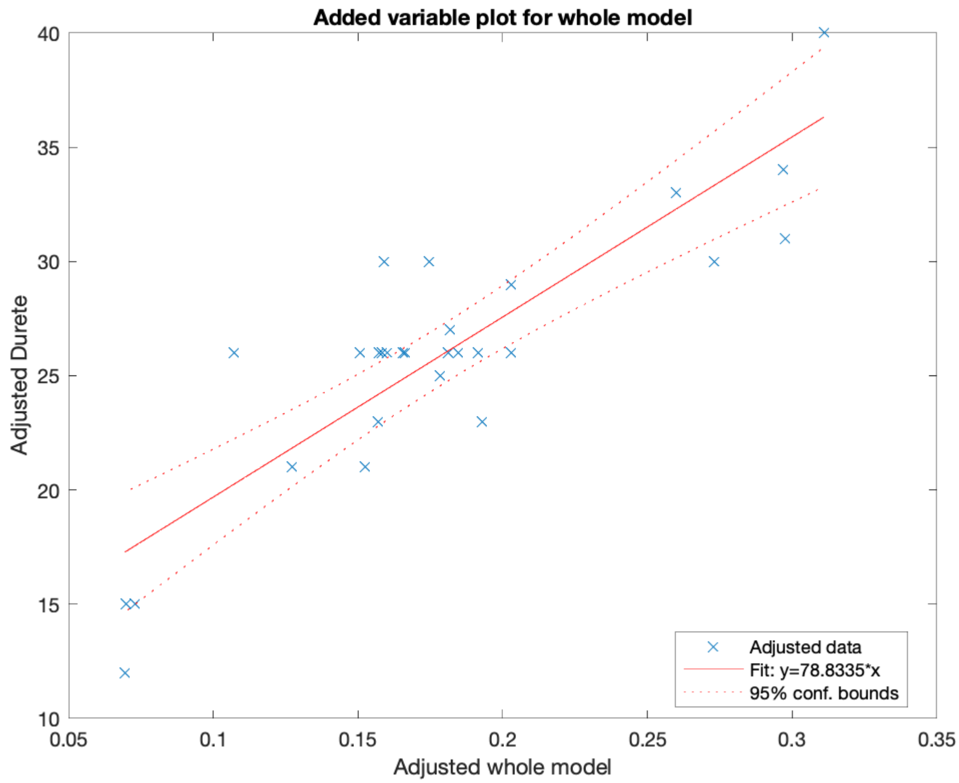
```
H=[1 0 0 0 0 0 0 0
   0 1 0 0 0 0 0 0
   0 0 0 0 0 0 1 0
   0 0 0 0 0 0 0 1];

C=[12,12,1.5,80];
[p,F,r]=coefTest(mdl_hardness_1,H,C)
```

```
p = 1.0000
F = 0.0031
r = 4
```

Use of plot(mdl)

This function is a version of "model vs data" plot. It compares the mdel without the constant with the data
It let us observe if both are coherent. Is is very basic. A sophistication is possible with the next routine *plotadded(mdl)* which will let us to select the regressors.

```
plot(mdl_hardness_1)
```

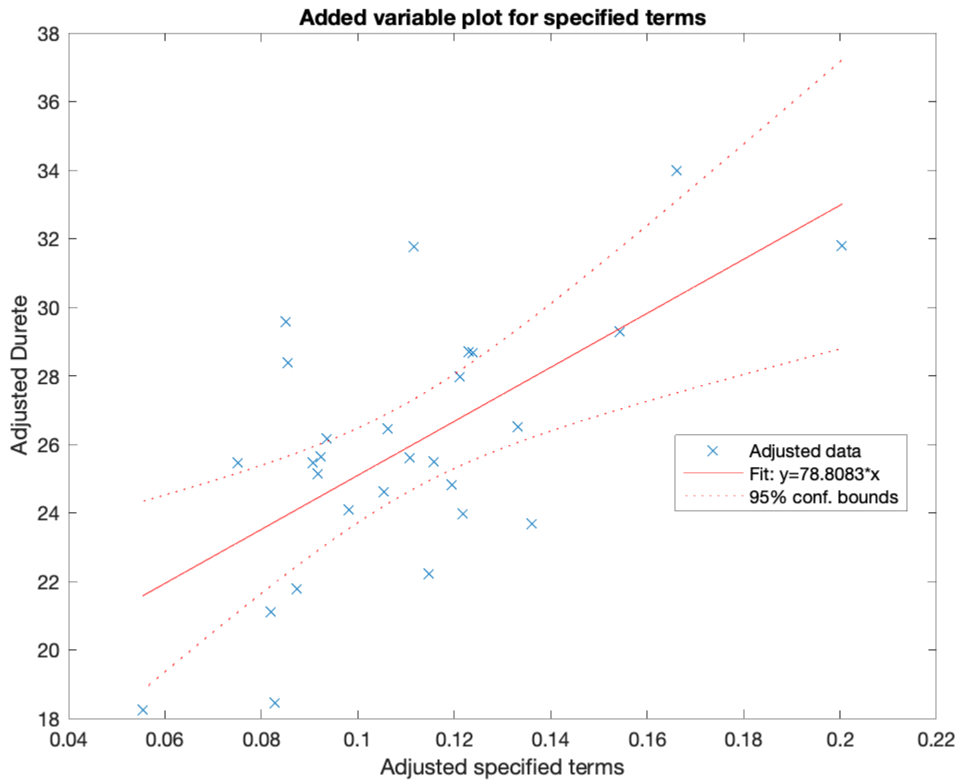


Use of plotAdded(mdl, coef)

