

# Introduction to Statistical Modeling

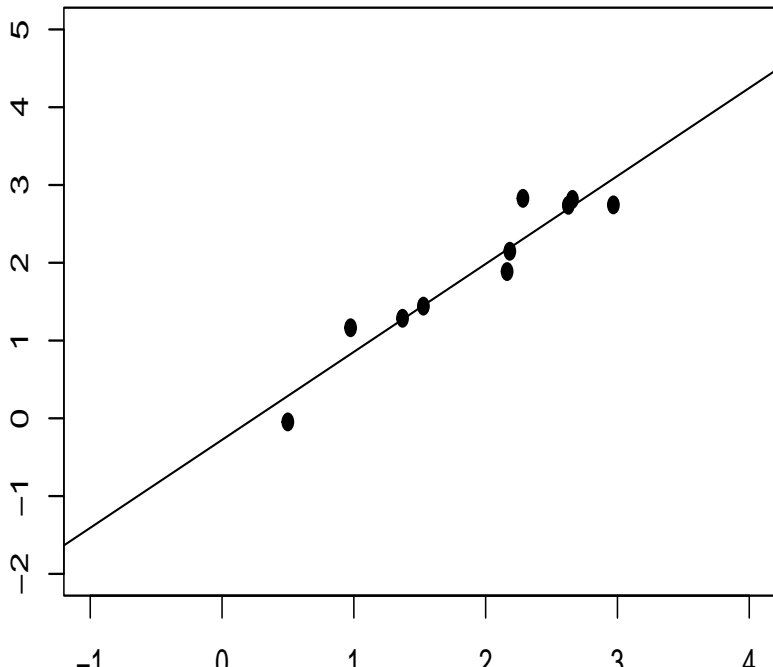
## Outliers

Joris Vankerschaver

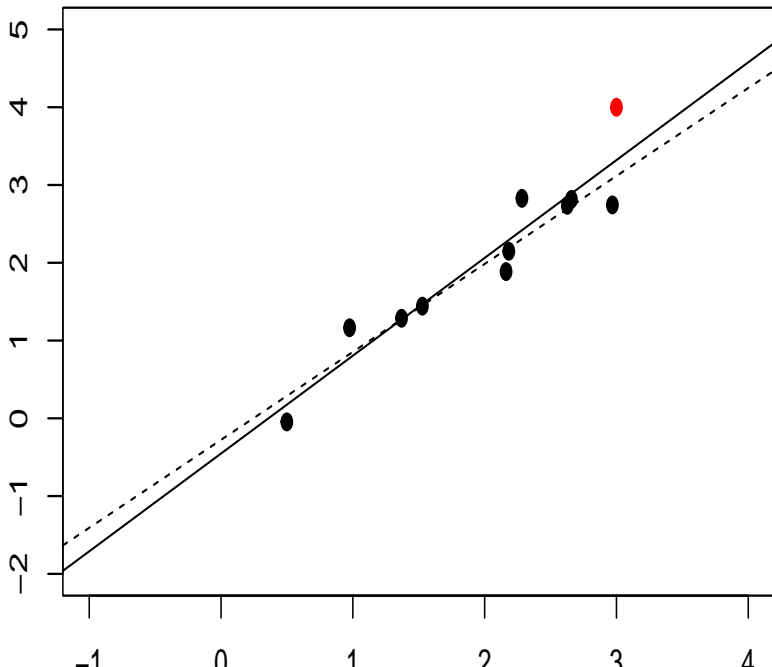
## Outliers / Influential observations

- Dataset often contains extreme values for outcome  $Y$  and/or predictors  $X$
- These **can** influence regression line strongly (but don't have to)

