Markéta Novotná

Regresní modelování – Projekt

Popis datového souboru:

Datový soubor pšenice.txt obsahuje koncentrace 5 těžkých kovů (Cd, Cu, Nu, Pb a Zn) v rostlinách (pšenice zrno) a v půdě. Dále pH půdy (ph_vym), obsah organického uhlíku v půdě (C_org) a hmotnost zkoumaných rostlin (hmota_abs). Koncentrace jsou uvedeny v mg/kg, hmota v g a organický uhlík v mg/g. Koncentrace kovů v půdě byla měřena dvěma metodami (kov_HNO3, kov_AR) v 2 M HNO3 a v Aqua regia.

Podle dostupné literatury by se na obsahu kovů v rostlinách měla nejvíce podílet jeho koncentrace v půdě, pH půdy a množství organického uhlíku v půdě.

Prohlídka dat:

```
       Cd                Cu               Ni                Pb               Zn
 Min.   :0.0200    Min.   :0.996    Min.   :0.2000    Min.   :0.2000    Min.   : 15.60
 1st Qu.:0.0500    1st Qu.:2.680    1st Qu.:0.2000    1st Qu.:0.2000    1st Qu.: 22.60
 Median :0.0800    Median :3.400    Median :0.3000    Median :0.2000    Median : 28.60
 Mean   :0.1392    Mean   :3.507    Mean   :0.3253    Mean   :0.3839    Mean   : 32.25
 3rd Qu.:0.1500    3rd Qu.:4.260    3rd Qu.:0.3200    3rd Qu.:0.8000    3rd Qu.: 38.20
 Max.   :0.8300    Max.   :6.530    Max.   :1.1000    Max.   :0.8300    Max.   :108.00

   hmota_abs        ph_vym           C_org
 Min.   :106.5    Min.   :4.800    Min.   :0.920
 1st Qu.:277.8    1st Qu.:6.025    1st Qu.:1.150
 Median :404.9    Median :6.575    Median :1.310
 Mean   :401.1    Mean   :6.455    Mean   :1.451
 3rd Qu.:472.6    3rd Qu.:6.900    3rd Qu.:1.650
 Max.   :851.4    Max.   :7.600    Max.   :2.810
 NA's   : 25.0
     Cd_HNO3          Cu_HNO3          Ni_HNO3          Pb_HNO3          Zn_HNO3
 Min.   :0.1250    Min.   :  3.475    Min.   : 1.425    Min.   :  9.485    Min.   :  7.485
 1st Qu.:0.1975    1st Qu.:  7.060    1st Qu.: 2.985    1st Qu.: 15.450    1st Qu.:  19.175
 Median :0.2450    Median :  9.367    Median : 5.797    Median : 20.788    Median :  26.194
 Mean   :0.9802    Mean   : 22.209    Mean   : 7.088    Mean   : 80.335    Mean   :  97.393
 3rd Qu.:0.9350    3rd Qu.: 14.834    3rd Qu.: 9.318    3rd Qu.: 34.932    3rd Qu.:  83.500
 Max.   :6.5933    Max.   :167.650    Max.   :22.726    Max.   :799.000    Max.   :1034.250

    Cd_AR            Cu_AR            Ni_AR            Pb_AR            Zn_AR
 Min.   :0.1050    Min.   :  6.36    Min.   :  7.746    Min.   :  9.675    Min.   :  32.33
 1st Qu.:0.2250    1st Qu.: 15.34    1st Qu.: 16.865    1st Qu.: 16.950    1st Qu.:  64.81
 Median :0.3375    Median : 22.57    Median : 22.220    Median : 24.925    Median :  80.05
 Mean   :1.1781    Mean   : 37.37    Mean   : 25.526    Mean   : 87.547    Mean   : 174.19
 3rd Qu.:1.0500    3rd Qu.: 33.95    3rd Qu.: 27.245    3rd Qu.: 37.766    3rd Qu.: 166.30
 Max.   :7.7538    Max.   :230.43    Max.   :114.700    Max.   :813.525    Max.   :1208.00
```

jednotlivé proměnné:

"Cd", "Cu", "Ni", "Pb", "Zn" – koncentrace kovu v rostlinách

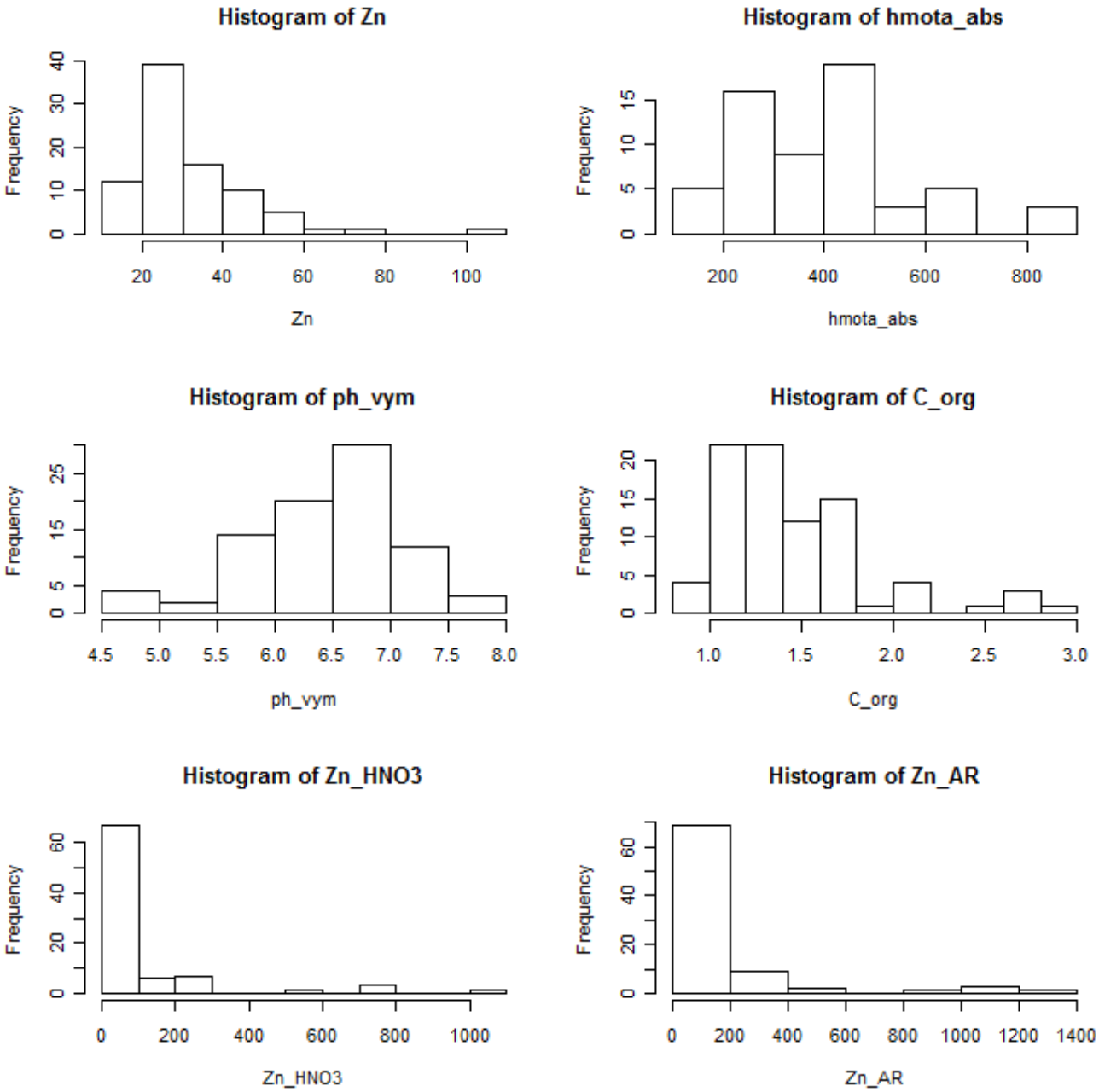
"hmota_abs", "ph_vym", "C_org"