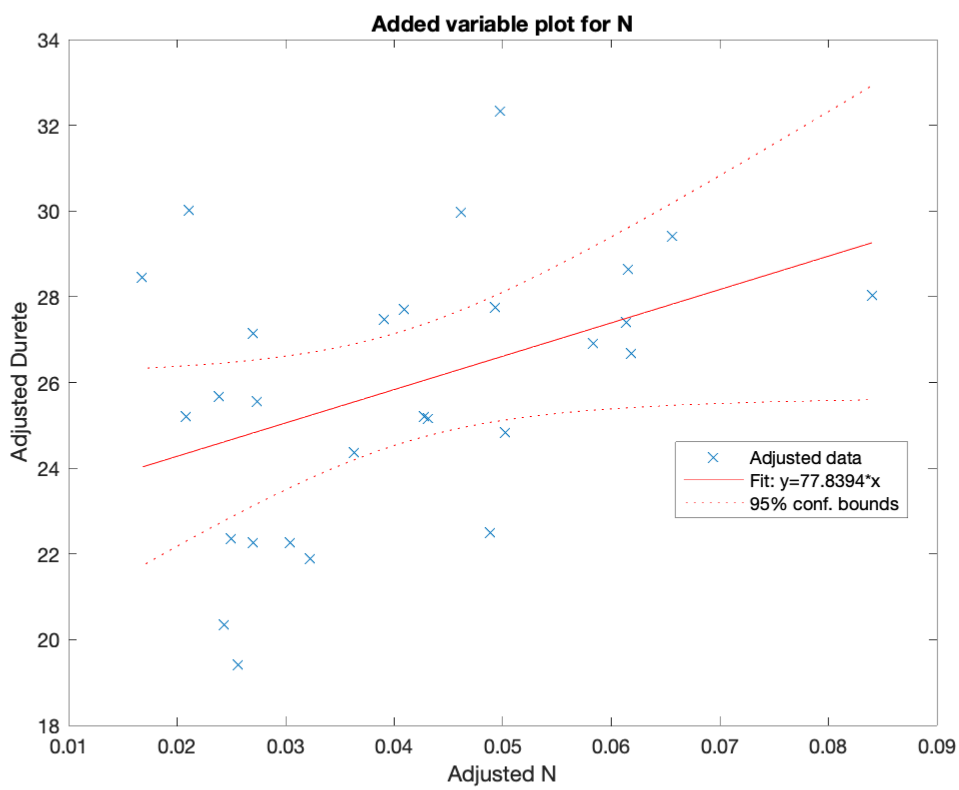
Without specification this routine is equivalent to plot(mdl). But it value resid in the possibility to include only some of the regressors.

In the example below, only the second regressor, C, and the seventh, N have been considered and is is possible to observe that the figure is quite equivalent to the whole model and showing that there is an important part of the data that is not taken in consideration by the model (the points outside the confidence hyperbole).

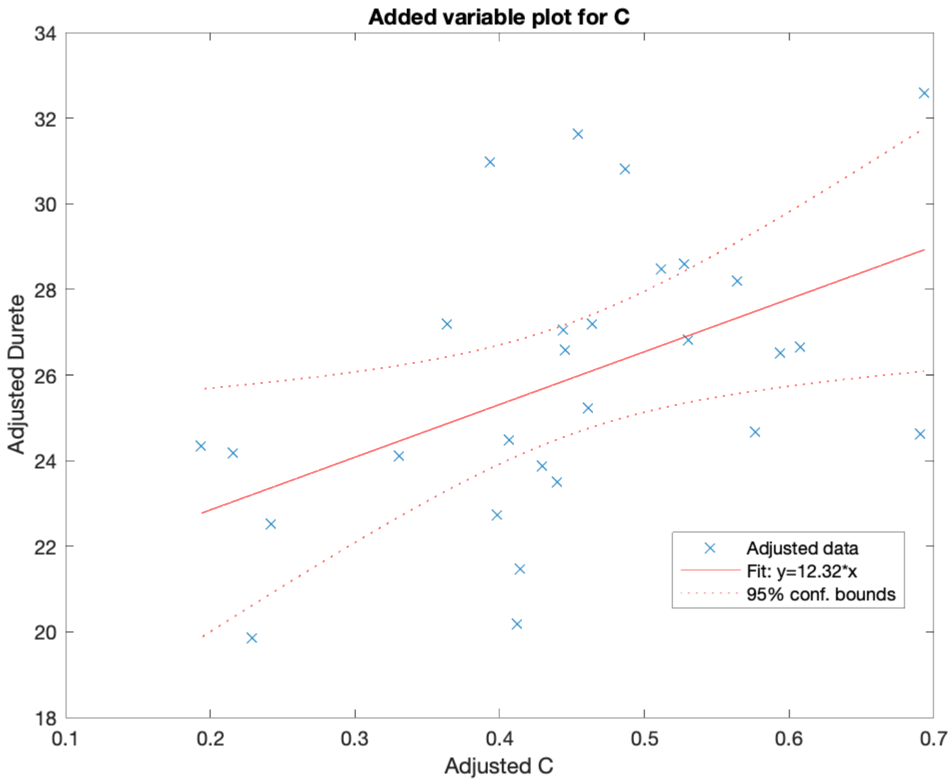
```
plotAdded(mdl_hardness_1,[2,8])
```




```
plotAdded(mdl_hardness_1,[8])
```



```
plotAdded(mdl_hardness_1,[2])
```