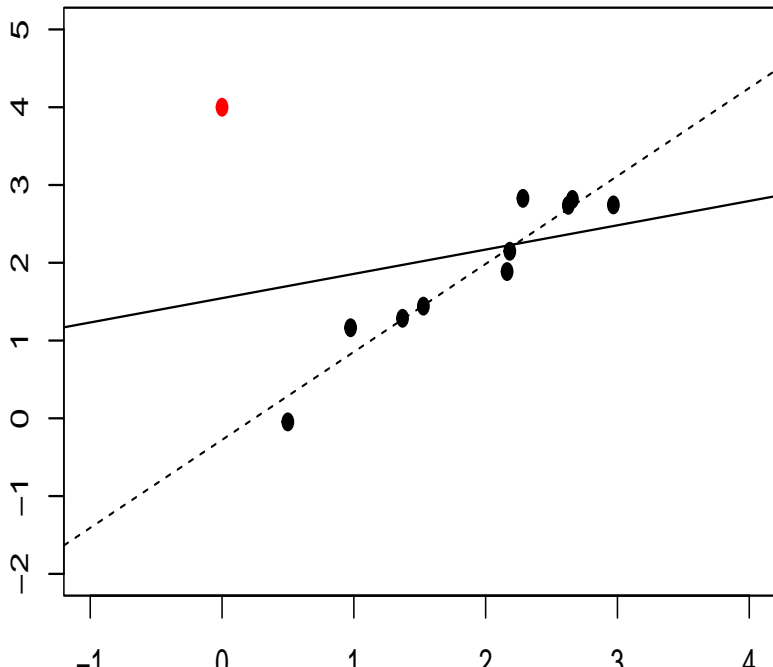
## Influence of influential observations



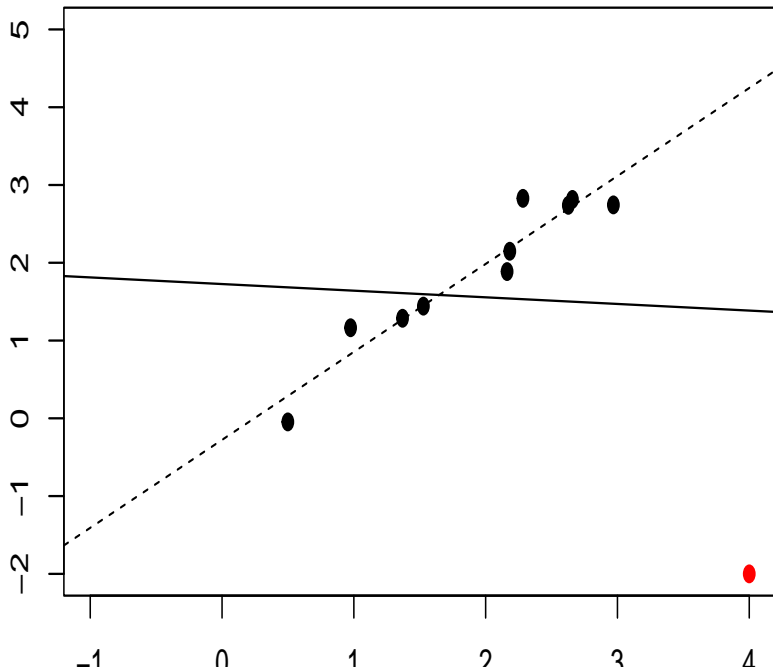
## Influence of influential observations



## Influence of influential observations



## Influence of influential observations



# Tracking influential observations

## Residuals:

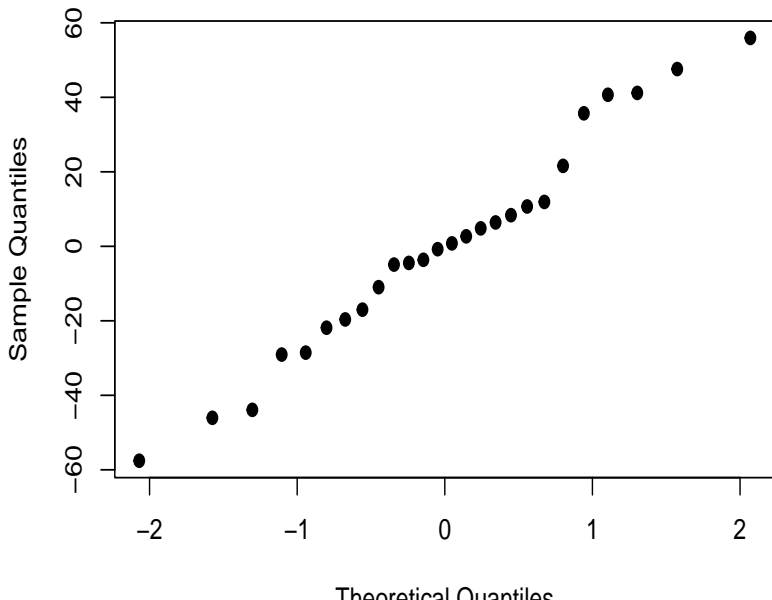
- Indicate how far outcome deviates from regression line
- Normally distributed with mean 0 and variance  $\sigma^2 = MSE$ .

