STAT410 Assignment 3

Ben Murarotto

April 2025

Contents

1 Soil dataset.

Cadmium exposure in pregnant women can lead to increased risk of birth defects in foetuses (Geng & Wang 2019). Diet is one of the main sources of human exposure to Cadmium with rice identified as one of the main contributors. A study1 has been undertaken to explore the components in soil which can be used to predict Cadmium accumulation in Rice. 18 soil samples were collected from different rice fields and a number of factors were recorded from each soil sample. The data set, soil.csv, contains 12 variables: - Soil: Unique identifier from where the soil was collected. This factor should not be used in the modelling. - pH: the pH of the soil. - SOM: the organic matter in the soil (g/kg) - EC: the electrical conductivity of the soil (ms/cm) - Clay: the amount of clay in the soil (g/kg) - Fe: iron in the soil (g/kg) - TN: total nitrogen in the soil (g/kg) - Mn: MnO content (g/kg) - TP: Total phosphorus (g/kg) - CEC: Cation exchange (cmol/kg) - AL: aluminium in the soil (g/kg) - Cd: the amount of cadmium extracted from the rice plants (mg/kg) which is the response variable.

```
soil.df<-read.csv("soil.csv", header=T)

summary(soil.df)
```

```
##      Soil                 pH              EC              SOM
##  Length:18          Min.   :4.250   Min.   :0.0400   Min.   : 6.13
##  Class :character   1st Qu.:5.452   1st Qu.:0.1200   1st Qu.:12.04
##  Mode  :character   Median :5.865   Median :0.1350   Median :22.57
##                     Mean   :6.143   Mean   :0.1639   Mean   :20.98
##                     3rd Qu.:6.800   3rd Qu.:0.2125   3rd Qu.:27.43
##                      Max.   :8.090   Max.   :0.3200   Max.   :33.43  
##        TN               TP              Clay             CEC       
##  Min.   :0.220   Min.   :0.3900   Min.   : 47.00   Min.   : 8.33  
##  1st Qu.:1.210   1st Qu.:0.5925   1st Qu.: 84.03   1st Qu.:13.22  
##  Median :1.605   Median :0.7900   Median :126.40   Median :15.98  
##  Mean   :1.459   Mean   :1.3378   Mean   :153.38   Mean   :18.45  
##  3rd Qu.:1.795   3rd Qu.:1.0575   3rd Qu.:208.18   3rd Qu.:20.86  
##  Max.   :2.540   Max.   :6.4200   Max.   :337.40   Max.   :37.83  
##        Fe               Mn               Al               Cd        
##  Min.   : 3.590   Min.   :0.0300   Min.   : 0.130   Min.   :0.1000  
##  1st Qu.: 7.287   1st Qu.:0.1375   1st Qu.: 1.095   1st Qu.:0.1725  
##  Median :13.290   Median :0.3150   Median : 5.695   Median :0.2050  
##  Mean   :13.076   Mean   :0.3050   Mean   : 4.633   Mean   :0.2217  
##  3rd Qu.:16.035   3rd Qu.:0.4375   3rd Qu.: 6.947   3rd Qu.:0.2775  
##  Max.   :27.450   Max.   :0.8200   Max.   :12.030   Max.   :0.3700  
```

```r
head(soil.df, 3)
```

```
##   Soil   pH   EC   SOM   TN   TP  Clay   CEC    Fe   Mn   Al   Cd
## 1   S1 4.25 0.32 21.84 2.30 0.92 109.1 21.00 27.45 0.11 7.31 0.10
## 2   S2 4.56 0.22 22.57 1.37 0.77 334.0 16.48 21.31 0.03 6.17 0.20
## 3   S3 5.10 0.28 26.90 1.21 0.42 141.6  9.50  3.65 0.06 7.03 0.11
```