"Cd_HNO3", "Cu_HNO3", "Ni_HNO3", "Pb_HNO3", "Zn_HNO3" – koncentrace kovu v půdě

"Cd_AR", "Cu_AR", "Ni_AR", "Pb_AR", "Zn_AR" – koncentrace kovu v půdě

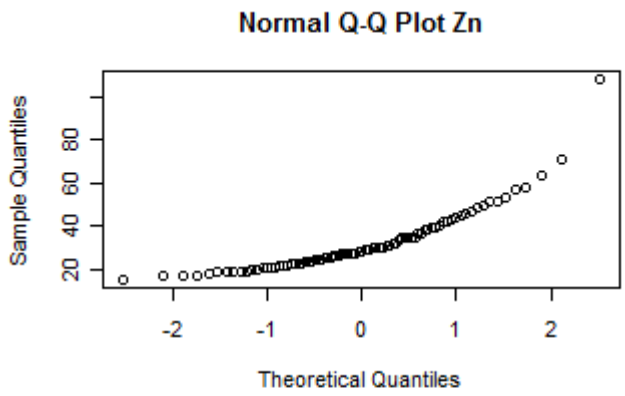
Pro tvorbu modelu jsem si vybrala pouze jeden kov – zinek a k němu příslušné vysvětlující proměnné.
Závisle proměnná – koncentrace zinku v rostlině (Zn).
Vysvětlující proměnné – koncentrace zinku v půdě (Zn_HNO3, Zn_AR), pH půdy (ph_vym) a organický uhlík v půdě (C_org).

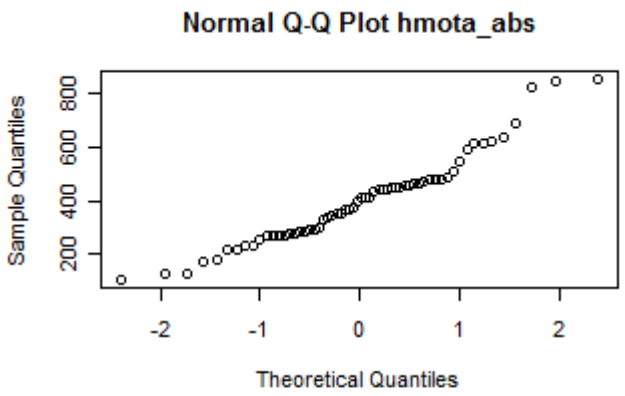
Histogramy:

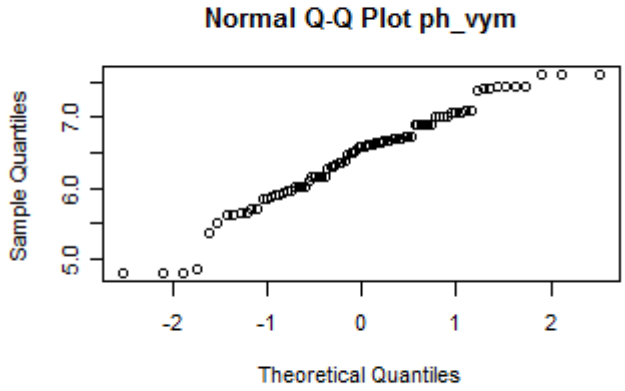


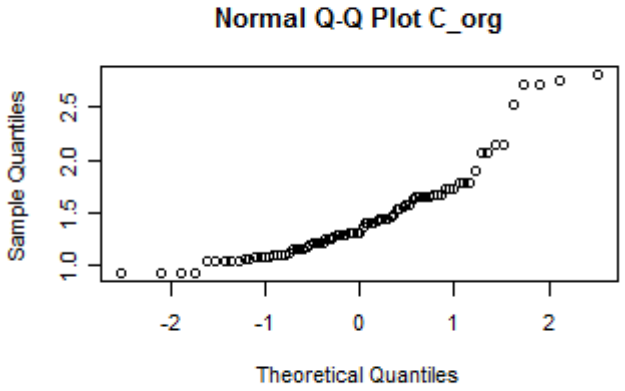
Rozložení hlavně koncentrace Zn v půdě je velmi nesymetrické. Možná by bylo vhodné použít logaritmickou transformaci dat.

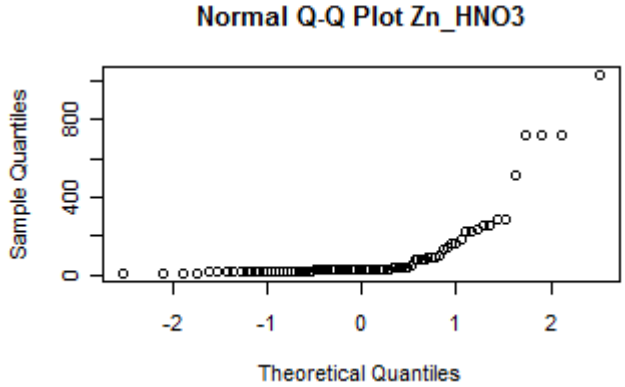
NP-ploty:

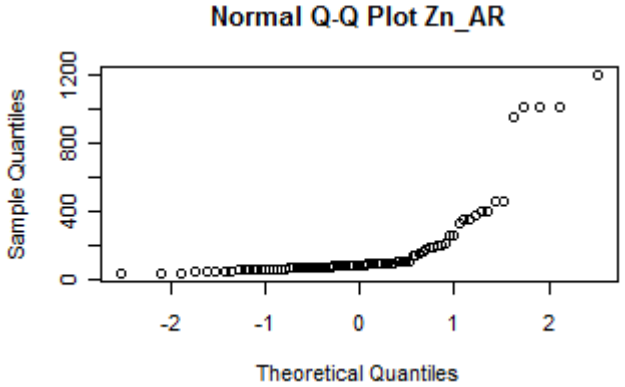






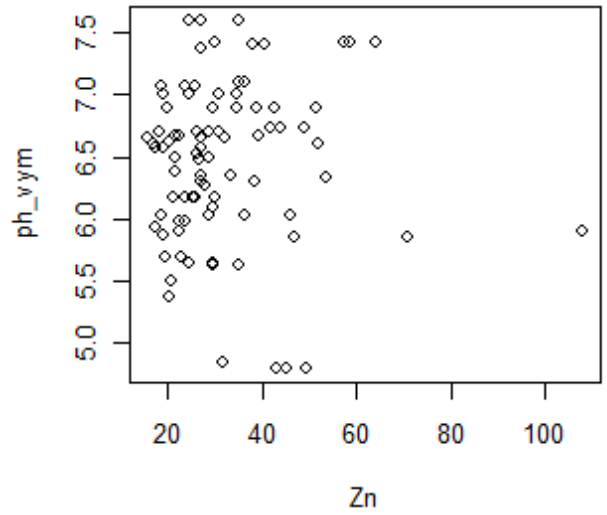


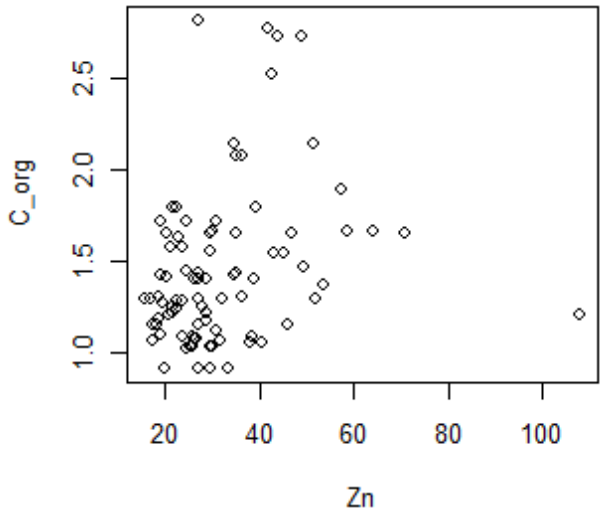


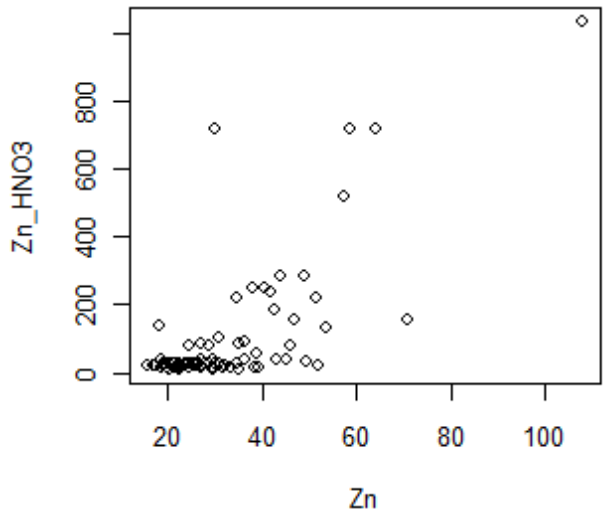


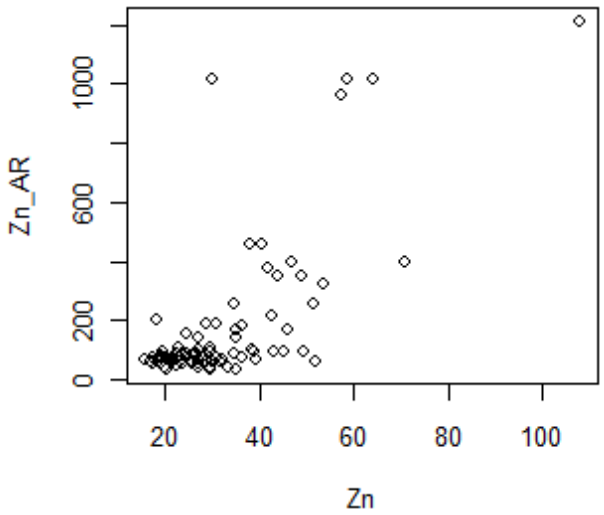
Normalita proměnných je většinou porušena. Nejlépe vypadá hmota_abs a ph_vym.

Závislost Zn na vysvětlujících proměnných:







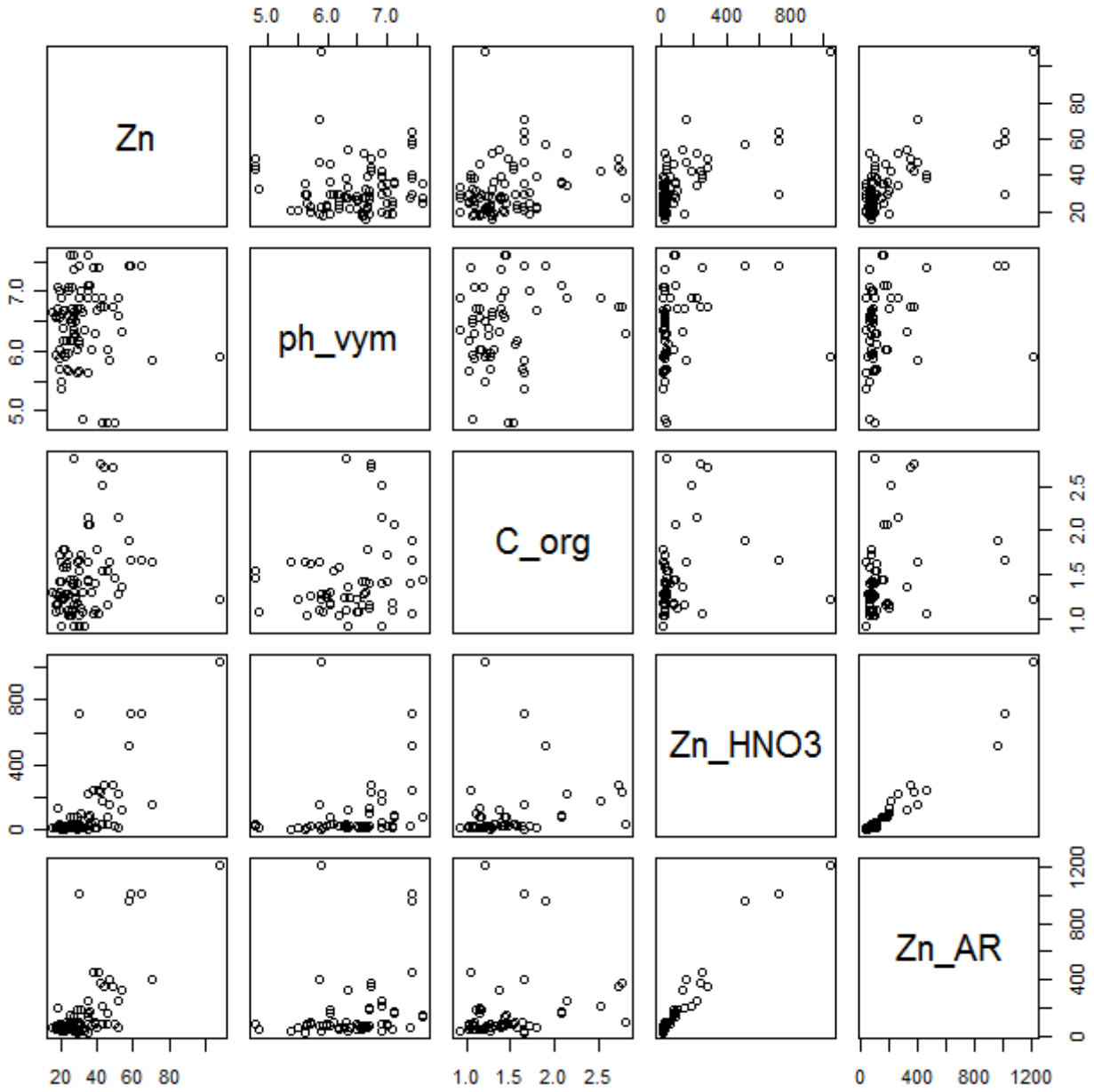


Je vidět přímá lineární závislost koncentrace Zn v rostlinách na organickém uhlíku i na koncentraci kovu v půdě. I když není příliš silná, viz korelační koeficienty:

> cor(Zn,C_org)
[1] 0.2633924
> cor(Zn,Zn_HNO3)
[1] 0.7188914
> cor(Zn,Zn_AR)
[1] 0.7151234

Větší hodnota u Zn_HNO3 (AR) je hlavně způsobena jednou odlehlou hodnotou.