Hence, can be used to identify extreme outcomes:

- 95% of residuals expected in interval  $[-2\sigma, 2\sigma]$
- Observations where residual is much larger are probably outliers.

# Exteme outcomes in analysis larches?

## Normal Q-Q Plot

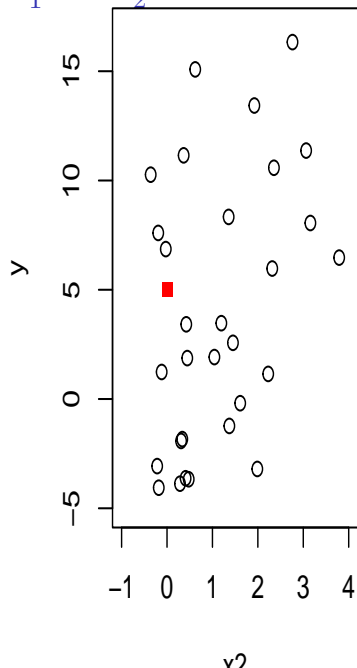
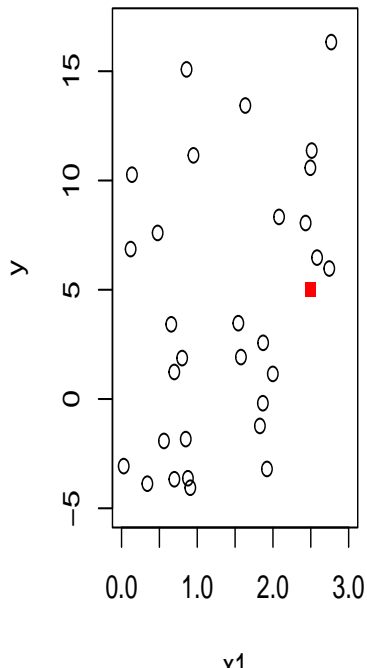




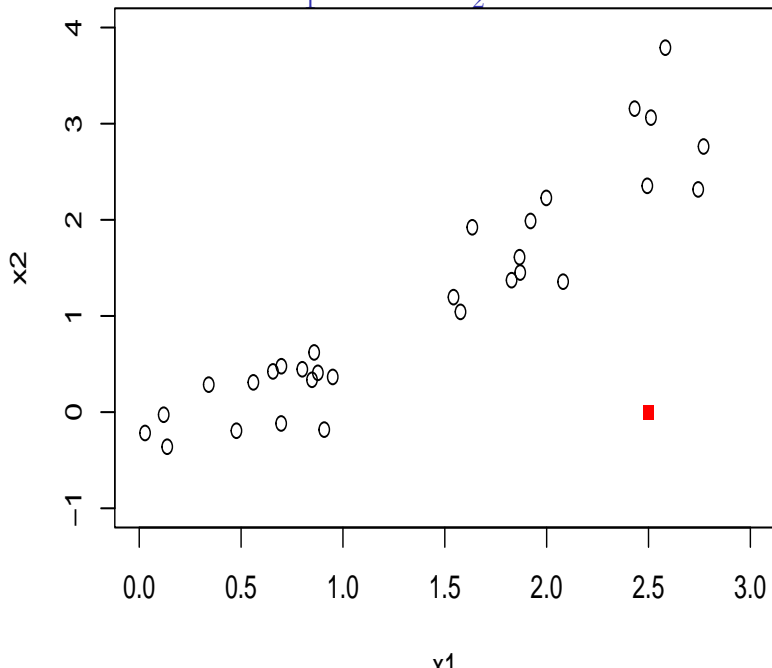
# Tracking influential observations

- **Scatterplots** of outcome in function of predictors can be used to identify extreme outcomes and predictors
- When multiple predictors, these plots have serious shortcomings