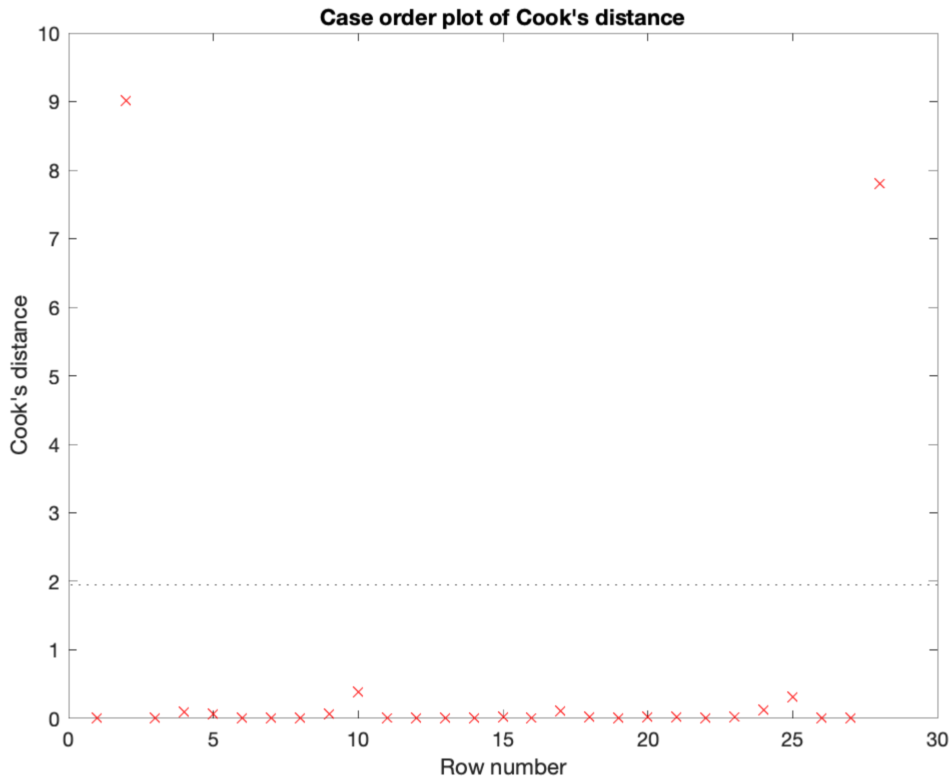
Use of plotDiagnostics

The routine plotDiagnostics(mdl,type) gives access to several graphics putting in evidence points whose leverage on the data is important and that would merit a more intensif scrutiny to be sure that the conclusions are not driven by erroneous data. Cook's distance is the scaled change in fitted values, which is useful for identifying outliers in the X values (observations for predictor variables). Cook's distance shows the influence of each observation on the fitted response values. An observation with Cook's distance larger than three times the mean Cook's distance might be an outlier. We see in the first graphic that two points (2, 28) have effectively a Cook's distance which is higher than 2 (the Cook's distance is 0.65), in fact 8 and 9

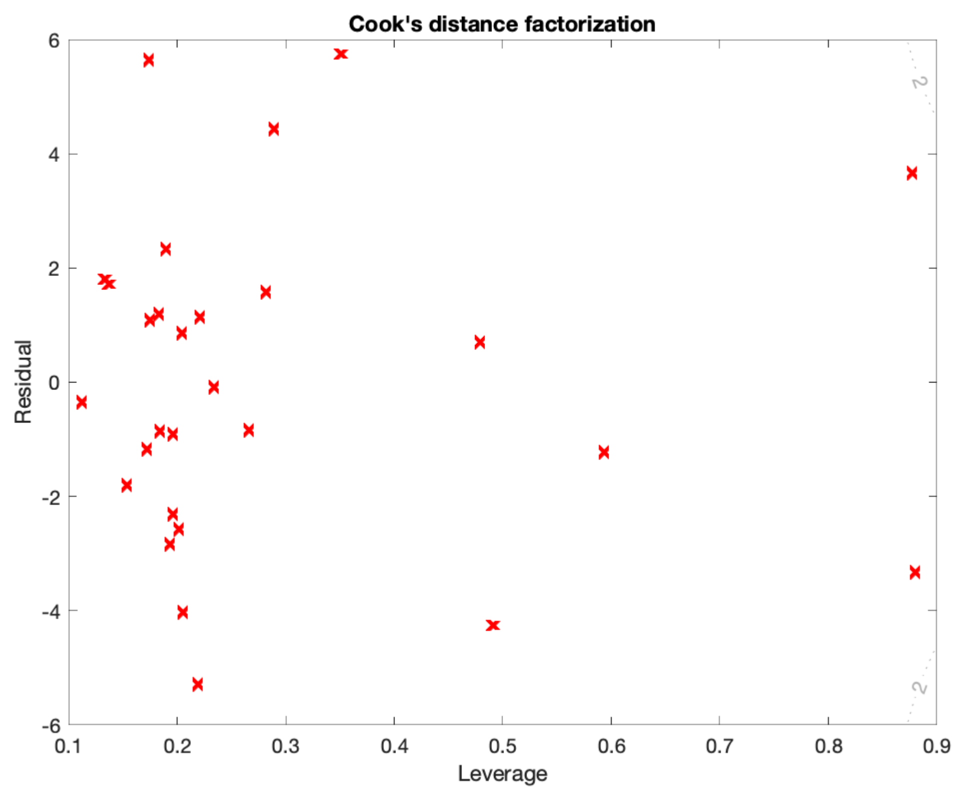
```
M=mean(mdl_hardness_1.Diagnostics.CooksDistance)
```

```
M = 0.6479
```