Posouzení multikolinearity:



Koncentrace Zn v půdě měřená dvěma různými metodami je velmi korelovaná:

> cor(Zn_HNO3,Zn_AR)
[1] 0.9785904

Proto nebude možné do modelu použít obě proměnné zároveň. Bude lepší vytvořit model zvlášť s proměnnou Zn_HNO3 a Zn_AR.

Korelace dalších proměnných mezi sebou nejsou příliš výrazné, proto je můžeme použít dohromady do jednoho modelu:

> cor(ph_vym,C_org)
[1] 0.1734307
> cor(ph_vym,Zn_HNO3)
[1] 0.2939546
> cor(ph_vym,Zn_AR)
[1] 0.3002152
> cor(C_org,Zn_AR)
[1] 0.2453882
> cor(C_org,Zn_HNO3)
[1] 0.2596277

Model 1 – všechny vysvětlující proměnné:

```
Call:
lm(formula = Zn ~ ph_vym + C_org + Zn_HNO3, model = TRUE, x = TRUE,
    y = TRUE)

Residuals:
    Min      1Q  Median      3Q     Max
-35.4363  -6.1898  -0.3238   4.8963  30.9421

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 55.465396  10.813878   5.129 1.95e-06 ***
ph_vym      -5.316463   1.659700  -3.203  0.00194 **
C_org        3.593463   2.512344   1.430  0.15647
Zn_HNO3      0.060440   0.006147   9.832 1.81e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.469 on 81 degrees of freedom
Multiple R-squared: 0.5767,     Adjusted R-squared: 0.5611
F-statistic: 36.79 on 3 and 81 DF,  p-value: 4.189e-15
```

Celkově je model 1 významný (p-value: 4.189e-15) a vysvětluje 57,67 % variability. Všechny vysvětlující proměnné jsou významné kromě C_org.

Podobné výsledky dostaneme i při použití Zn_AR:

```
Call:
lm(formula = Zn ~ ph_vym + C_org + Zn_AR, model = TRUE, x = TRUE,
    y = TRUE)

Residuals:
    Min      1Q  Median      3Q     Max
-35.7491  -6.3180  -0.3555   4.9088  26.9996

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 53.781471  10.779731   4.989 3.40e-06 ***
ph_vym      -5.455568   1.665368  -3.276  0.00155 **
C_org        4.007870   2.506143   1.599  0.11367
Zn_AR        0.045163   0.004603   9.812 1.98e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.479 on 81 degrees of freedom
Multiple R-squared: 0.5758,     Adjusted R-squared: 0.5601
F-statistic: 36.65 on 3 and 81 DF,  p-value: 4.569e-15
```

Protože je proměnná organický uhlík nevýznamná, zkusíme jednodušší model jen se dvěma prediktory a zjistíme, jestli by tato proměnná nešla z modelu vypustit.

Model 2 – bez C_org:

Výsledky pro Zn_HNO3:

```
Call:
lm(formula = Zn ~ ph_vym + Zn_HNO3, model = TRUE, x = TRUE, y = TRUE)

Residuals:
   Min     1Q Median     3Q    Max
-36.138  -6.025  -0.578   5.775  31.688

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 58.878805  10.614309   5.547 3.46e-07 ***
ph_vym      -5.066708   1.660981  -3.050  0.00308 **
Zn_HNO3      0.062388   0.006032  10.342  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.529 on 82 degrees of freedom
Multiple R-squared: 0.566,      Adjusted R-squared: 0.5555
F-statistic: 53.48 on 2 and 82 DF,  p-value: 1.365e-15
```

a pro Zn_AR:

```
Call:
lm(formula = Zn ~ ph_vym + Zn_AR, model = TRUE, x = TRUE, y = TRUE)

Residuals:
    Min      1Q  Median      3Q     Max
-36.4612  -6.2977  -0.4891   5.8618  25.5856

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 57.479764  10.628275   5.408 6.15e-07 ***
ph_vym      -5.168240   1.671298  -3.092  0.00271 **
Zn_AR        0.046678   0.004547  10.266 2.23e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.569 on 82 degrees of freedom
Multiple R-squared: 0.5624,     Adjusted R-squared: 0.5518
F-statistic:  52.7 on 2 and 82 DF,  p-value: 1.918e-15
```

Vidíme, že model pouze se dvěma prediktory vysvětluje jenom o 1% méně variability než složitější model. Zkusíme modely porovnat:

```
> anova(model1a,model2a)
Analysis of Variance Table

Model 1: Zn ~ ph_vym + C_org + Zn_HNO3
Model 2: Zn ~ ph_vym + Zn_HNO3
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     81 7262.6
2     82 7446.1 -1   -183.43 2.0458 0.1565
> anova(model1b,model2b)
Analysis of Variance Table
```