## Multivariate outliers: $Y$ versus $X_1$ or $X_2$



## Multivariate outliers: $X_1$ versus $X_2$



# Leverage

- Diagnostic measure to identify influential predictor-observations
- Data point has high leverage if it has “extreme” predictor values (low or high)

Mathematically:

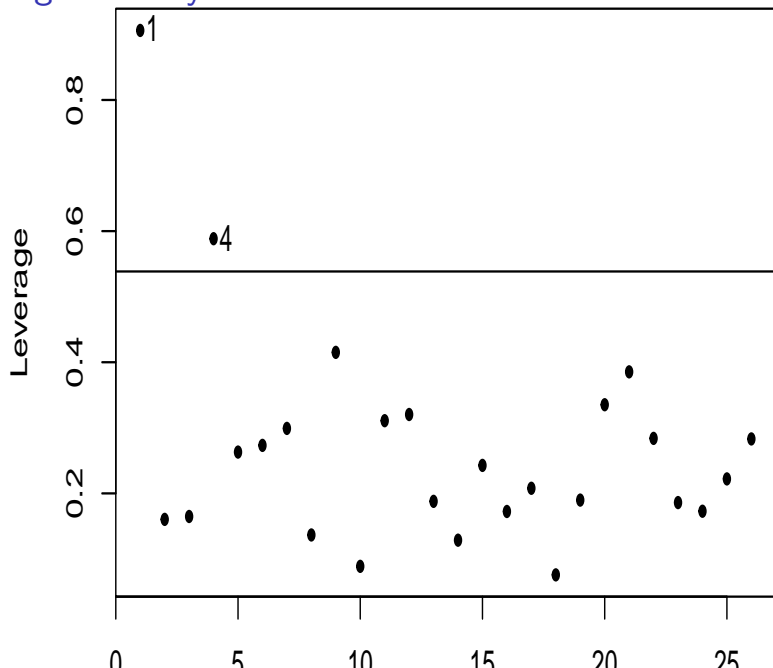
- Weighted distance between predictor for observation  $i$  and mean predictor.
- How much the  $i$ th observed value affects the  $i$ th fitted value:

$$h_{ii} = \frac{\partial \hat{y}_i}{\partial y_i}.$$

# Interpretation of leverage

- If high leverage for  $i^{th}$  observation, then
  - it has predictor values that deviate strongly from the mean
  - it **possibly** has large influence on regression coefficients and predictions
- Leverage is on average  $p/n$  with  $p$  number of unknown parameters
- **Extreme leverage**: larger than  $2p/n$

## Leverage in analysis of larches



## Cook's distance

- Diagnostic measure for influence of  $i^{th}$  observation on all predictions / estimated coefficients.
- Cook's distance for  $i^{th}$  observation is obtained by comparing each prediction  $\hat{Y}_j$  with prediction  $\hat{Y}_{j(i)}$  that would be obtained **if  $i^{th}$  observation was deleted**:

$$D_i = \frac{\sum_{j=1}^n (\hat{Y}_j - \hat{Y}_{j(i)})^2}{p \cdot \text{MSE}}$$

## Interpretation Cook's distance

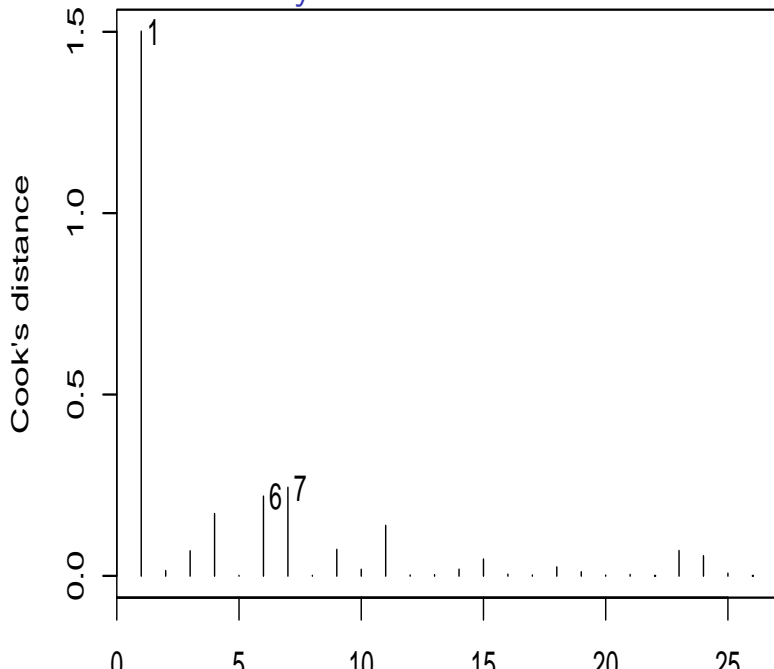
- If Cook's distance  $D_i$  large, then  $i^{th}$  observation has large influence on predictions and coefficients
- **Extreme Cook's distance:** exceeds 50% percentile of  $F_{p,n-p}$ -distribution

Example:

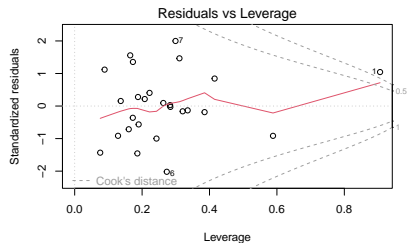
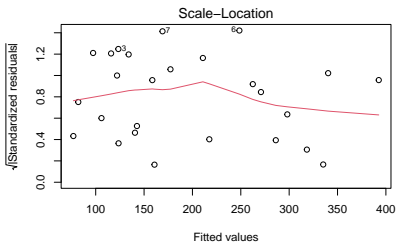
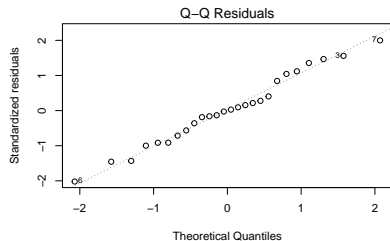
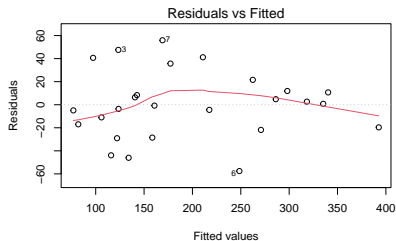
- In analysis of larches is  $p = 7, n = 26$  and the 50% percentile of  $F_{p,n-p}$ -distribution 0.94
- Cook's distance of first observation is 1.5 and corresponds to 77% percentile
- Conclusion: first observation has large influence on estimated regression coefficients



## Cook's distance in analysis of larches



# Analysis of larches: residual plots



# DFBETAs

On what coefficient(s) will first observation have large influence?