1.1 Exploratory analysis

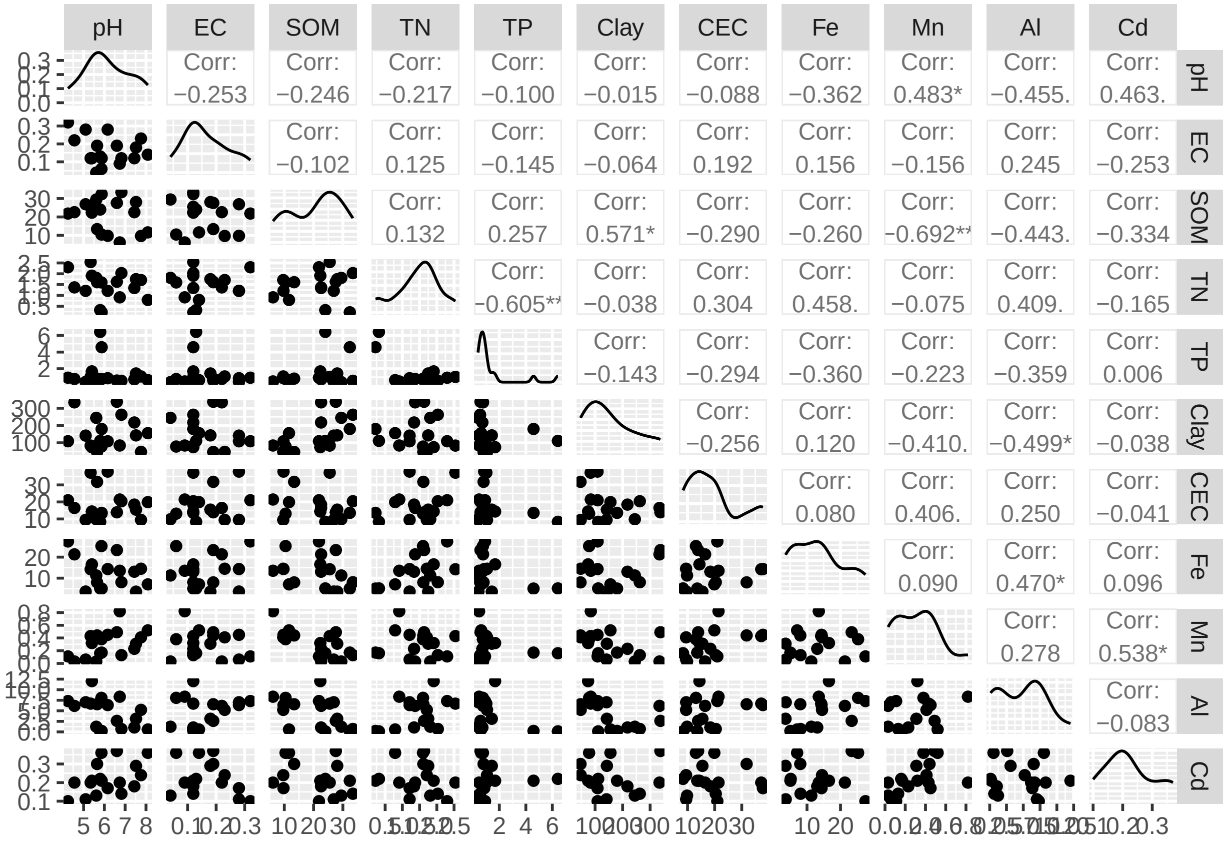
Using the ggpairs() function to plot the data let’s assess correlations between the predictors and the response variable and any correlations between the predictors.

```r
library(GGally)
```

```
## Warning: package 'GGally' was built under R version 4.4.3
```

```r
library(ggplot2)

ggpairs(
  soil.df, columns= 2:12,
  upper = list(continuous = wrap("cor", size = 3, method = "pearson"))
)
```



Assessing the GGpairs plots we notice a few things: 1. Outliers - namely in the TN variable there are two observations that create high leverage when plotted with other variables 2. Correlations - Variables TP and TN have a strong negative correlation (-0.605). The response variable Cd has a moderately strong correlation with Mn (0.538) and pH (0.463). pH seems to also have correlation with multiple heavy metal factors such as Al and Fe. SOM and Mn are strongly negatively correlated (-0.692).

