Forestry Example

CZ

Fall, 2014

Forestry Example

The objective of the experiment was to predict total needle area (TNA) of a seedling based on

- trunk size: caliper (CAL)
- height (HT)
- product (interaction) of CAL and HT (HTCAL)

Read Data

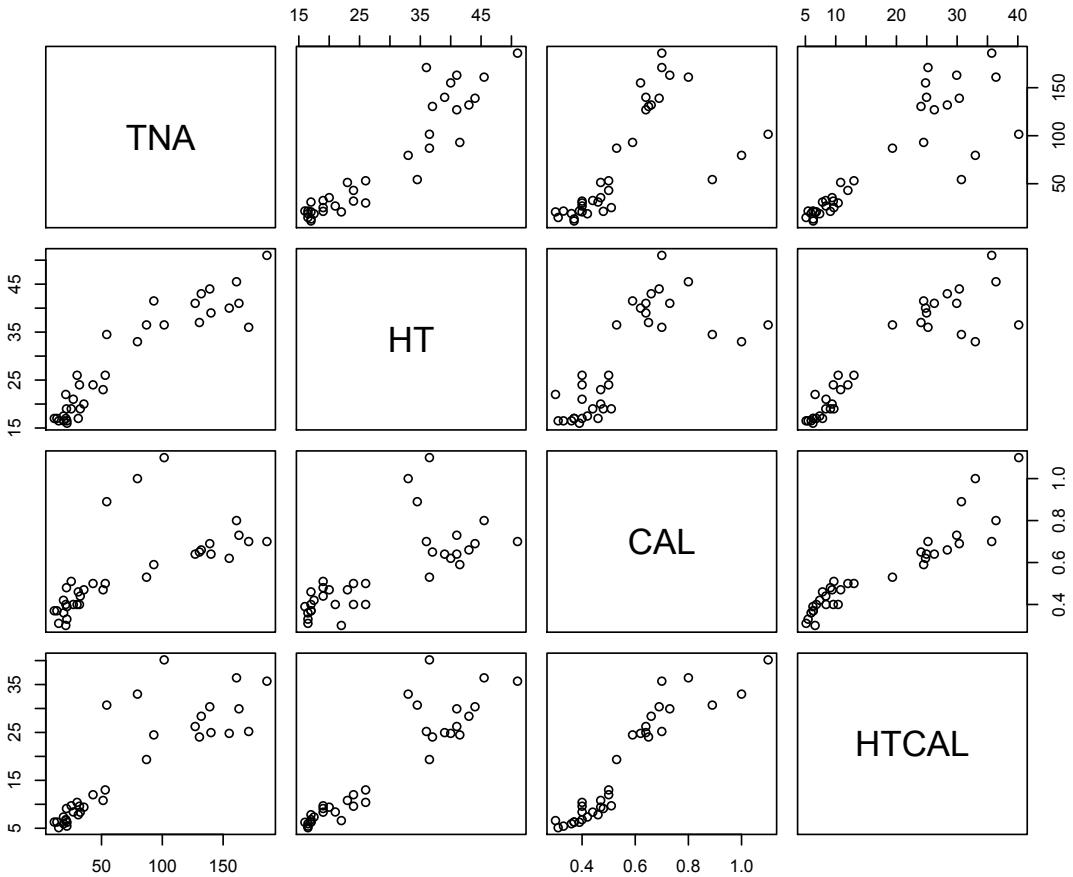
```
> dat <- read.table("forestry.txt", header=TRUE)
> head(dat)
     TNA   HT CAL HTCAL
1 101.51 36.5 1.1 40.15
2  79.54 33.0 1.0 33.00
3  20.62 22.0 0.3  6.60
4  53.07 26.0 0.5 13.00
5  43.02 24.0 0.5 12.00
6  31.88 24.0 0.4  9.60
> dim(dat)
[1] 35  4
```

Look at data

```
> summary(dat)
      TNA               HT              CAL              HTCAL
 Min.   : 11.18   Min.   :16.00   Min.   :0.3000   Min.   : 5.115
 1st Qu.: 21.55   1st Qu.:18.25   1st Qu.:0.4000   1st Qu.: 7.585
 Median : 43.02   Median :24.00   Median :0.5000   Median :10.810
 Mean   : 70.35   Mean   :28.39   Mean   :0.5491   Mean   :17.096
 3rd Qu.:128.75   3rd Qu.:38.00   3rd Qu.:0.6550   3rd Qu.:25.720
 Max.   :186.00   Max.   :51.00   Max.   :1.1000   Max.   :40.150
```