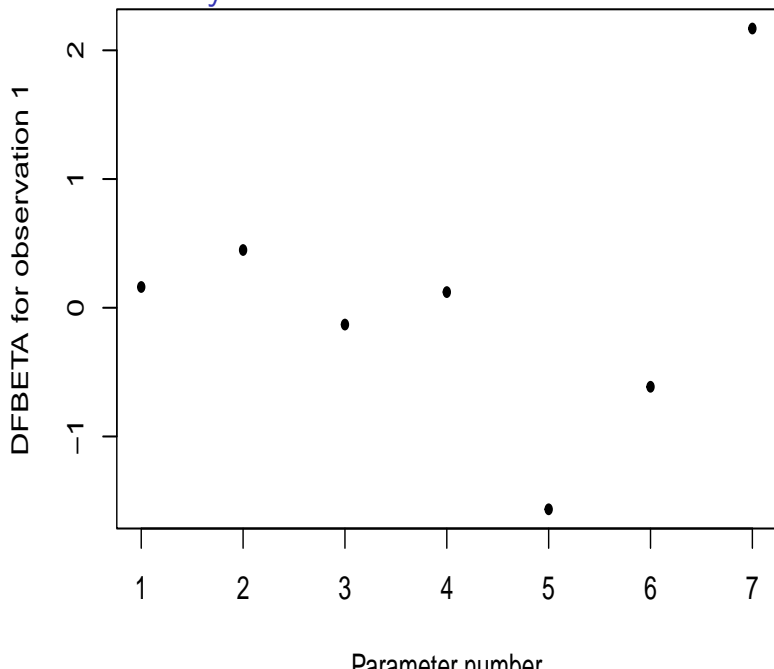
- Diagnostic measure for influence of  $i^{th}$  observation **on each regression coefficient separately**
- DFBETAs for  $i^{th}$  observation and  $j^{th}$  coefficient is obtained by comparing  $j^{th}$  coefficient  $\hat{\beta}_j$  with coefficient  $\hat{\beta}_{j(i)}$  from model **if  $i^{th}$  observation would have been deleted**

$$DFBETA_{j(i)} = \frac{\hat{\beta}_j - \hat{\beta}_{j(i)}}{SD(\hat{\beta}_j)}$$

# Interpretation DFBETAs

- Sign indicates if deleting observation  $i$  causes an increase ( $\text{DFBETA} < 0$ ) or decrease ( $\text{DFBETA} > 0$ ) in each coefficient
- **Extreme DFBETAs:** exceeds 1 in small to moderate datasets, and  $2/\sqrt{n}$  in large datasets

## DFBETAs in analysis of larches



## DFBETAs in analysis of larches

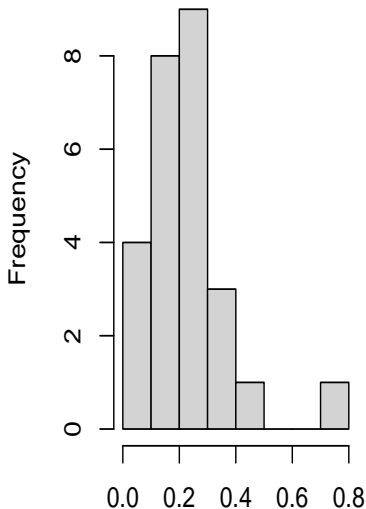
First observation has **large influence on interaction between phosphorus and residual ash**:

- current coefficient is -598.08 (SE 290.02);
- DFBETA is 2.17;
- after deletion of first observation, interaction between phosphorus and residual ash will be around

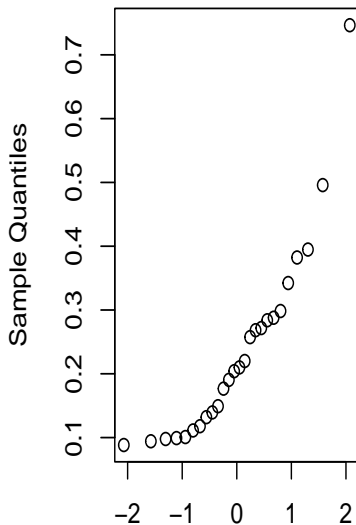
$$-598.08 - 2.17 \times 290.02 = -1227.42$$

## Histogram and QQ-plot of interaction

Histogram of phosphor \* resid



Normal Q-Q Plot



## Analysis of larches after deletion 1<sup>st</sup> observation

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	101.8433	170.89955	0.5959248	0.558645291
nitrogen2	-194.2135	105.36608	-1.8432260	0.081826133
phosphor2	-911.1361	686.55995	-1.3271035	0.201063465
potassium2	132.9597	41.41191	3.2106631	0.004847492
residu2	332.9542	159.95466	2.0815538	0.051930532
nitrogen2:phosphor2	1179.3050	429.71839	2.7443671	0.013331564
phosphor2:residu2	-1225.6303	665.87615	-1.8406279	0.082222531