1.2 Fitting main effects.

Here we are going to fit a main effects (first order) model using all the soil factors (excluding the sample identifier, Soil). Using the car library we will check the four indicators of multicollinearity between your predictors.

```
mod1 <- lm(data = soil.df, Cd~pH + EC + SOM + Clay + Fe + TN + Mn + TP + CEC + Al)
summary(mod1)
```

```
##
## Call:
## lm(formula = Cd ~ pH + EC + SOM + Clay + Fe + TN + Mn + TP +
##     CEC + Al, data = soil.df)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.129079 -0.046648 -0.000818  0.045757  0.103180
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.0620412  0.4231262   0.147    0.888
## pH           0.0121652  0.0501118   0.243    0.815
## EC          -0.0659571  0.3400633  -0.194    0.852
## SOM         -0.0025690  0.0078842  -0.326    0.754
## Clay         0.0002815  0.0006573   0.428    0.681
## Fe           0.0008109  0.0068283   0.119    0.909
## TN           0.0313064  0.1006195   0.311    0.765
## Mn           0.2359345  0.2464208   0.957    0.370
## TP           0.0174837  0.0333276   0.525    0.616
## CEC         -0.0019703  0.0040858  -0.482    0.644
## Al          -0.0019112  0.0154341  -0.124    0.905
##
## Residual standard error: 0.09556 on 7 degrees of freedom
## Multiple R-squared:  0.4667, Adjusted R-squared:  -0.2952
## F-statistic: 0.6125 on 10 and 7 DF,  p-value: 0.7676
```

```
library(car)
```

```
## Warning: package 'car' was built under R version 4.4.3
```

```
## Loading required package: carData
```

```
vif(mod1)
```

```
##       pH       EC      SOM     Clay       Fe       TN       Mn       TP
## 5.549991 1.309856 8.556292 6.689521 4.809548 7.179751 4.935959 5.175509
##      CEC       Al
## 2.500158 5.460923
```