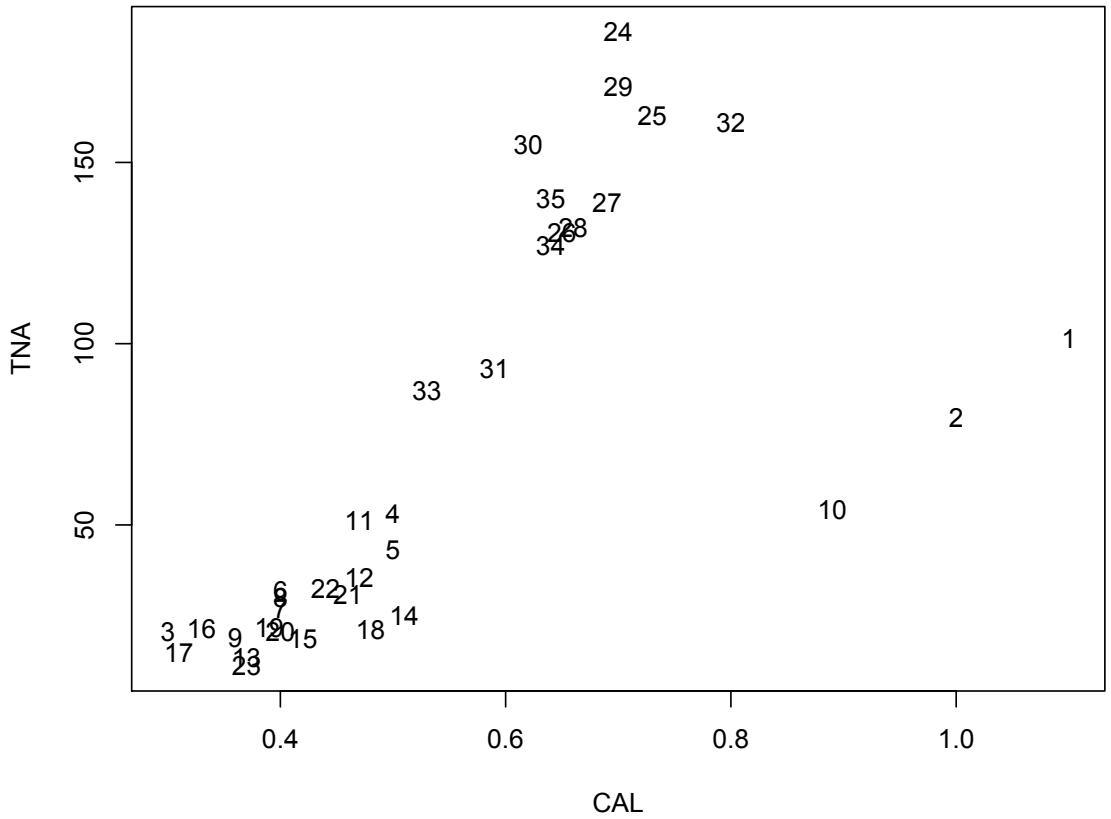
Look at data

```
> pairs(dat)
```



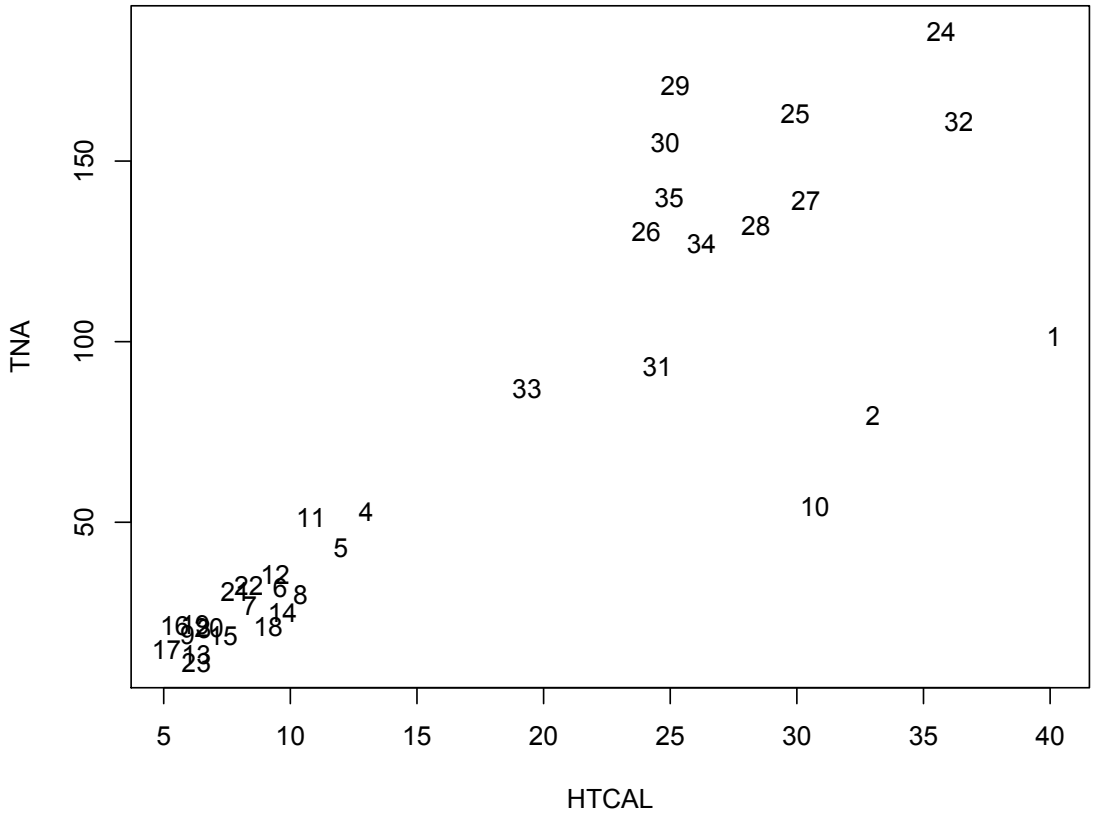
Scatterplot

```
> plot(dat$CAL, dat$TNA, type="n", xlab="CAL", ylab="TNA")
> text(dat$CAL, dat$TNA)
```



Scatterplot

```
> plot(dat$HTCAL, dat$TNA, type="n", xlab="HTCAL", ylab="TNA")
> text(dat$HTCAL, dat$TNA)
```



Data Checking

Look at cases 1, 2 and 10

```
> dat[c(1,2,10),]
       TNA   HT  CAL  HTCAL
1   101.51 36.5 1.10 40.150
2    79.54 33.0 1.00 33.000
10   54.30 34.5 0.89 30.705
```

- It was found they are recording errors
- We will remove them for subsequent analyses

```
>dat2=dat[-c(1,2,10),]
```

Model Fitting

Consider the model:

$$TNA = \beta_0 + \beta_1 CAL + \beta_2 HT + \beta_3 HTCAL + \epsilon$$

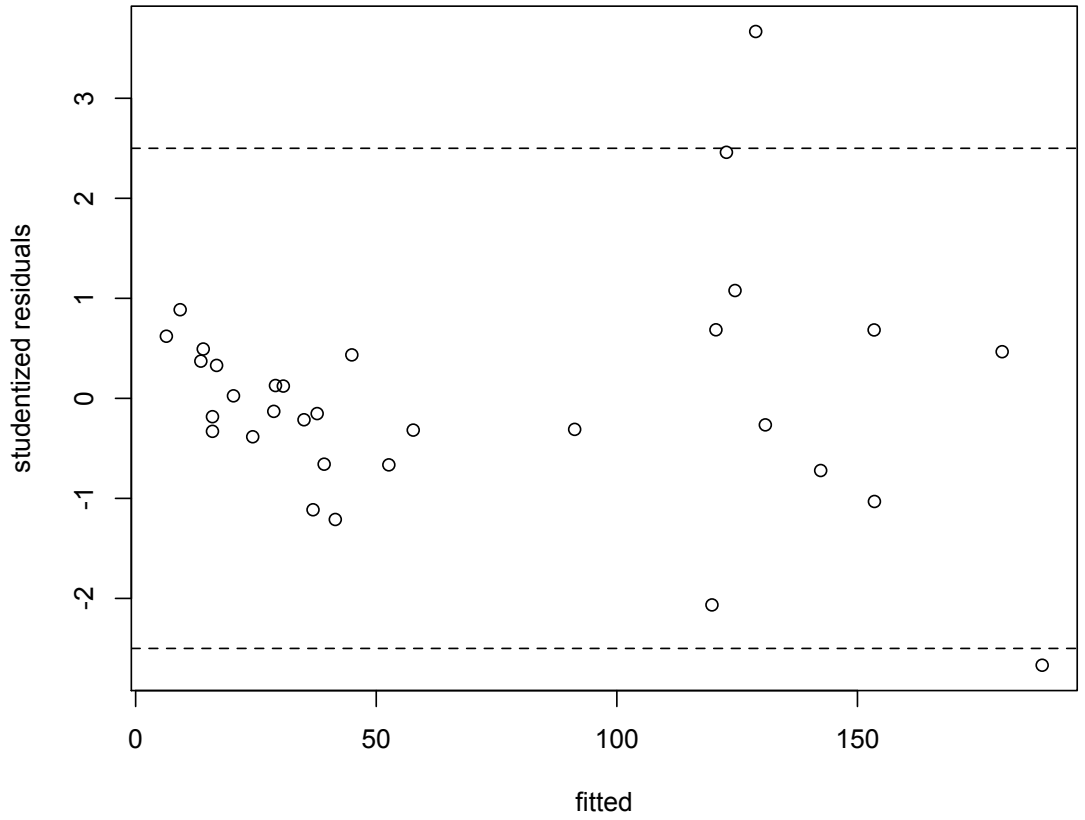
```
> fit2 <- lm(TNA~CAL+HT+HTCAL, data=dat2)
> summary(fit2)
Residuals:
    Min      1Q  Median      3Q     Max
-27.387  -6.534  -2.072   6.424  42.147

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -43.5282    40.9721  -1.062    0.297
CAL          70.6631    88.1728   0.801    0.430
HT            0.3328     1.6250   0.205    0.839
HTCAL         4.4023     2.6954   1.633    0.114

Residual standard error: 14.94 on 28 degrees of freedom
Multiple R-squared: 0.9423, Adjusted R-squared: 0.9361
F-statistic: 152.5 on 3 and 28 DF,  p-value: < 2.2e-16
```

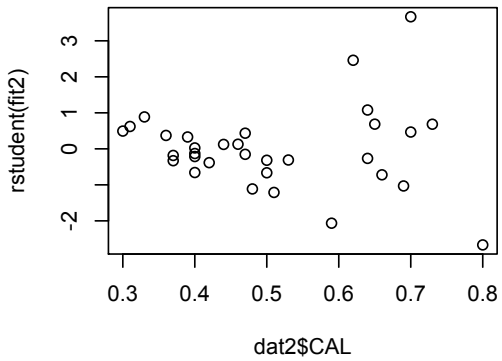
Studentized Residuals vs Fitted

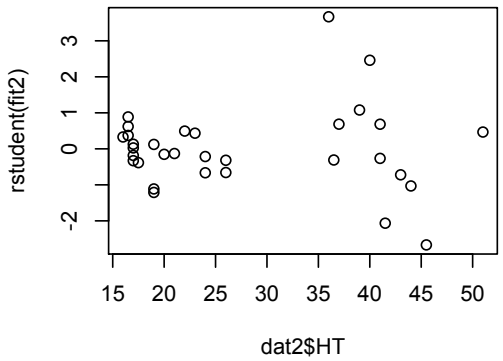
```
> plot(fitted(fit2), rstudent(fit2), xlab="fitted",
 ylab="studentized residuals")
> abline(h=c(-2.5,2.5),lty=2)
```

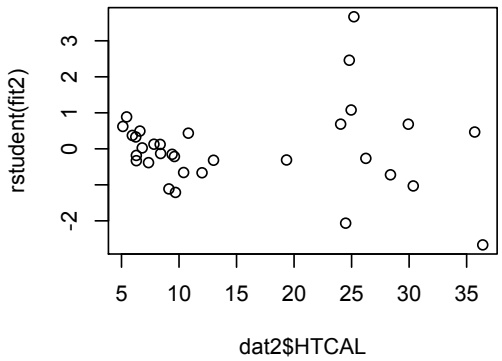


Studentized Residuals vs. Each Predictor

```
> par(mfrow=c(2,2))
> plot(dat2$CAL,rstudent(fit2))
> plot(dat2$HT,rstudent(fit2))
> plot(dat2$HTCAL,rstudent(fit2))
```