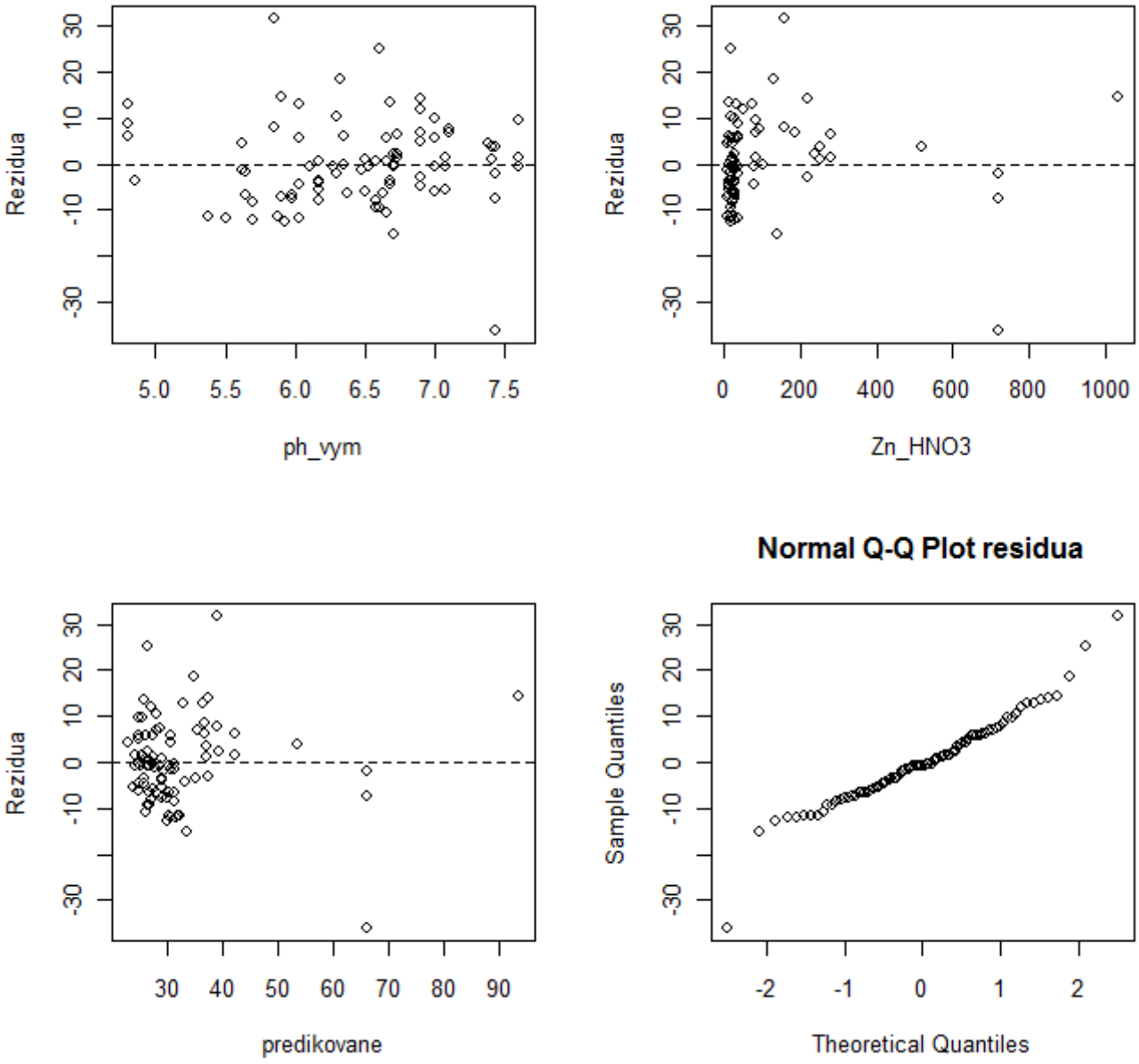
Model 1: Zn ~ ph_vym + C_org + Zn_AR
Model 2: Zn ~ ph_vym + Zn_AR

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	81	7278.3				
2	82	7508.1	-1	-229.81	2.5575	0.1137

V obou případech nezamítáme hypotézu o vhodnosti jednoduššího modelu. Proto budeme dále pracovat s modelem, který obsahuje jenom 2 vysvětlující proměnné.
Pro model 2a, který vyšel nejlépe, provedeme analýzu reziduí.

Analýza reziduí – model 2a

Je potřeba ověřit linearitu reziduí – grafy oproti nezávislým proměnným a homogenitu rozptylu – graf rezidua vs. predikované hodnoty.



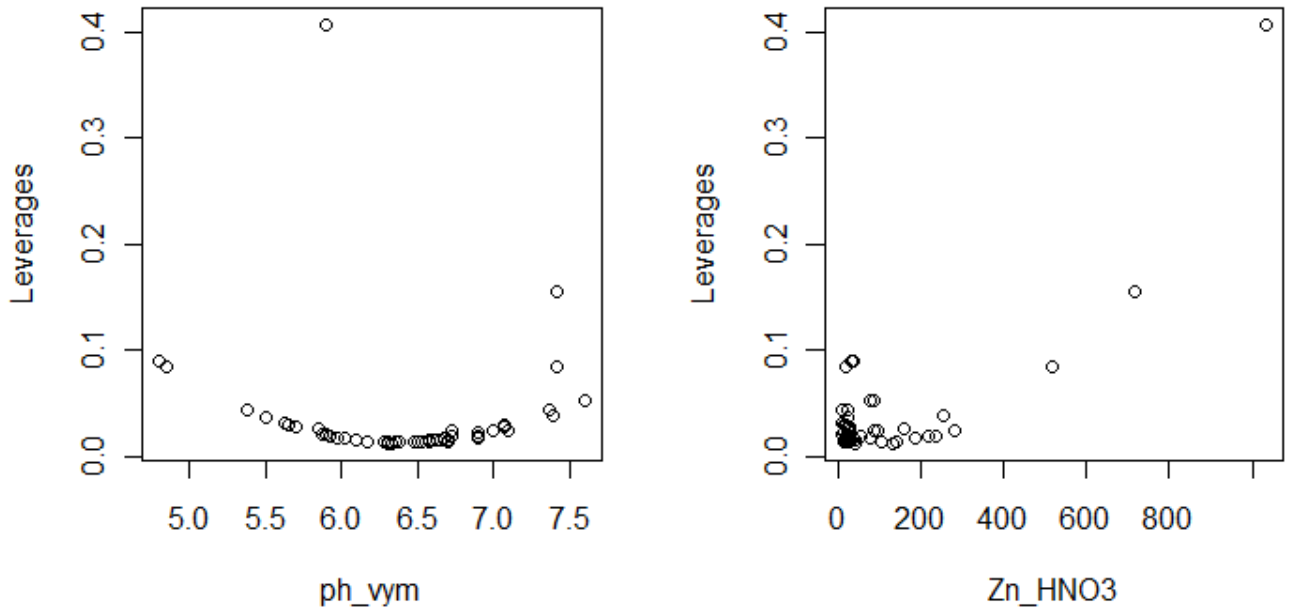
Rezidua jsou rovnoměrně rozptýlena kolem nulové hodnoty. Žádná závislost tam není vidět. U proměnné Zn_HNO3 jsou body převážně kolem nuly a několik odlehlých pozorování nad 400 mg/kg.
NP-plot reziduí nesvědčí o přílišném porušení normality.

K odstranění možných problémů se nabízí možnost data transformovat. Nejprve se ale podíváme na odlehlá a vlivná pozorování.

Hledání zvláštních bodů:

Odlehlá pozorování (extrémní hodnoty závisle proměnné) můžeme vidět na grafech reziduí na předchozí stránce. Vyskytuje se zde jedno výrazně odlehlé pozorování v grafu rezidua vs. Zn_HNO3 (má hodnotu 1034 mg/kg).