Lets remove TP and TN from our model due to their high VIF score and correlation with each other whilst having low correlation with the response. We will also remove SOM and as it has a high VIF. Mn and pH will stay in our final model as it explains a lot of the response and is correlated with other variables which we could remove. Let's fit a second main effects model using Mn, EC, pH and Fe.

```
mod2 <- lm(data=soil.df, Cd ~ Mn + EC + pH + Fe)
summary(mod2)
```

```
##
## Call:
## lm(formula = Cd ~ Mn + EC + pH + Fe, data = soil.df)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.118477 -0.055728 -0.009858  0.053911  0.091159
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.013192   0.149409   0.088    0.931
## Mn           0.132266   0.104959   1.260    0.230
## EC          -0.159661   0.241375  -0.661    0.520
## pH           0.026496   0.021681   1.222    0.243
## Fe           0.002412   0.002773   0.870    0.400
##
## Residual standard error: 0.07473 on 13 degrees of freedom
## Multiple R-squared:  0.3942, Adjusted R-squared:  0.2078
## F-statistic: 2.115 on 4 and 13 DF,  p-value: 0.1371
```

```
vif(mod2)
```

```
##       Mn       EC       pH       Fe 
## 1.464028 1.078887 1.698409 1.297099
```

```
mod3<-lm(data=soil.df, Cd~(Mn + EC + pH + Fe)^2 + I(Mn^2) + I(pH^2) + I(EC^2) + I(Fe^2))
summary(mod3)
```

```
## 
## Call:
## lm(formula = Cd ~ (Mn + EC + pH + Fe)^2 + I(Mn^2) + I(pH^2) + 
##     I(EC^2) + I(Fe^2), data = soil.df)
## 
## Residuals:
##          1          2          3          4          5          6          7 
##  0.0002358 -0.0003785 -0.0001544 -0.0002246 -0.0001331 -0.0003315  0.0007604 
##          8          9         10         11         12         13         14 
##  0.0032272  0.0005865 -0.0035025 -0.0003317 -0.0005003 -0.0001603  0.0024046 
##         15         16         17         18 
## -0.0015940 -0.0013627  0.0008270  0.0006321 
## 
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  1.135e+00  1.445e-01   7.858 0.004294 ** 
## Mn          -6.707e-01  7.289e-02  -9.202 0.002714 ** 
## EC           2.741e+00  1.859e-01  14.744 0.000677 ***
## pH          -2.712e-01  4.559e-02  -5.949 0.009499 ** 
## Fe          -3.779e-02  2.430e-03 -15.547 0.000578 ***
## I(Mn^2)     -2.835e-02  5.845e-02  -0.485 0.660856    
## I(pH^2)      1.747e-02  3.891e-03   4.490 0.020614 *  
## I(EC^2)     -7.626e+00  2.346e-01 -32.501 6.40e-05 ***
## I(Fe^2)      1.359e-03  3.904e-05  34.816 5.21e-05 ***
## Mn:EC        3.879e+00  2.085e-01  18.606 0.000339 ***
## Mn:pH        1.023e-01  1.753e-02   5.833 0.010040 *  
## Mn:Fe       -1.979e-02  2.497e-03  -7.927 0.004185 ** 
## EC:pH       -1.368e-01  2.616e-02  -5.230 0.013604 *  
## EC:Fe       -5.504e-02  2.478e-03 -22.215 0.000200 ***
## pH:Fe        2.895e-03  3.274e-04   8.842 0.003049 ** 
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
## 
## Residual standard error: 0.003447 on 3 degrees of freedom
## Multiple R-squared:  0.9997, Adjusted R-squared:  0.9983 
## F-statistic: 720.1 on 14 and 3 DF,  p-value: 7.516e-05
```

```
anova(mod3)
```

```
## Analysis of Variance Table
## 
## Response: Cd
##           Df   Sum Sq  Mean Sq  F value    Pr(>F)    
## Mn         1 0.034641 0.034641 2914.753 1.400e-05 ***
## EC         1 0.003522 0.003522  296.374 0.0004270 ***
## pH         1 0.004853 0.004853  408.343 0.0002649 ***
## Fe         1 0.004225 0.004225  355.534 0.0003257 ***
## I(Mn^2)    1 0.019112 0.019112 1608.149 3.412e-05 ***
## I(pH^2)    1 0.001363 0.001363  114.665 0.0017412 **
## I(EC^2)    1 0.007616 0.007616  640.865 0.0001352 ***
## I(Fe^2)    1 0.019410 0.019410 1633.169 3.334e-05 ***
## Mn:EC      1 0.007951 0.007951  668.985 0.0001268 ***
## Mn:pH      1 0.007747 0.007747  651.885 0.0001318 ***
## Mn:Fe      1 0.000218 0.000218   18.314 0.0234363 *
## EC:pH      1 0.000265 0.000265   22.282 0.0180087 *
## EC:Fe      1 0.007962 0.007962  669.943 0.0001265 ***
## pH:Fe      1 0.000929 0.000929   78.187 0.0030488 **
## Residuals  3 0.000036 0.000012
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
mod4<-lm(data=soil.df, Cd ~ Mn + EC + pH + Fe + I(EC^2) + I(Fe^2) + Mn:EC + pH:Fe)
summary(mod4)
```

```
##
## Call:
## lm(formula = Cd ~ Mn + EC + pH + Fe + I(EC^2) + I(Fe^2) + Mn:EC +
##     pH:Fe, data = soil.df)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.057908 -0.021535 -0.009867  0.028516  0.069287
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.2342474  0.1698724   1.379  0.20121
## Mn          -0.0804591  0.1236115  -0.651  0.53137
## EC           1.6724728  0.7953042   2.103  0.06481 .
## pH          -0.0005706  0.0232710  -0.025  0.98097
## Fe          -0.0314000  0.0139095  -2.257  0.05039 .
## I(EC^2)     -7.1326921  2.1704245  -3.286  0.00943 **
## I(Fe^2)      0.0011110  0.0002824   3.934  0.00344 **
## Mn:EC        1.9677101  0.8931785   2.203  0.05507 .
## pH:Fe        0.0005519  0.0016661   0.331  0.74806
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0469 on 9 degrees of freedom
## Multiple R-squared:  0.8348, Adjusted R-squared:  0.688
## F-statistic: 5.685 on 8 and 9 DF,  p-value: 0.008773
```