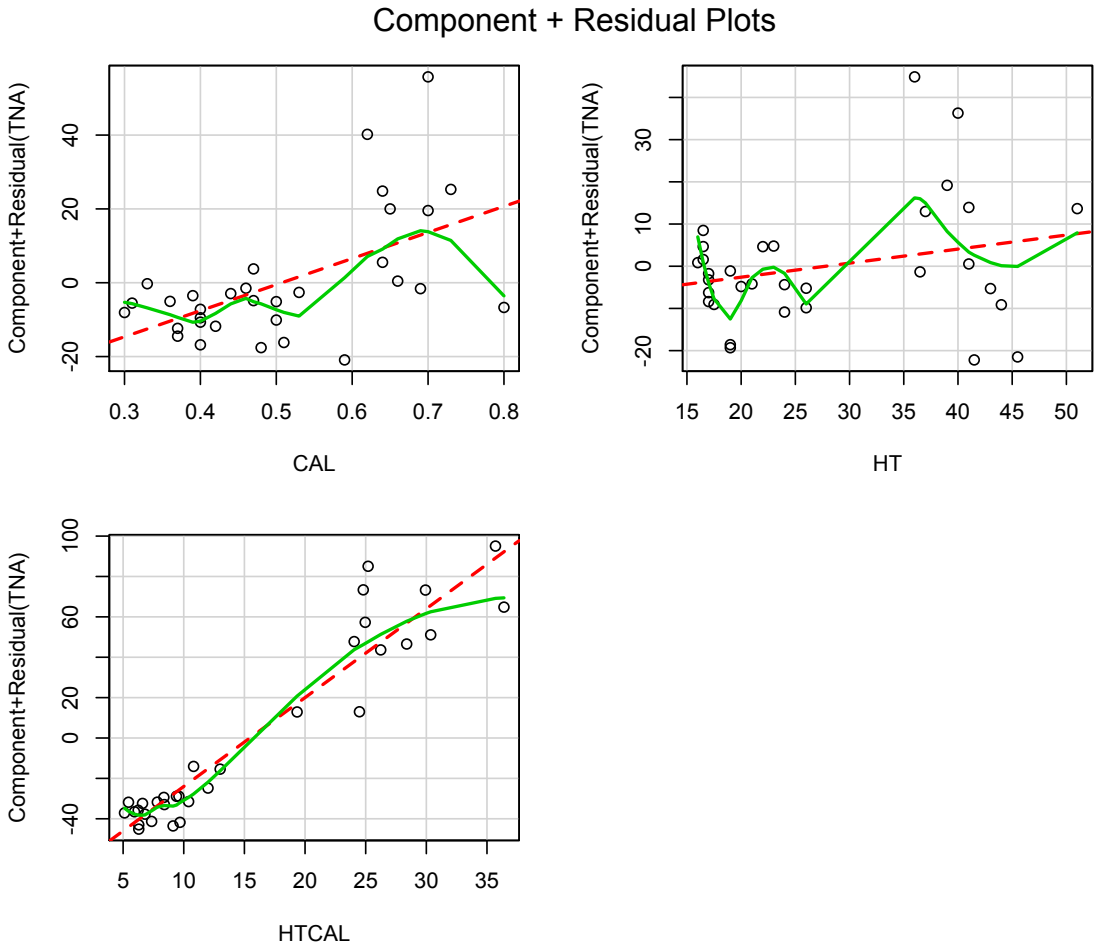






Partial Residuals vs. Each Predictor

```
> library(car)
> crPlots(fit2)
```

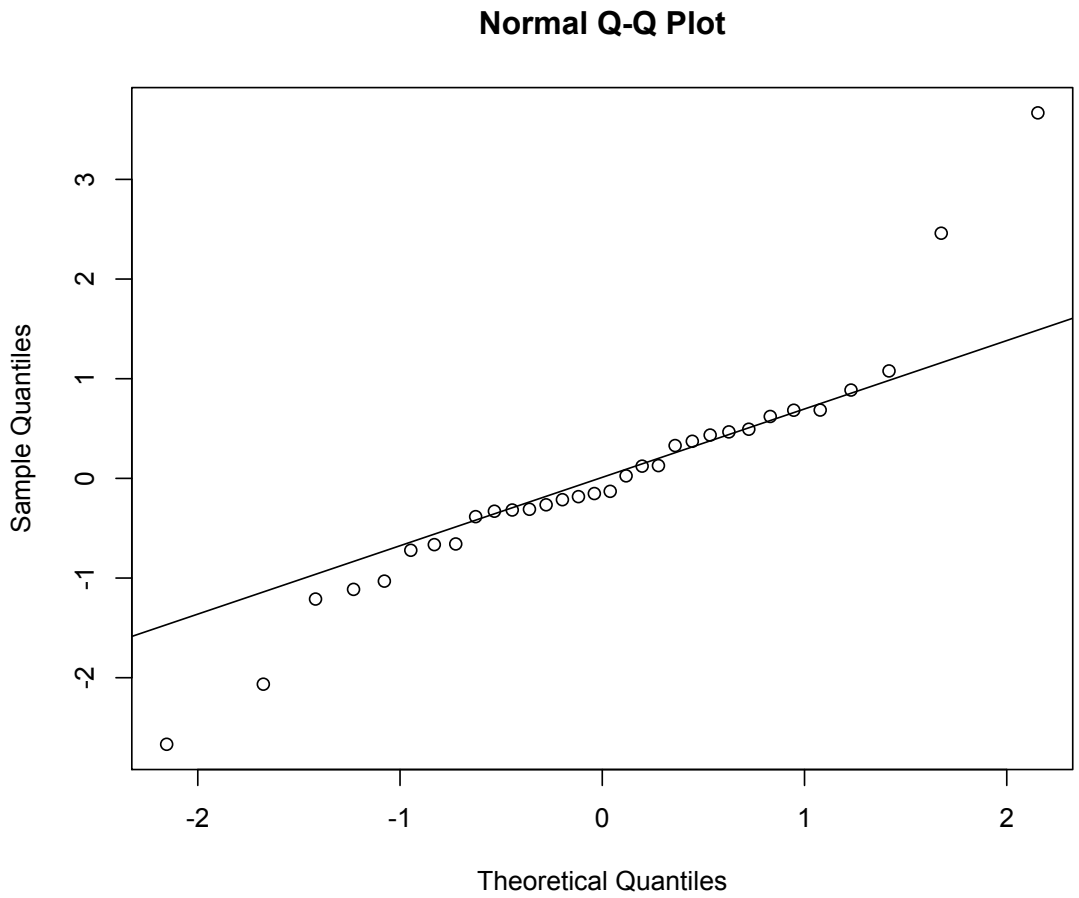


Summary of Residual Plots

- (Studentized) Residuals vs. Fitted Values
  Purpose: check whether $E(\epsilon) = 0$
  - Random pattern (no pattern), the above assumption is correct. Next, check non-constant variance, outliers...
  - Non-random pattern, need to check which term is nonlinear.
- (Studentized) Residuals vs. Each Predictor $x_j, j = 1, \ldots, p$
  Purpose: check if $E(y)$ is linear in x_j
  - Random pattern (no pattern), y is linear in x_j
  - Non-random pattern, y is not linear in x_j
- OR Partial residuals vs. Each Predictor $x_j, j = 1, \ldots, p$
  Purpose: check if $E(y)$ is linear in x_j
  - Linear pattern, y is linear in x_j
  - Non-linear pattern, y is not linear in x_j