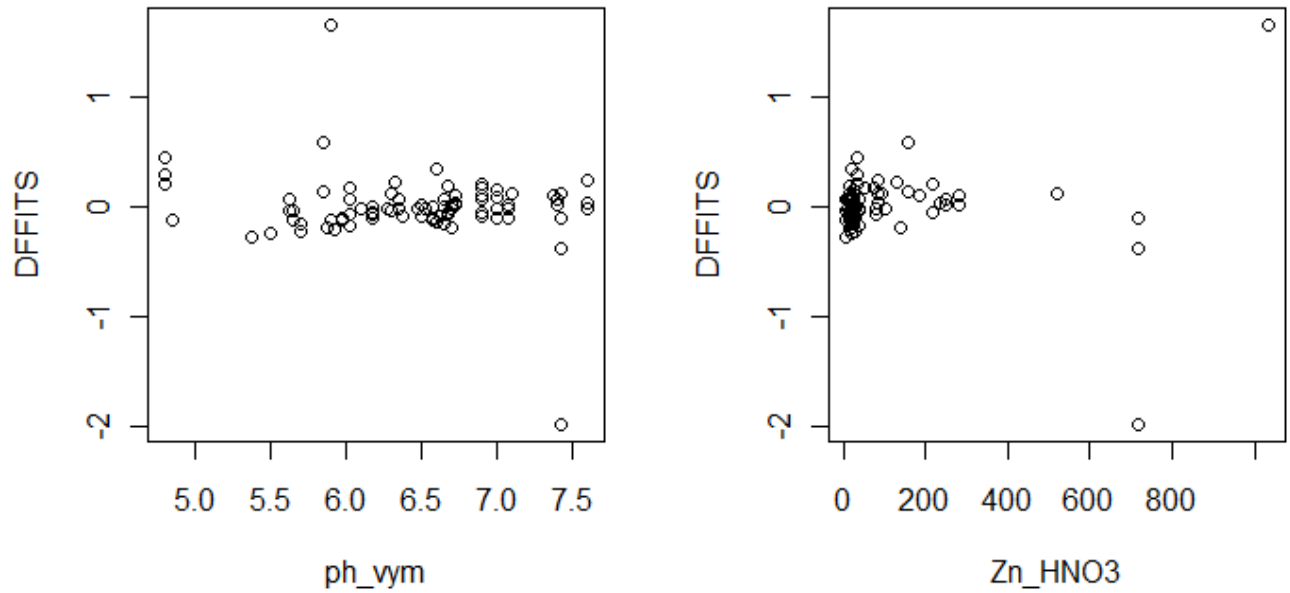
Vlivná pozorování (extrémní hodnoty nezávisle proměnných) se zjišťují pomocí delečních diagnostik. Nejdříve si určíme pákové body:



U obou nezávisle proměnných se vyskytuje vždy jeden velmi odlehlý bod. Konkrétně se jedná o:

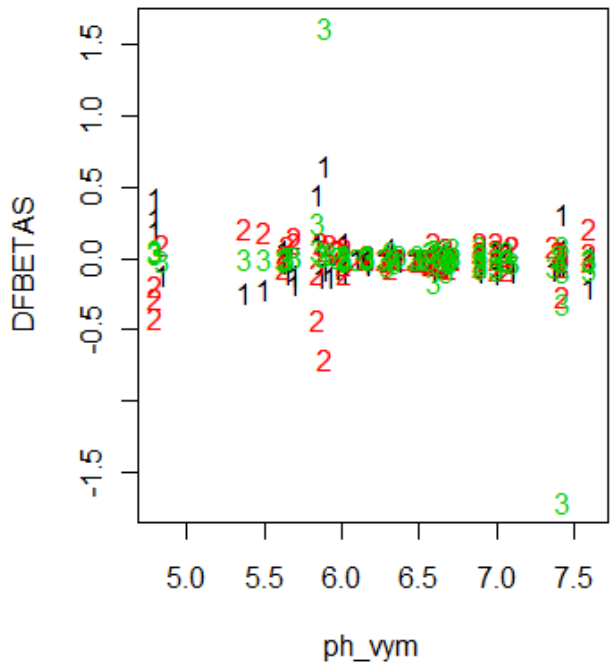
> ph_vym[abs(hatvalues(model2a))>0.4]
[1] 5.9
> Zn_HNO3[abs(hatvalues(model2a))>0.4]
[1] 1034.25

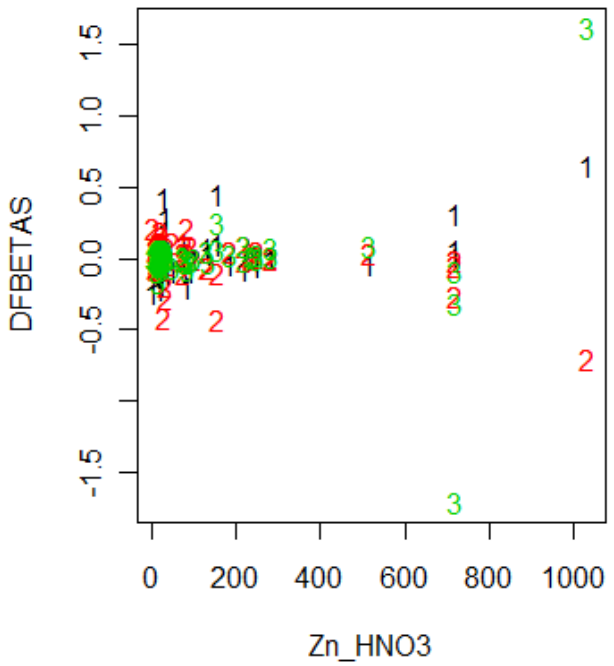
Dále lze použít DFFITS, DFBETAS a Cookovu vzdálenost:



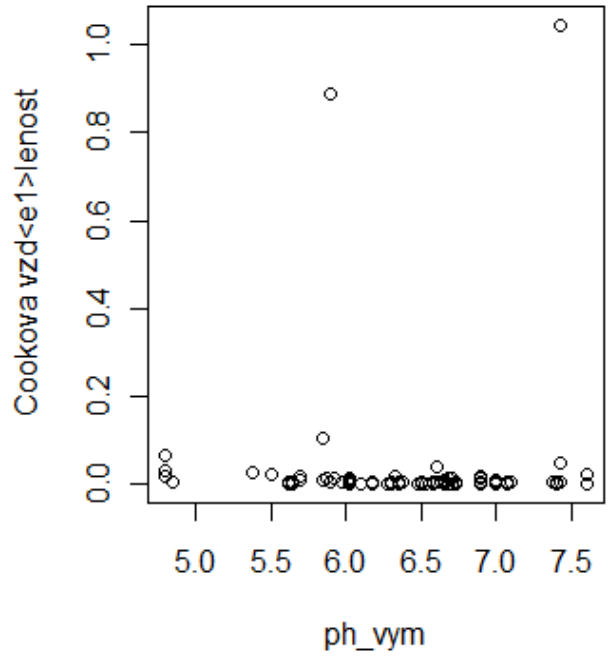
V grafu DFFITS se již objeví 2 vlivné body. Jsou to:

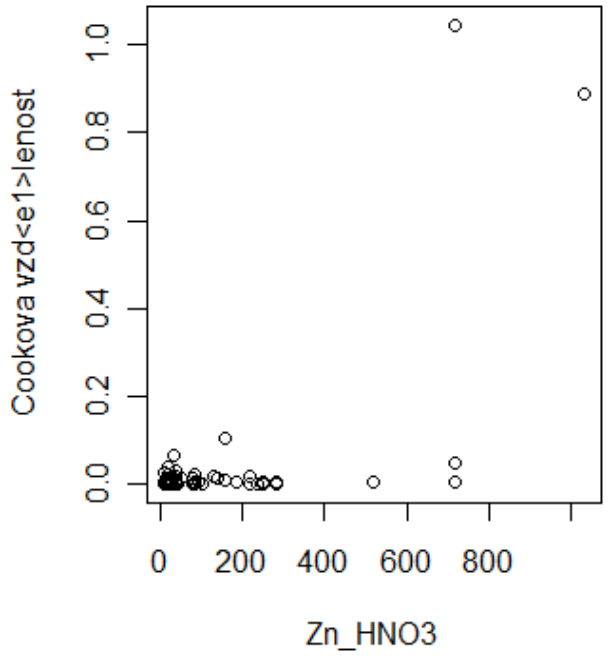
> ph_vym[abs(dffits(model2a))>1]
[1] 5.900 7.425
> Zn_HNO3[abs(dffits(model2a))>1]
[1] 1034.25 717.75





Z grafu DFBETAS je vidět, že „problém“ nastává především ve třetím parametru β_2, který je u proměnné Zn_HNO3.