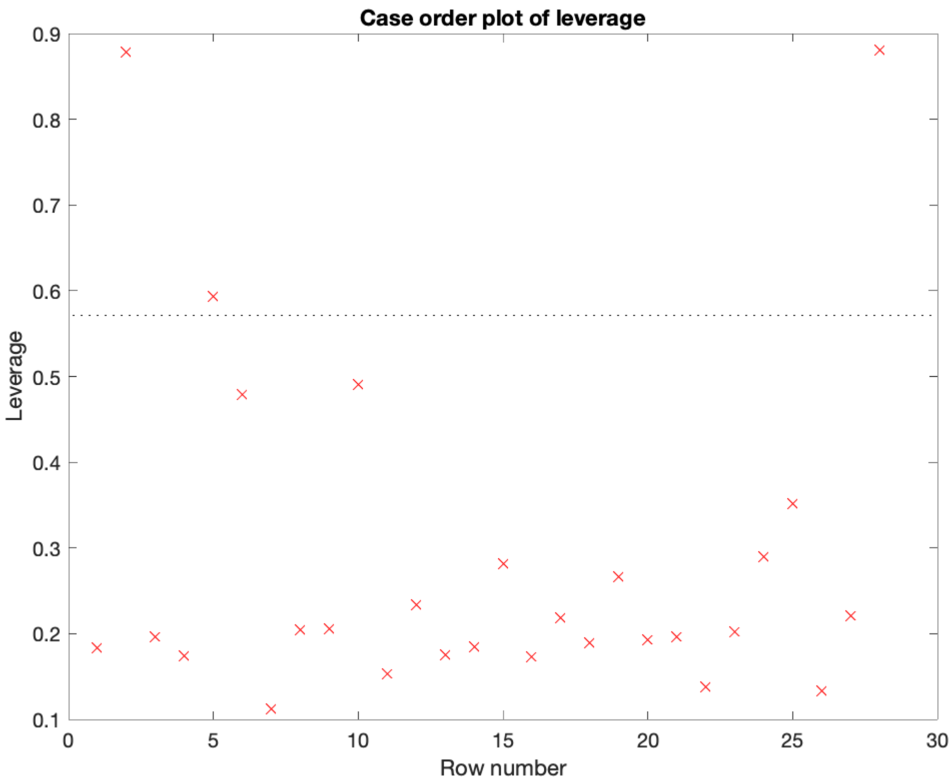
```
plotDiagnostics(mdl_hardness_1,"cookd")
```




```
plotDiagnostics(mdl_hardness_1,"contour")
```



```
plotDiagnostics(mdl_hardness_1,"leverage")
```



Only the significant factors of the first round

Inference of a linear model of the hardness based on factors that show p-values close or smaller than 5%

```
mdl_durete_2=fitlm(Steel_tbl,'Durete~C+Mn+Mo+N');
disp(mdl_durete_2)
```

```
Linear regression model:
    Durete ~ 1 + C + Mn + Mo + N

Estimated Coefficients:
                   Estimate      SE        tStat       pValue
                   ________    _______    _______    __________

    (Intercept)      19.337     2.9542     6.5458    1.1153e-06
    C                12.834     3.7373     3.4342     0.0022628
    Mn             -0.62282    0.21438    -2.9053     0.0079736
    Mo               1.4431    0.55807     2.5858      0.016526
    N                80.022     35.742     2.2389      0.035129


Number of observations: 28, Error degrees of freedom: 23
Root Mean Squared Error: 3.3
```

```
R-squared: 0.721,  Adjusted R-Squared: 0.672
F-statistic vs. constant model: 14.9, p-value = 3.93e-06
```