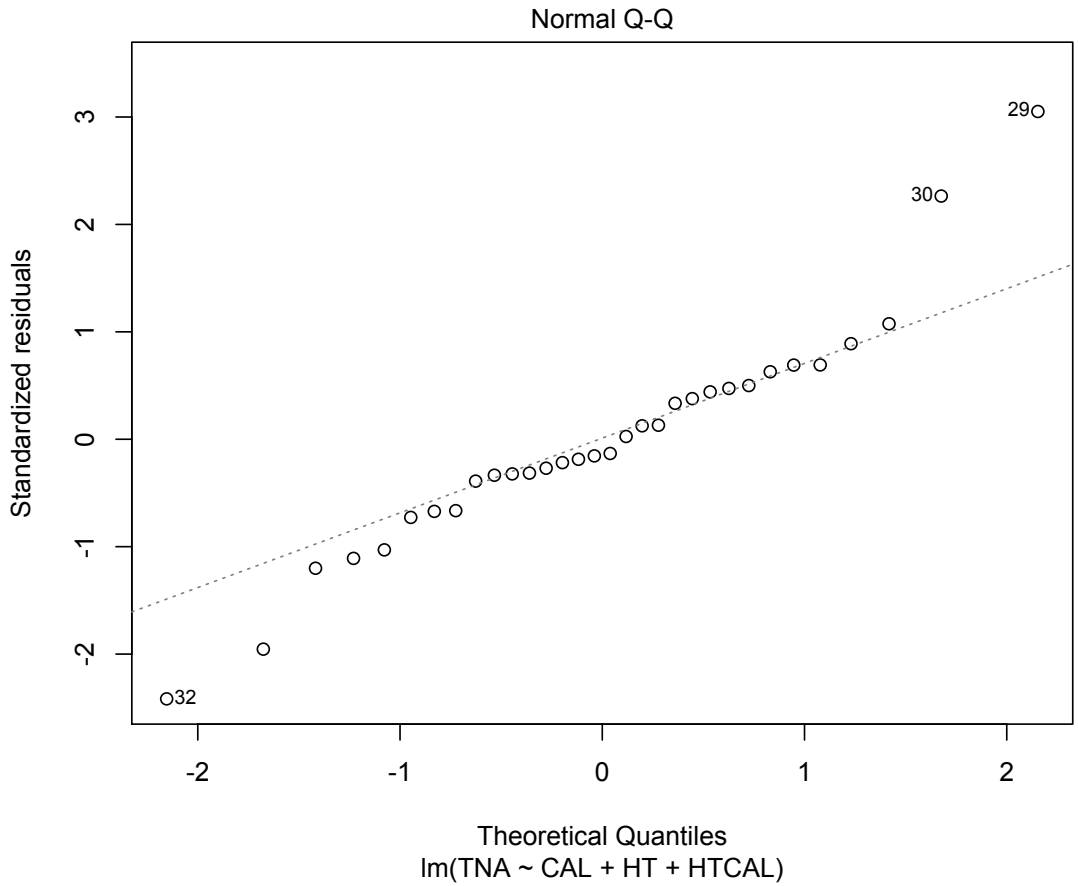
QQ Plot

```
> qqnorm(rstudent(fit2))
> qqline(rstudent(fit2))
```



QQ Plot

```
> plot(fit2, which=2)
```



Normality Test

- Shapiro-Wilk Normality Test:
  H_0: the (studentized) residuals follow a normal distribution
  H_a: the (studentized) residuals do not follow a normal distribution
- In R,

```
> shapiro.test(rstudent(fit2))
Shapiro-Wilk normality test

data:  rstudent(fit2)
W = 0.9158, p-value = 0.01604
```

- Since p-value$= 0.01604 < 0.05$, Reject H_0.

Leverage

```
> hatvalues(fit2)
        3         4         5         6         7         8         9
0.22584297 0.08295463 0.08516022 0.08810006 0.05319807 0.13408625 0.10333210 0.0644
       12        13        14        15        16        17        18
0.07980348 0.08588250 0.16550782 0.06615480 0.15089373 0.19603094 0.10656738 0.0882
       20        21        22        23        24        25        26
0.07027305 0.10956585 0.06083412 0.08588250 0.29579006 0.14472101 0.08140189 0.1089
       28        29        30        31        32        33        34
0.09263875 0.14538859 0.09234465 0.15984114 0.42486893 0.19216122 0.08390370 0.0752

> 2*4/32
[1] 0.25
> which(hatvalues(fit2)>0.25)
[1] 24 32
```

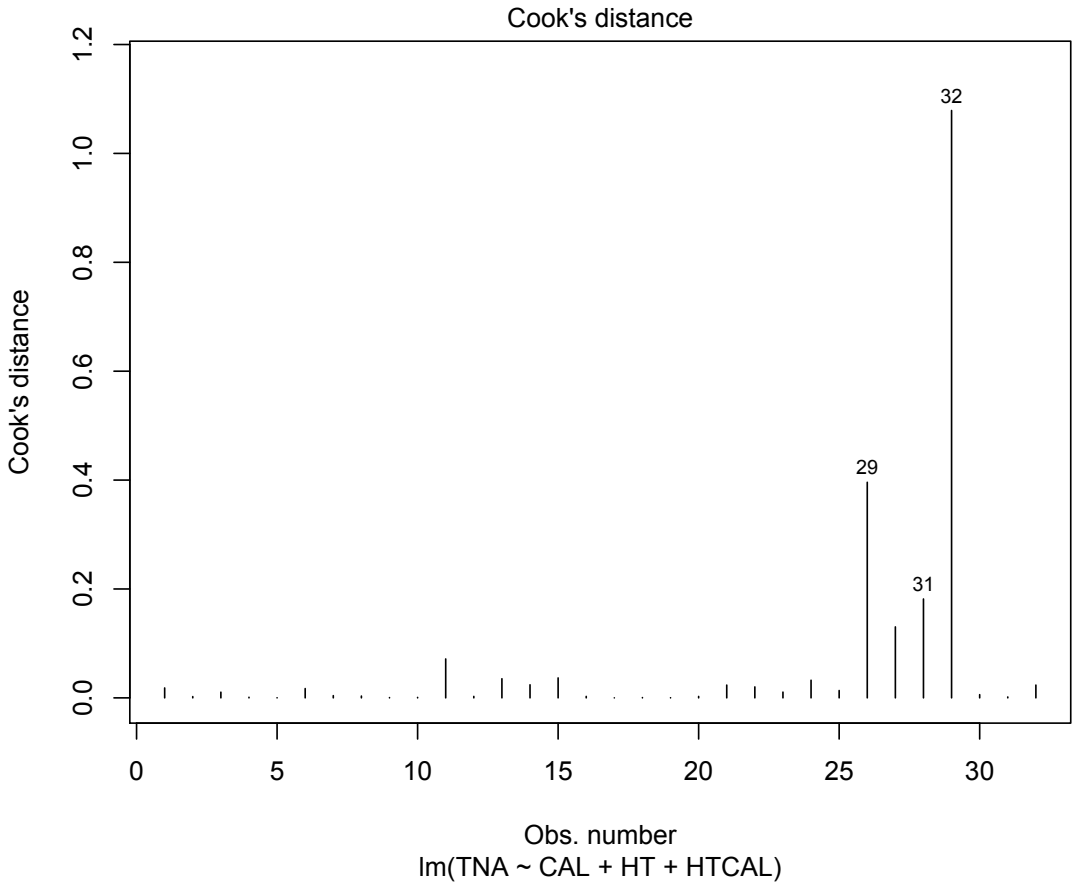
Cook's Distance

```
> cooks.distance(fit2)
           3            4            5            6            7            8
1.823248e-02 2.352843e-03 1.049554e-02 1.144257e-03 2.461355e-04 1.710854e-02 4.118
          11           12           13           14           15           16
3.347968e-03 5.203396e-04 8.172448e-04 7.150411e-02 2.693299e-03 3.515950e-02 2.404
          18           19           20           21           22           23
3.667164e-02 2.709213e-03 1.248025e-05 5.240002e-04 2.532045e-04 2.634907e-03 2.349
          25           26           27           28           29           30
2.016396e-02 1.060849e-02 3.242375e-02 1.351241e-02 3.959810e-01 1.303862e-01 1.816
          32           33           34           35
1.078827e+00 5.904776e-03 1.664538e-03 2.351611e-02

> qf(0.5,4,32-4)
[1] 0.8598354
> which(cooks.distance(fit2)>0.86)
[1] 32
```