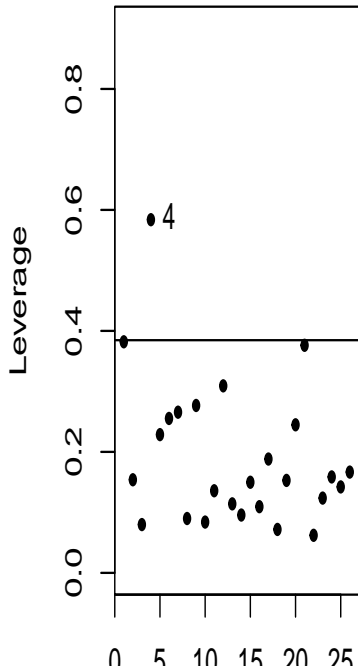
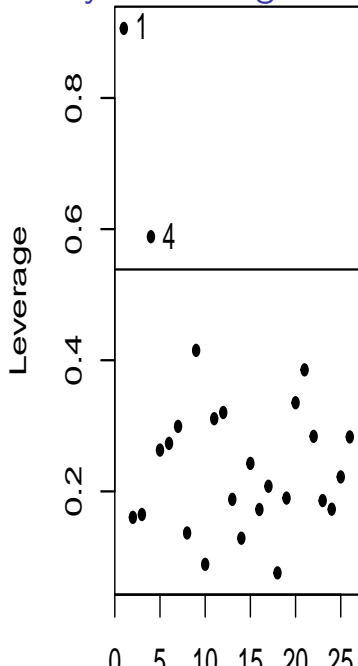
## Analysis of larches after deletion interaction

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	134.47244	182.5540	0.7366172	0.469908260
nitrogen	-66.35461	94.9178	-0.6990744	0.492556067
phosphor	-1024.58992	736.3183	-1.3915041	0.179357766
potassium	128.83662	44.2072	2.9143806	0.008574138
residu	23.51194	36.0929	0.6514284	0.522185637
nitrogen:phosphor	661.49644	370.7815	1.7840600	0.089594772

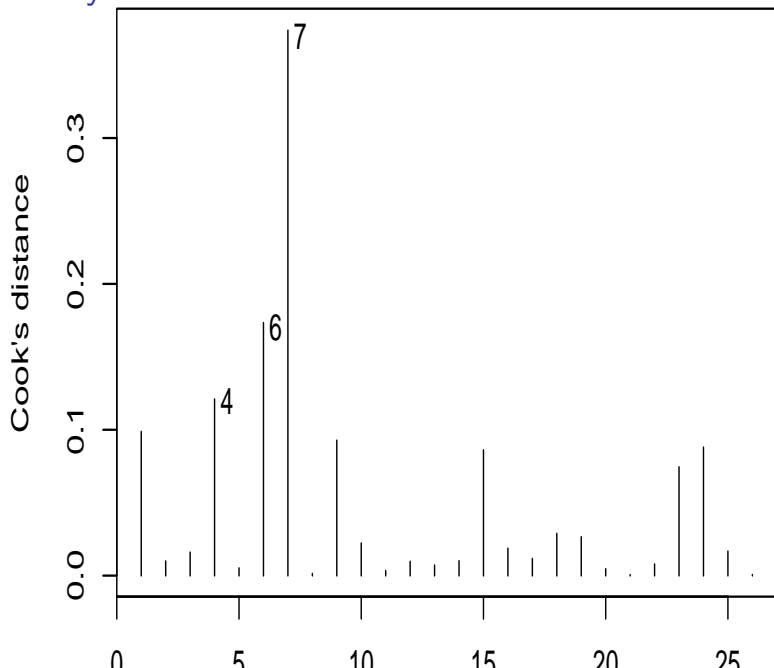
## Analysis of larches: final model

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	160.66283	175.61424	0.9148622	0.370649894
nitrogen	-76.49677	92.34000	-0.8284250	0.416746264
phosphor	-1120.70470	711.42841	-1.5752881	0.130135986
potassium	138.06170	41.29966	3.3429260	0.003084272
nitrogen:phosphor	724.38231	353.05353	2.0517634	0.052870451

## Final analysis: leverage



## Final analysis: Cook's distance



# Final analysis: residual plots