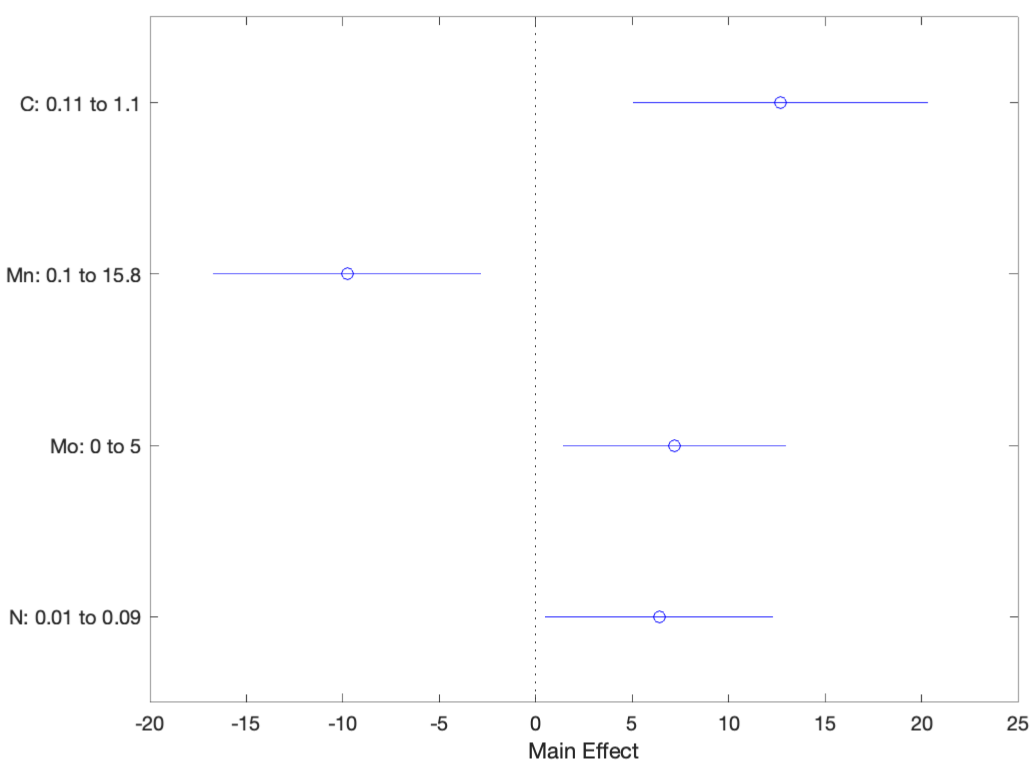
ANALYSIS

1. The remaining four linear coefficients now have p-values ��below the usual 5% limit
2. The pValue of the constant has also improved significantly
3. The values ��of tStat are all greater than the criterion t(5%/2,N-P=23)=2.07
4. A model based on the concentration of C, Mn, Mo and N seems acceptable: a p_value for the model of about $4\ 10^{-6}$
5. At this stage it is interesting to look at the diagnostics

---

Plot of the effects with their confidence intervals

```
figure
plotEffects(mdl_durete_2)
```



List of confidence intervals

```
CI=table(coefCI(mdl_durete_2),...
    'VariableNames',{'CI'},...
    'RowNames',{'const','C','Mn','Mo','N'});
```

```
disp(CI)
```

```
                 CI
         __________________

 const    13.226     25.449
 C        5.1034     20.566
 Mn      -1.0663   -0.17935
 Mo      0.28861     2.5975
 N        6.0847     153.96
```

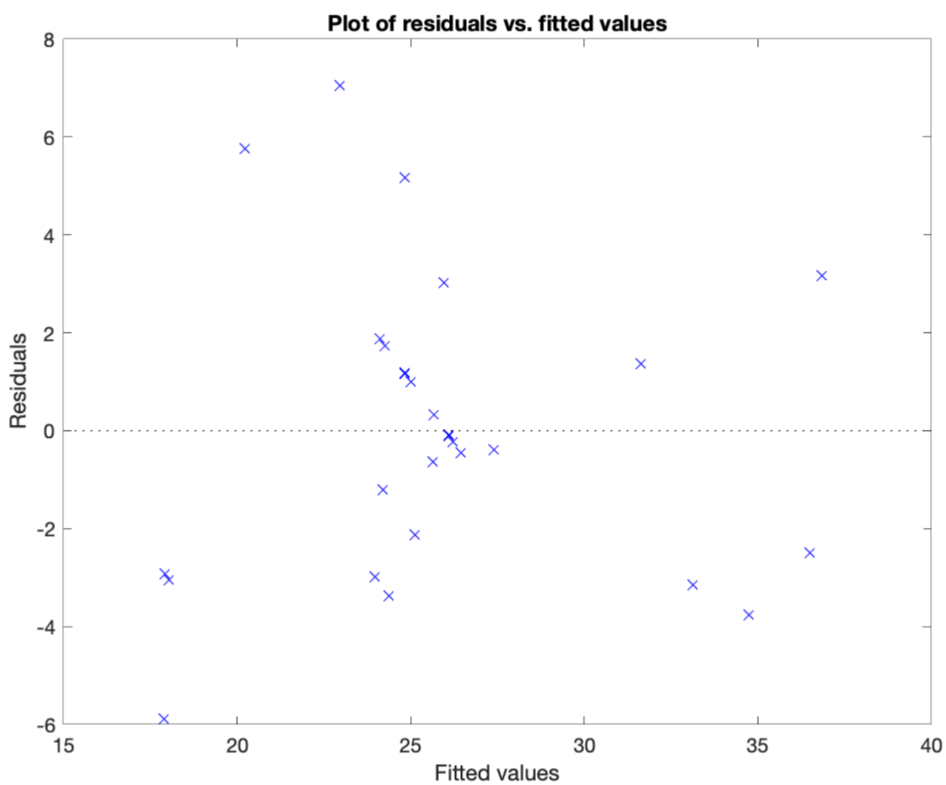
ANALYSE

1. The four linear coefficients have relatively wide confidence intervals with respect to their value
2. The prediction capacity of the model is therefore very limited.
3. At this stage it is interesting to look at the diagnostics