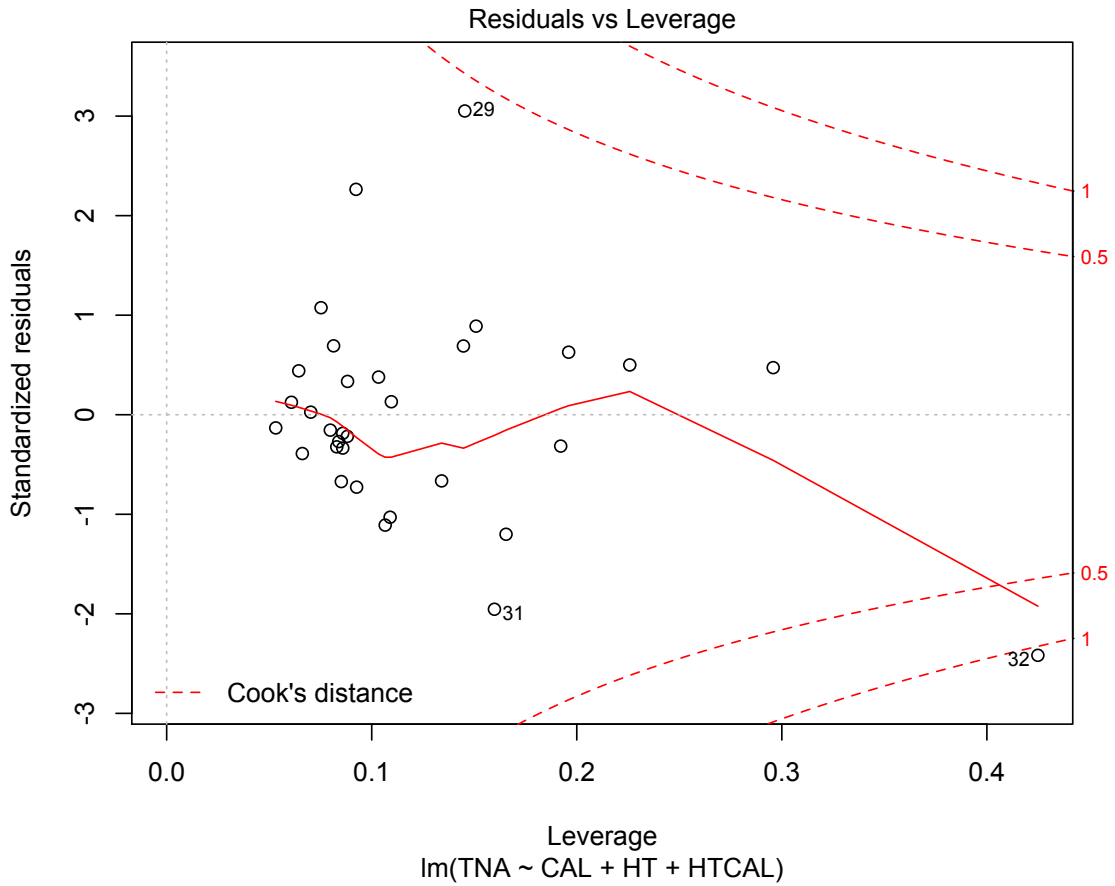
Cook's Distance

```
> plot(fit2, which=4)
```