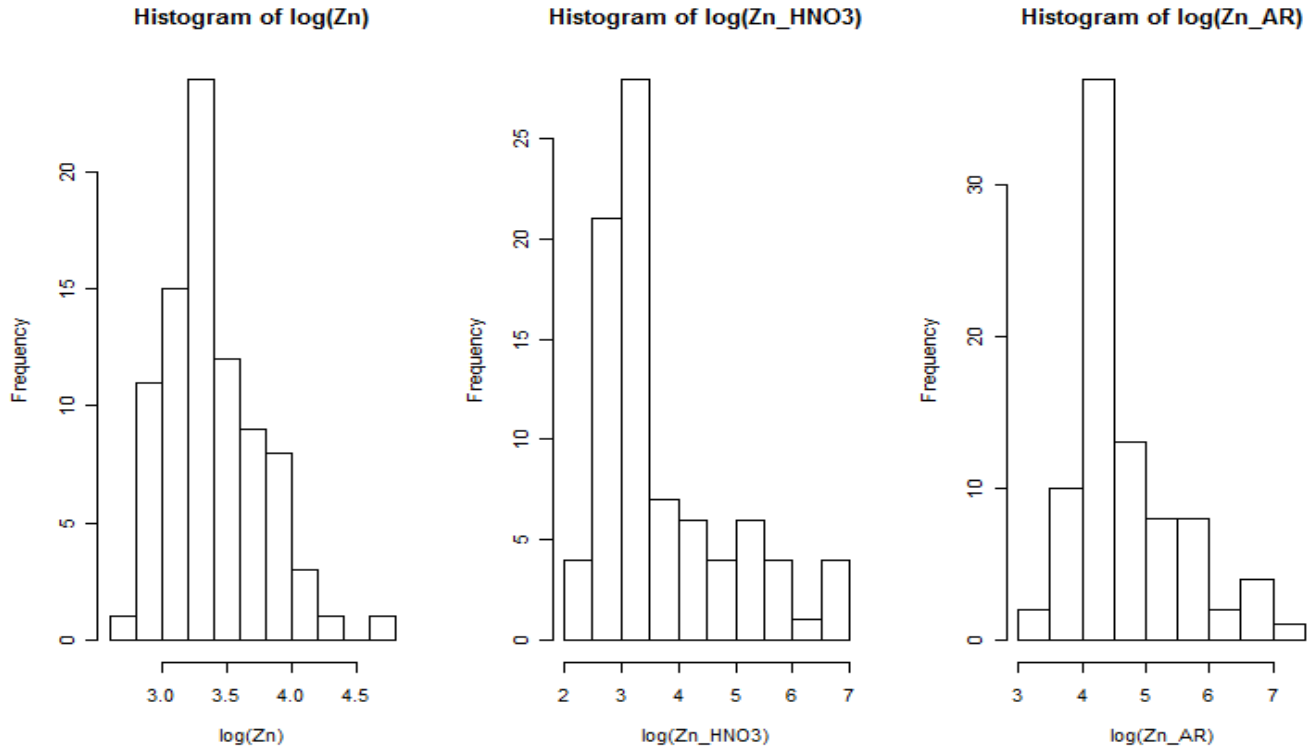
Předchozí závěry potvrzuje i graf Cookových vzdáleností.

Nalezené odlehlé hodnoty by bylo možné odstranit a podívat se, jestli se změní celkový model.

Další možností, jak se vypořádat s nepříliš dobrými výsledky při analýze reziduí a nesymetričností vstupních proměnných je transformace dat. V literatuře se nejvíce pracuje s logaritmickou transformací a to především u koncentrací kovu jak v půdě, tak i v rostlině (např. log(Mplant)=a+b*log(Msoil)+c*pH).

Transformace dat:

Využijeme tedy log(Zn) a log(Zn_HNO3), případně (log(Zn_AR)).
Nejprve se podíváme, jak se změní rozložení hodnot po logaritmické transformaci:



Transformací jsme dosáhli větší symetričnosti dat.
Vytvoříme nové dva modely.

Modely t1 a t2 (všechny proměnné):

Call:
lm(formula = log(Zn) ~ ph_vym + C_org + log(Zn_HNO3), model = TRUE,
x = TRUE, y = TRUE)

Residuals:
Min 1Q Median 3Q Max
-0.72390 -0.17071 0.01623 0.17987 0.75780

Coefficients:
Estimate Std. Error t value Pr(>|t|)
*(Intercept) 3.44261 0.29877 11.523 < 2e-16 ****
*ph_vym -0.15300 0.04880 -3.135 0.00239 ***
C_org 0.04527 0.07586 0.597 0.55239
*log(Zn_HNO3) 0.23644 0.03008 7.860 1.42e-11 ****
---
*Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1*

Residual standard error: 0.2694 on 81 degrees of freedom
Multiple R-squared: 0.491, Adjusted R-squared: 0.4722
F-statistic: 26.05 on 3 and 81 DF, p-value: 6.787e-12

Call:
lm(formula = log(Zn) ~ ph_vym + C_org + log(Zn_AR), model = TRUE,
x = TRUE, y = TRUE)

Residuals:
Min 1Q Median 3Q Max

-0.62480 -0.19185 -0.01207 0.18732 0.75203

Coefficients:
Estimate Std. Error t value Pr(>|t|)
*(Intercept) 2.66615 0.30181 8.834 1.69e-13 ****
*ph_vym -0.12396 0.04784 -2.591 0.0113 **
C_org 0.06906 0.07511 0.919 0.3606
*log(Zn_AR) 0.30614 0.03936 7.778 2.05e-11 ****
---
*Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1*

Residual standard error: 0.2706 on 81 degrees of freedom
Multiple R-squared: 0.4864, Adjusted R-squared: 0.4674
F-statistic: 25.57 on 3 and 81 DF, p-value: 9.741e-12

Podobně jako u předchozích modelů vychází člen C_org nevýznamně. Také se asi o 10 % zhoršilo procento vysvětlené variability.
Zkusíme model jen se dvěma prediktory bez organického uhlíku v půdě.