---

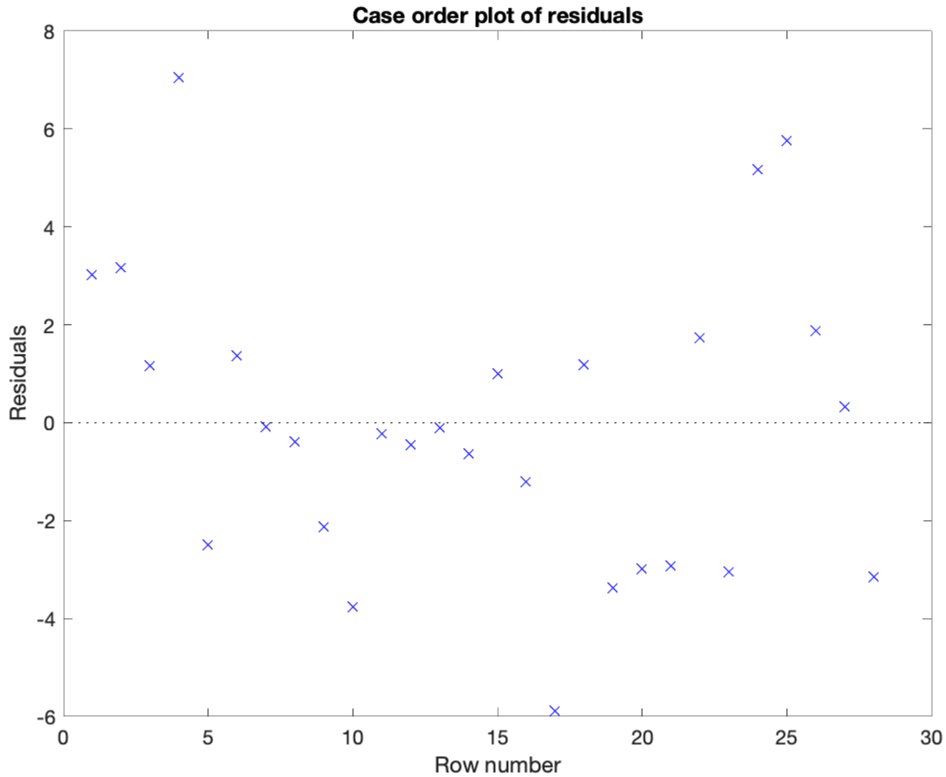
Plot of residues vs model

```
figure
plotResiduals(mdl_durete_2,'fitted')
```



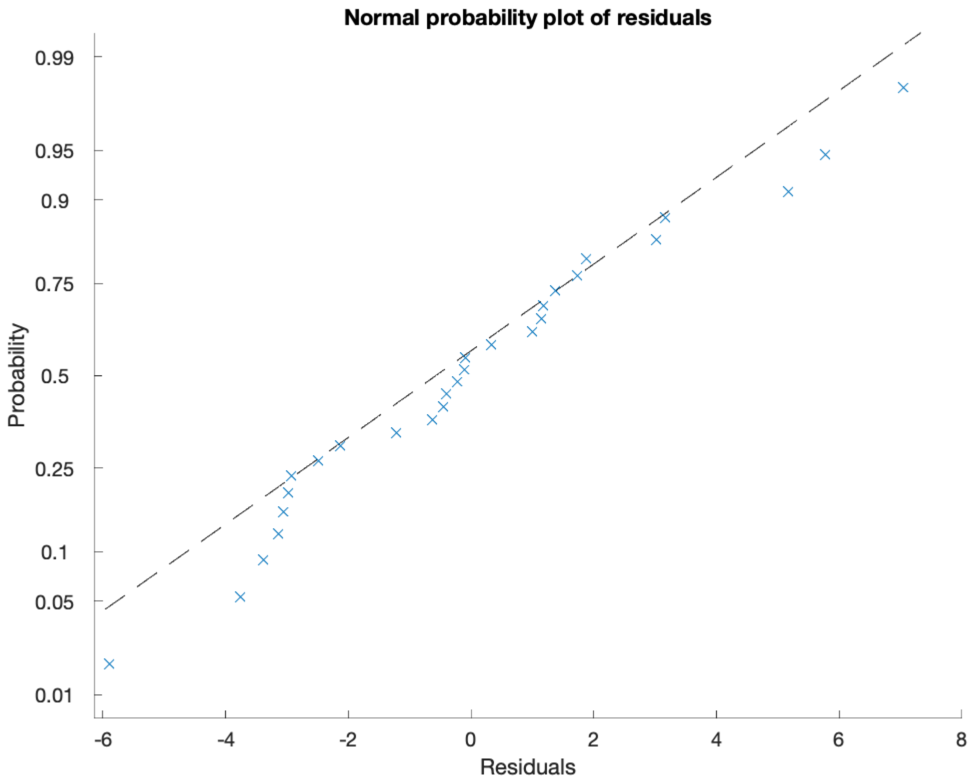
Plot of the residues by order of experiments

```
figure
plotResiduals(mdl_durete_2,'caseorder')
```