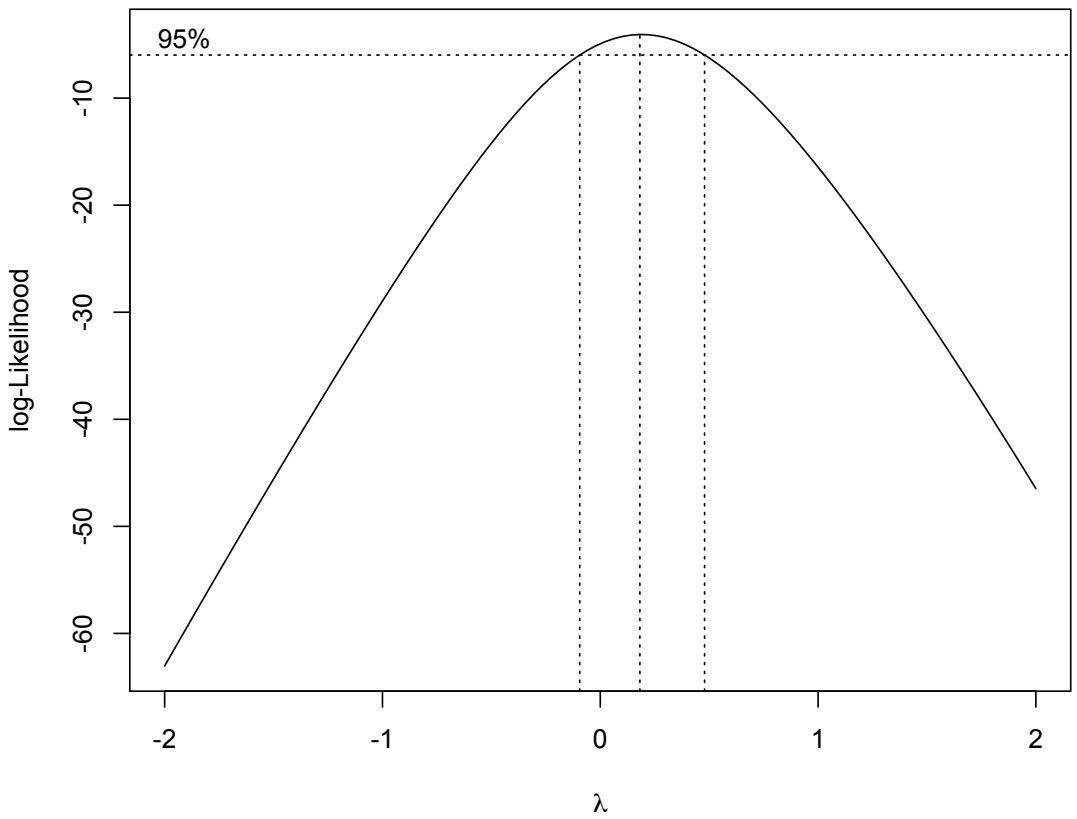
Residual, Leverage & Cook's D

```
> plot(fit2, which=5)
```



Box Cox Transformation

```
> library(MASS)
> boxcox(fit2)
```



Transformed Model

Consider the model:

$$log(TNA) = \beta_0 + \beta_1 CAL + \beta_2 HT + \beta_3 HTCAL + \epsilon$$

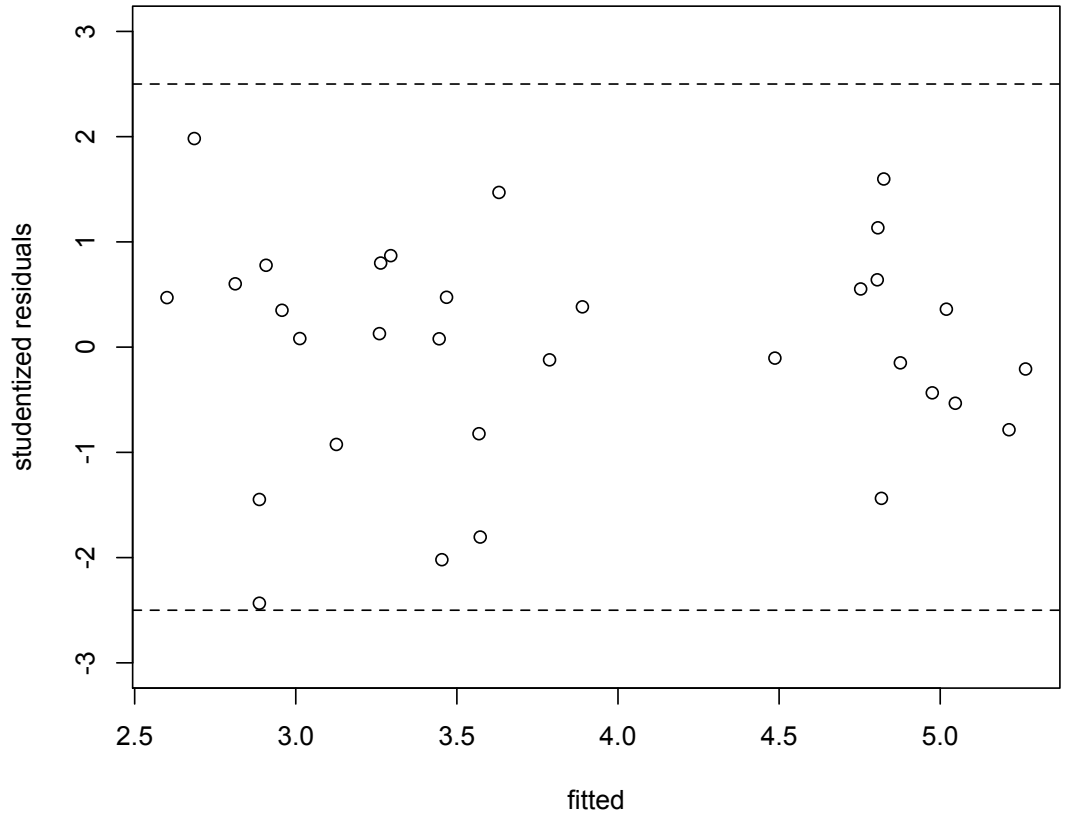
```
Residuals:
     Min       1Q   Median       3Q      Max
-0.47292 -0.11734  0.02279  0.12950  0.38300

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.44651    0.60440  -0.739  0.46619
CAL          6.02168    1.30068   4.630 7.63e-05 ***
HT           0.10505    0.02397   4.382  0.00015 ***
HTCAL       -0.10816    0.03976  -2.720  0.01108 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.2204 on 28 degrees of freedom
Multiple R-squared: 0.9471, Adjusted R-squared: 0.9414
F-statistic: 167.1 on 3 and 28 DF,  p-value: < 2.2e-16
```