$$\begin{aligned}\hat{Cd} = {} & 0.2342 - 0.0805 \cdot \mathrm{Mn} + 1.6725 \cdot \mathrm{EC} - 0.0006 \cdot \mathrm{pH} - 0.0314 \cdot \mathrm{Fe} \\ & - 7.1327 \cdot \mathrm{EC}^2 + 0.0011 \cdot \mathrm{Fe}^2 \\ & + 1.9677 \cdot (\mathrm{Mn} \cdot \mathrm{EC}) + 0.0006 \cdot (\mathrm{pH} \cdot \mathrm{Fe})\end{aligned}$$

```
formL<-formula(~1, data=soil.df)
formU<-formula(Cd~(Mn + EC + pH + Fe)^2 + I(Mn^2) + I(pH^2) + I(EC^2) + I(Fe^2))
start.mod.b<-mod3
step.mod.b<-step(start.mod.b,
                 direction = "backward",
                 scope = list(lower=formL, upper=formU))
```

```
## Start:  AIC=-206.38
## Cd ~ (Mn + EC + pH + Fe)^2 + I(Mn^2) + I(pH^2) + I(EC^2) + I(Fe^2)
##
##            Df Sum of Sq       RSS     AIC
## - I(Mn^2)  1 0.0000028 0.0000385 -207.02
## <none>                 0.0000357 -206.38
## - I(pH^2)  1 0.0002396 0.0002752 -171.59
## - EC:pH    1 0.0003250 0.0003607 -166.72
## - Mn:pH    1 0.0004043 0.0004400 -163.15
## - Mn:Fe    1 0.0007469 0.0007825 -152.78
## - pH:Fe    1 0.0009292 0.0009649 -149.01
## - Mn:EC    1 0.0041144 0.0041501 -122.75
## - EC:Fe    1 0.0058651 0.0059007 -116.42
## - I(EC^2)  1 0.0125539 0.0125895 -102.78
## - I(Fe^2)  1 0.0144059 0.0144416 -100.30
##
## Step:  AIC=-207.02
## Cd ~ Mn + EC + pH + Fe + I(pH^2) + I(EC^2) + I(Fe^2) + Mn:EC +
##     Mn:pH + Mn:Fe + EC:pH + EC:Fe + pH:Fe
##
##            Df Sum of Sq       RSS      AIC
## <none>                 0.0000385 -207.017
## - EC:pH    1 0.0003303 0.0003687 -168.325
## - Mn:Fe    1 0.0007956 0.0008340 -153.633
## - I(pH^2)  1 0.0009207 0.0009591 -151.118
## - pH:Fe    1 0.0009316 0.0009701 -150.913
## - Mn:pH    1 0.0017685 0.0018070 -139.717
## - EC:Fe    1 0.0084598 0.0084983 -111.849
## - Mn:EC    1 0.0109894 0.0110279 -107.159
## - I(EC^2)  1 0.0190013 0.0190398  -97.329
## - I(Fe^2)  1 0.0243991 0.0244375  -92.836
```

Backward stepwise model selection was performed using the complete second order model for predictors Mn, EC, pH, and Fe. The null model was defined as the intercept only model, and the upper model included all linear, quadratic plus two way interaction terms. Using AIC, one term (I(Mnˆ2)) was removed, improving the AIC from –206.38 to –207.02. The adjusted R squared of the final model remained extremely high, indicating an excellent fit while slightly simplifying the model. The final regression equation includes the effects of Mn, EC, pH, Fe, and their nonlinear and interactions.

```
summary(step.mod.b)
```

```
##
## Call:
## lm(formula = Cd ~ Mn + EC + pH + Fe + I(pH^2) + I(EC^2) + I(Fe^2) +
##     Mn:EC + Mn:pH + Mn:Fe + EC:pH + EC:Fe + pH:Fe, data = soil.df)
##
## Residuals:
##          1          2          3          4          5          6          7
##  0.0002613 -0.0005416 -0.0002815 -0.0005589  0.0005043 -0.0009081  0.0003166
##          8          9         10         11         12         13         14
##  0.0036744  0.0008857 -0.0028725  0.0001490 -0.0009713 -0.0003632  0.0022377
##         15         16         17         18
## -0.0015117 -0.0019355  0.0008846  0.0010308
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.189e+00  8.233e-02  14.444 0.000134 ***
## Mn          -6.575e-01  6.077e-02 -10.818 0.000414 ***
## EC           2.756e+00  1.648e-01  16.718 7.50e-05 ***
## pH          -2.891e-01  2.418e-02 -11.954 0.000281 ***
## Fe          -3.813e-02  2.089e-03 -18.257 5.29e-05 ***
## I(pH^2)      1.904e-02  1.946e-03   9.787 0.000611 ***
## I(EC^2)     -7.691e+00  1.730e-01 -44.460 1.53e-06 ***
## I(Fe^2)      1.371e-03  2.722e-05  50.381 9.29e-07 ***
## Mn:EC        3.958e+00  1.171e-01  33.812 4.56e-06 ***
## Mn:pH        9.463e-02  6.977e-03  13.564 0.000171 ***
## Mn:Fe       -1.942e-02  2.134e-03  -9.097 0.000810 ***
## EC:pH       -1.376e-01  2.348e-02  -5.862 0.004228 **
## EC:Fe       -5.569e-02  1.877e-03 -29.666 7.69e-06 ***
## pH:Fe        2.898e-03  2.944e-04   9.845 0.000597 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.0031 on 4 degrees of freedom
## Multiple R-squared:  0.9997, Adjusted R-squared:  0.9986
## F-statistic: 958.8 on 13 and 4 DF,  p-value: 2.506e-06
```