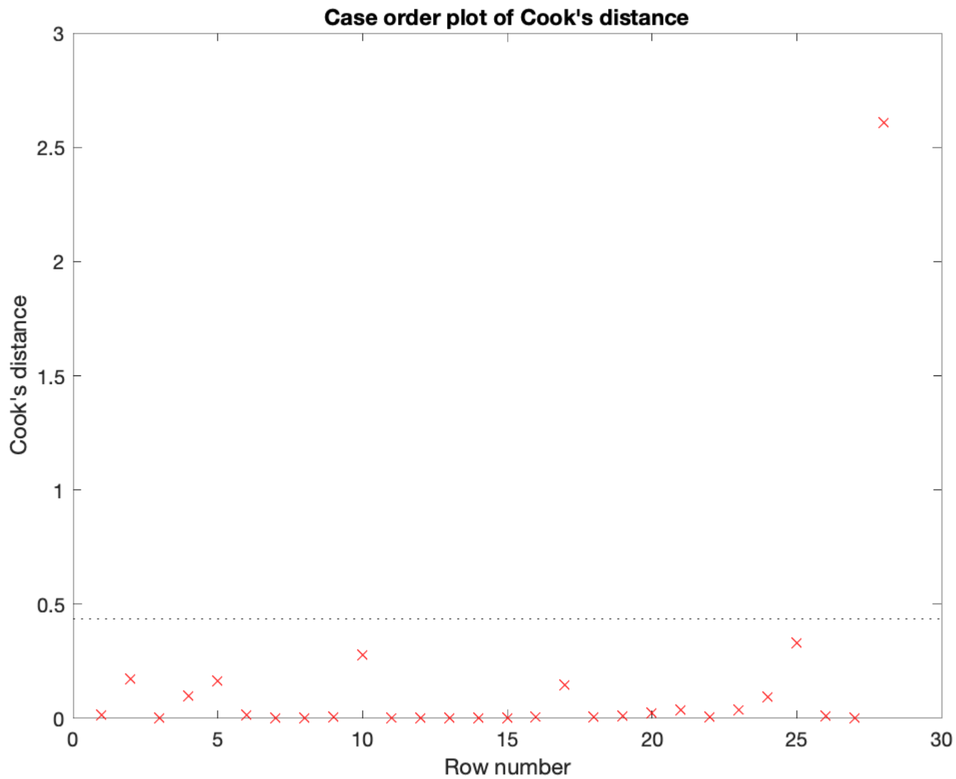
NormalPlot of residuals

```
figure
plotResiduals(mdl_durete_2,'probability')
```



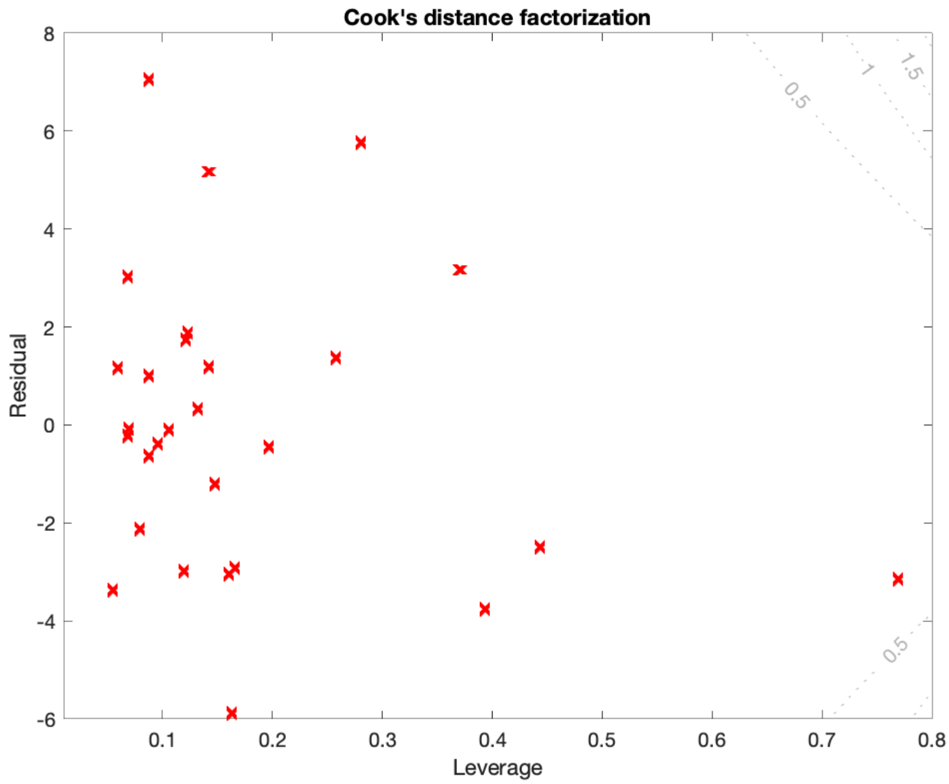
Cook distances

```
figure
plotDiagnostics(mdl_durete_2,'cookd')
```



Cook's distance vs residuals

```
figure
plotDiagnostics(mdl_durete_2,'contour')
```



ANALYSIS

1. No specific pattern in residual analysis
2. Case 28 (the last) at a distance from Cook large enough to suggest that it could be an outlier
3. On the other hand, the contourplot does not seem to confirm this hypothesis.
4. Let's see what happens when we remove this point