Modely t3 a t4 (bez C_org):

Call:
lm(formula = log(Zn) ~ ph_vym + log(Zn_HNO3), model = TRUE, x = TRUE,
y = TRUE)

Residuals:
Min 1Q Median 3Q Max
-0.74577 -0.17563 0.01287 0.17436 0.75587

Coefficients:
Estimate Std. Error t value Pr(>|t|)
*(Intercept) 3.48057 0.29077 11.970 < 2e-16 ****
*ph_vym -0.15278 0.04861 -3.143 0.00233 ***
*log(Zn_HNO3) 0.24353 0.02753 8.848 1.45e-13 ****
---
*Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1*

Residual standard error: 0.2683 on 82 degrees of freedom
Multiple R-squared: 0.4888, Adjusted R-squared: 0.4763
F-statistic: 39.2 on 2 and 82 DF, p-value: 1.127e-12

Call:
lm(formula = log(Zn) ~ ph_vym + log(Zn_AR), model = TRUE, x = TRUE,
y = TRUE)

Residuals:
Min 1Q Median 3Q Max
-0.65353 -0.16779 -0.01225 0.20357 0.74806

Coefficients:
Estimate Std. Error t value Pr(>|t|)
*(Intercept) 2.69102 0.30031 8.961 8.61e-14 ****
*ph_vym -0.12180 0.04774 -2.551 0.0126 **
*log(Zn_AR) 0.31926 0.03665 8.711 2.70e-13 ****
---
*Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1*

Residual standard error: 0.2703 on 82 degrees of freedom
Multiple R-squared: 0.4811, Adjusted R-squared: 0.4684
F-statistic: 38.01 on 2 and 82 DF, p-value: 2.088e-12

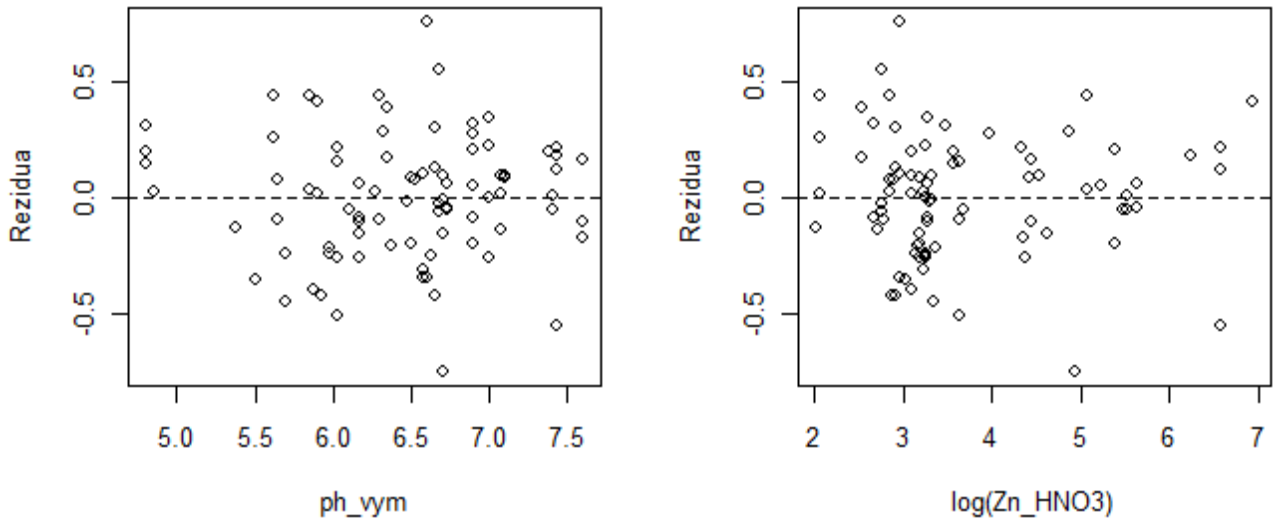
U obou modelů se dokonce mírně zvedla vyčerpaná variabilita. Jednodušší modely budou nejspíš lepší. Ověříme ještě pomocí porovnávání modelů:

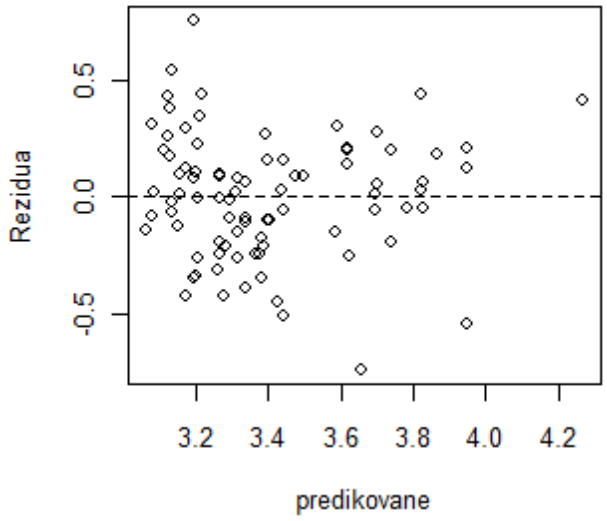
```
> anova(model_t1,model_t3)
Analysis of Variance Table

Model 1: log(Zn) ~ ph_vym + C_org + log(Zn_HNO3)
Model 2: log(Zn) ~ ph_vym + log(Zn_HNO3)
  Res.Df    RSS Df Sum of Sq     F Pr(>F)
1     81 5.8773
2     82 5.9031 -1 -0.025833 0.356 0.5524
> anova(model_t2,model_t4)
Analysis of Variance Table

Model 1: log(Zn) ~ ph_vym + C_org + log(Zn_AR)
Model 2: log(Zn) ~ ph_vym + log(Zn_AR)
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     81 5.9306
2     82 5.9925 -1 -0.061887 0.8452 0.3606
```

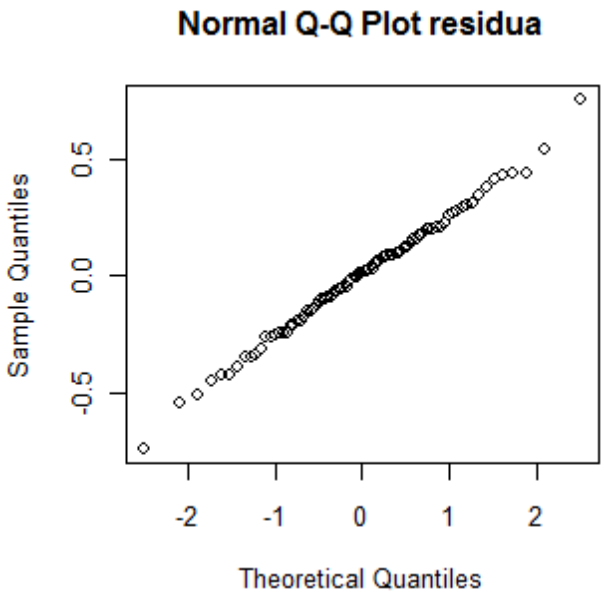
Nezamítáme hypotézu o vhodnosti jednoduššího modelu.

U „nejlepšího“ modelu (t3) můžeme provést analýzu reziduí.

Analýza reziduí – model t3







Je vidět, že díky logaritmické transformaci se zlepšil graf reziduí pro proměnnou Zn_HNO3 a predikované hodnoty. I NP-plot svědčí o normalitě více. Jinak rezidua vypadají velmi podobně jako v předchozím případě.

Závěr:

Zdá se, že model vysvětlující koncentraci Zn v rostlinách pomocí koncentrace Zn v půdě, pH půdy a organického uhlíku v půdě splňuje po logaritmické transformaci koncentrací v rostlinách a v půdě předpoklady pro dobře sestavený model. Jenom procento vysvětlené variability není příliš vysoké, a proto by možná stálo za to hledat ještě nějaké další proměnné, které by také mohly mít na koncentraci kovu v rostlinách vliv.