$$
\begin{aligned}
\hat{Cd} = {} & 0.2342 - 0.0805 \cdot \mathrm{Mn} + 1.6725 \cdot \mathrm{EC} - 0.0006 \cdot \mathrm{pH} - 0.0314 \cdot \mathrm{Fe} \\
& - 7.1327 \cdot \mathrm{EC}^2 + 0.0011 \cdot \mathrm{Fe}^2 \\
& + 1.9677 \cdot (\mathrm{Mn} \cdot \mathrm{EC}) + 0.0006 \cdot (\mathrm{pH} \cdot \mathrm{Fe})
\end{aligned}
$$

1.3 Stepwise model versus simplified second order model.

```
summary(mod4)
```

```
##
## Call:
## lm(formula = Cd ~ Mn + EC + pH + Fe + I(EC^2) + I(Fe^2) + Mn:EC +
##     pH:Fe, data = soil.df)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.057908 -0.021535 -0.009867  0.028516  0.069287
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.2342474  0.1698724   1.379  0.20121
## Mn          -0.0804591  0.1236115  -0.651  0.53137
## EC           1.6724728  0.7953042   2.103  0.06481 .
## pH          -0.0005706  0.0232710  -0.025  0.98097
## Fe          -0.0314000  0.0139095  -2.257  0.05039 .
## I(EC^2)     -7.1326921  2.1704245  -3.286  0.00943 **
## I(Fe^2)      0.0011110  0.0002824   3.934  0.00344 **
## Mn:EC        1.9677101  0.8931785   2.203  0.05507 .
## pH:Fe        0.0005519  0.0016661   0.331  0.74806
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
```

```
## Residual standard error: 0.0469 on 9 degrees of freedom
## Multiple R-squared:  0.8348, Adjusted R-squared:  0.688
## F-statistic: 5.685 on 8 and 9 DF,  p-value: 0.008773
```

```
cat("AIC simplified second order: ", AIC(mod4),"\n")
```

```
## AIC simplified second order:  -51.54279
```

```
cat("Adj R squared simp second order: ", summary(mod4)$adj.r.squared,"\n")
```

```
## Adj R squared simp second order:  0.687953
```

```
summary(step.mod.b)
```

```
##
## Call:
## lm(formula = Cd ~ Mn + EC + pH + Fe + I(pH^2) + I(EC^2) + I(Fe^2) +
##     Mn:EC + Mn:pH + Mn:Fe + EC:pH + EC:Fe + pH:Fe, data = soil.df)
##
## Residuals:
##          1          2          3          4          5          6          7
##  0.0002613 -0.0005416 -0.0002815 -0.0005589  0.0005043 -0.0009081  0.0003166
##          8          9         10         11         12         13         14
##  0.0036744  0.0008857 -0.0028725  0.0001490 -0.0009713 -0.0003632  0.0022377
##         15         16         17         18
## -0.0015117 -0.0019355  0.0008846  0.0010308
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.189e+00  8.233e-02  14.444 0.000134 ***
## Mn          -6.575e-01  6.077e-02 -10.818 0.000414 ***
## EC           2.756e+00  1.648e-01  16.718 7.50e-05 ***
## pH          -2.891e-01  2.418e-02 -11.954 0.000281 ***
## Fe          -3.813e-02  2.089e-03 -18.257 5.29e-05 ***
## I(pH^2)      1.904e-02  1.946e-03   9.787 0.000611 ***
## I(EC^2)     -7.691e+00  1.730e-01 -44.460 1.53e-06 ***
## I(Fe^2)      1.371e-03  2.722e-05  50.381 9.29e-07 ***
## Mn:EC        3.958e+00  1.171e-01  33.812 4.56e-06 ***
## Mn:pH        9.463e-02  6.977e-03  13.564 0.000171 ***
## Mn:Fe       -1.942e-02  2.134e-03  -9.097 0.000810 ***
## EC:pH       -1.376e-01  2.348e-02  -5.862 0.004228 **
## EC:Fe       -5.569e-02  1.877e-03 -29.666 7.69e-06 ***
## pH:Fe        2.898e-03  2.944e-04   9.845 0.000597 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.0031 on 4 degrees of freedom
## Multiple R-squared:  0.9997, Adjusted R-squared:  0.9986
## F-statistic: 958.8 on 13 and 4 DF,  p-value: 2.506e-06
```

```
cat("AIC stepwise: ", AIC(step.mod.b), "\n")
```

```
## AIC stepwise:  -153.9356
```