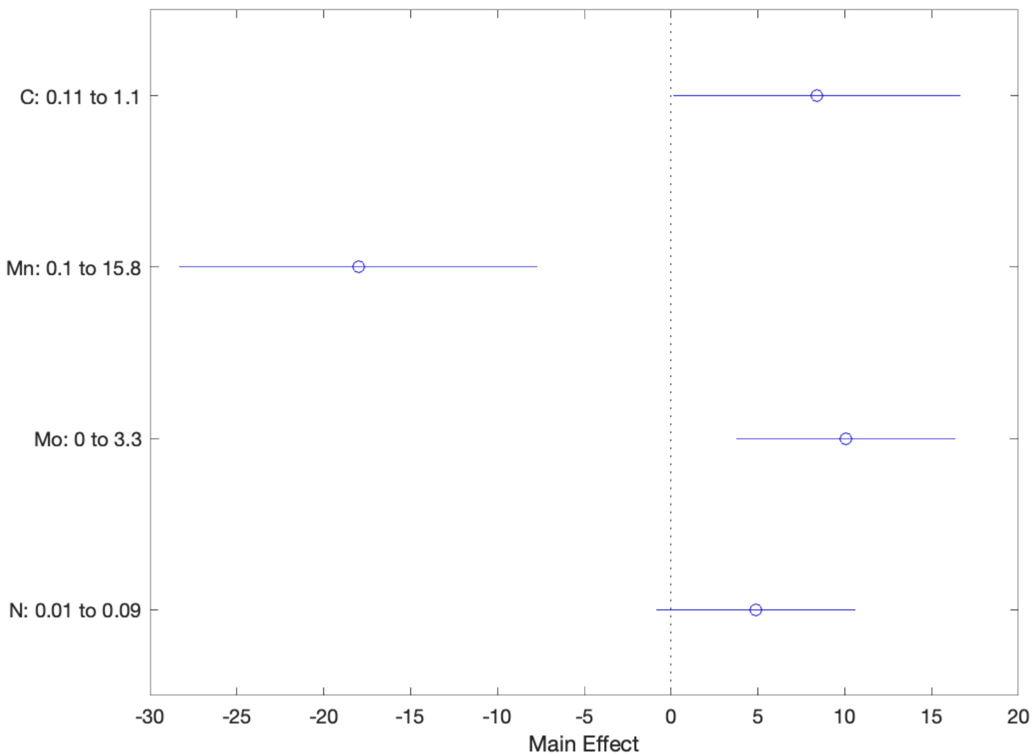
Without the oulier

```
[~,larg] = max(mdl_durete_2.Diagnostics.CooksDistance);

mdl_durete_3=fitlm(Steel_tbl,'Durete~C+Mn+Mo+N','Exclude',larg);
disp(mdl_durete_3)
```

```
Linear regression model:
    Durete ~ 1 + C + Mn + Mo + N

Estimated Coefficients:
                   Estimate      SE        tStat       pValue  
                   ________    ______     ______    __________

    (Intercept)     24.278     3.6026      6.739    8.9857e-07
    C               8.5055     4.0334     2.1088      0.046579
```

```
    Mn              -1.1471     0.31745    -3.6135     0.0015404
    Mo                3.055     0.91973     3.3216     0.0030987
    N                61.344      34.431     1.7817      0.088617


Number of observations: 27, Error degrees of freedom: 22
Root Mean Squared Error: 3.08
R-squared: 0.764,  Adjusted R-Squared: 0.721
F-statistic vs. constant model: 17.8, p-value = 1.2e-06
```

```
figure
plotEffects(mdl_durete_3)
```



confidence intervals

```
CI3=table(coefCI(mdl_durete_3),...
    'VariableNames',{'CI'},...
    'RowNames',{'const','C','Mn','Mo','N'});
disp(CI3)
```

```
                     CI
              ___________________

    const     16.806       31.749
    C         0.1407        16.87
    Mn       -1.8055     -0.48877
    Mo        1.1476       4.9624
    N        -10.061       132.75
```

ANALYSIS

1. It does not significantly improve the model

Step wise regression

```
mdl_durete_4 = stepwiselm(Steel_tbl,'Durete~C+Mn+Si+Cr+Co+Mo+N',...
    'PredictorVars',{'C','Mn','Si','Cr','Co','Mo','N'})
```

```
1. Adding Mn:Mo, FStat = 27.0714, pValue = 5.06193e-05
2. Adding Si:N, FStat = 9.3762, pValue = 0.0067135
3. Removing C, FStat = 1.2395, pValue = 0.28021
4. Removing Co, FStat = 2.3733, pValue = 0.13992
mdl_durete_4 =
Linear regression model:
    Durete ~ 1 + Cr + Mn*Mo + Si*N

Estimated Coefficients:
```

	Estimate	SE	tStat	pValue
(Intercept)	4.3942	4.7718	0.92086	0.3681
Mn	−0.89919	0.2152	−4.1783	0.00046362
Si	5.7279	1.3539	4.2306	0.00041031
Cr	0.81268	0.16808	4.835	0.00010054
Mo	−1.8265	0.47238	−3.8666	0.00096095
N	223.5	52.966	4.2197	0.00042091
Mn:Mo	0.38929	0.060135	6.4736	2.5978e−06
Si:N	−107.16	31.949	−3.3541	0.0031584

```
Number of observations: 28, Error degrees of freedom: 20
Root Mean Squared Error: 1.89
R-squared: 0.921,  Adjusted R-Squared: 0.893
F-statistic vs. constant model: 33.1, p-value = 1.18e-09
```

ANALYSIS

- The method quickly leads to a better result by integrating interaction term