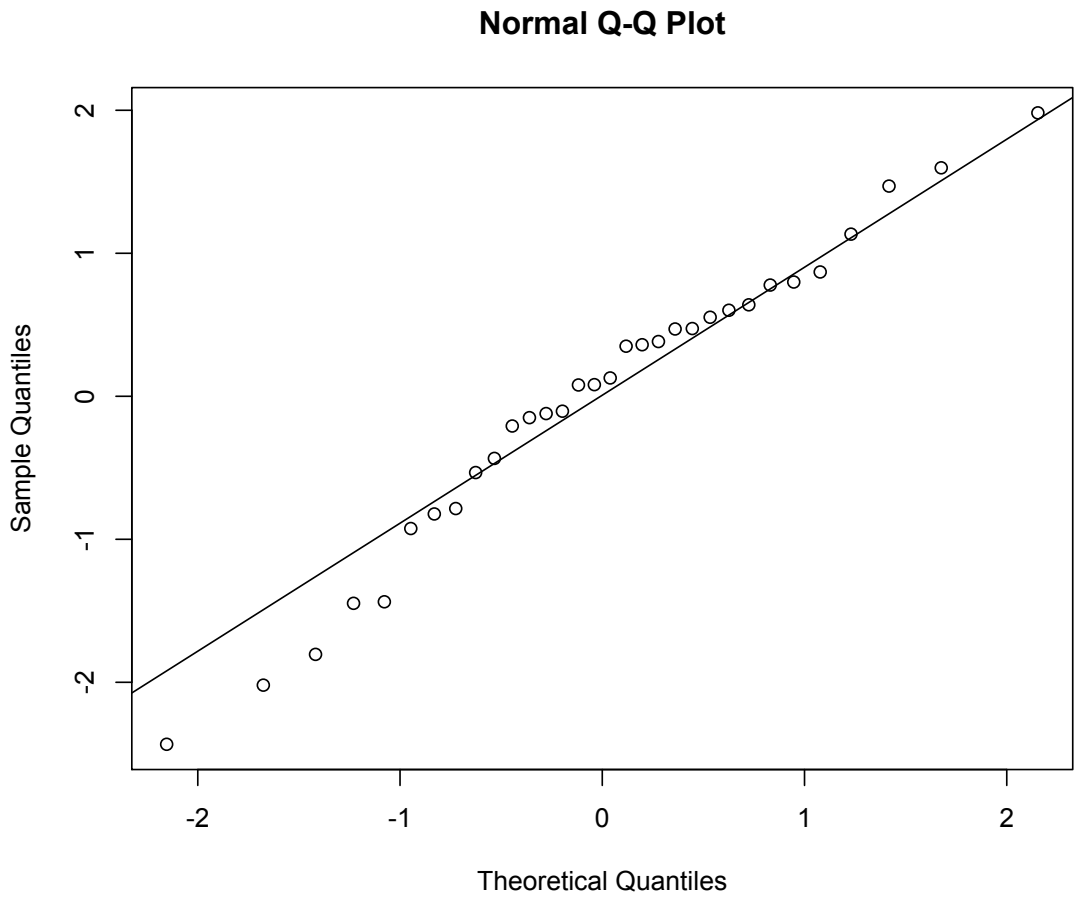
Residual vs fitted

```
> plot(fitted(fit3), rstudent(fit3), xlab="fitted",
ylab="studentized residuals")
> abline(h=c(-2.5, 2.5), lty=2)
```



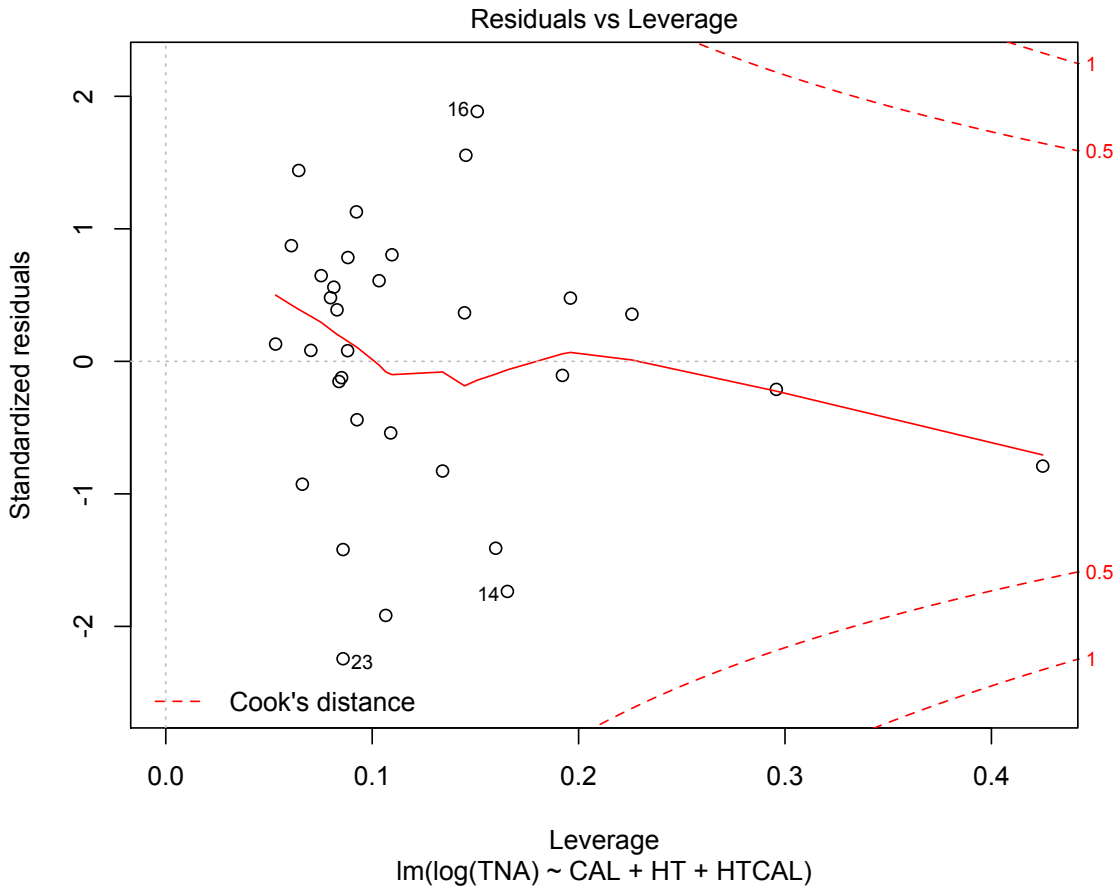
QQ Plot

```
> qqnorm(rstudent(fit3))
> qqline(rstudent(fit3))
```



Residual, Leverage & Cook's D

```
>  plot(fit3, which=5)
> qf(0.5,4,32-4)
[1] 0.8598354
> 2*4/32
[1] 0.25
```



Conclusion

The final fitted regression model is:

$$log(\widehat{TNA}) = -0.447 + 6.022CAL + 0.105HT - 0.108HTCAL$$

Based on the final model, we can:

- Interpretation
- Estimation
- Prediction