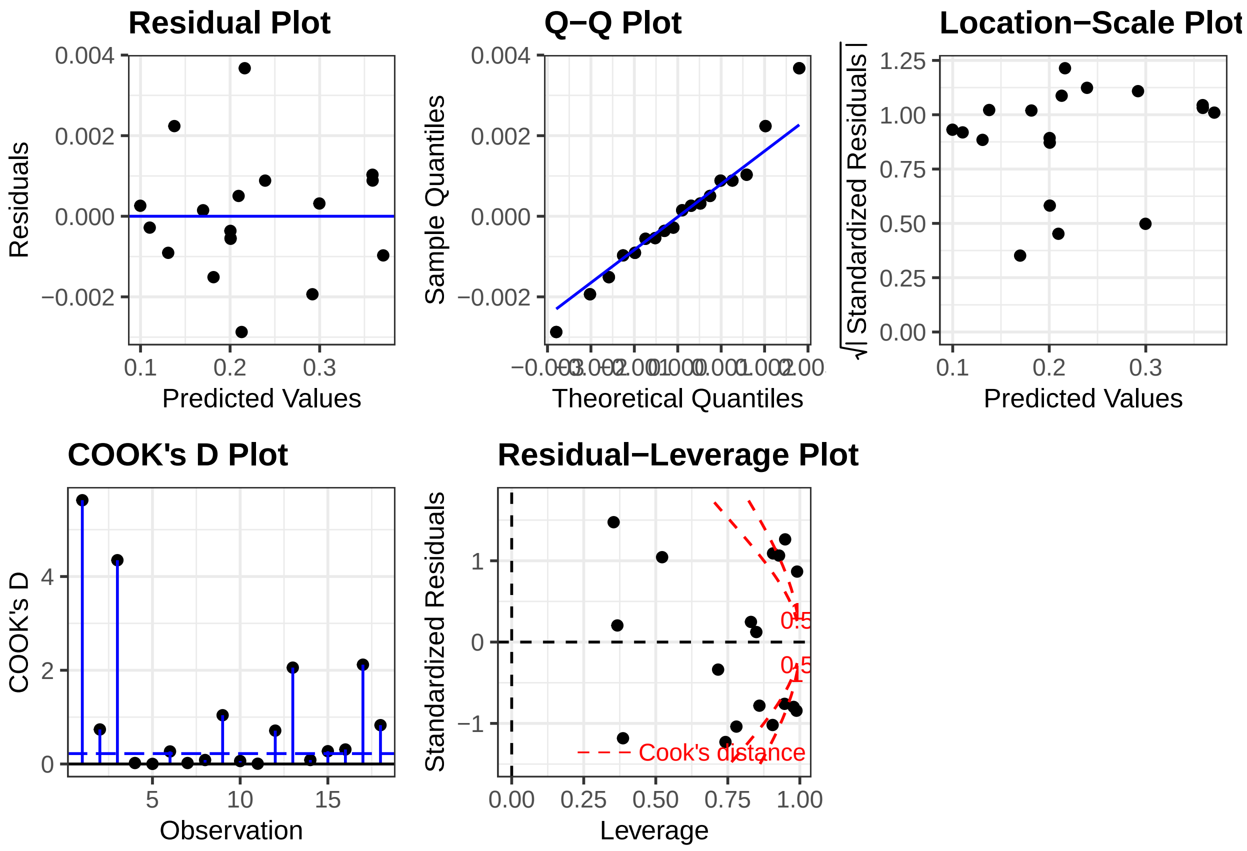
```
cat("Adj R squared stepwise: ", summary(step.mod.b)$adj.r.squared, "\n")
```

```
## Adj R squared stepwise:  0.9986365
```

To determine the best model for predicting cadmium concentration in rice, both the simplified second-order model and the stepwise-selected model were evaluated. The stepwise model had a significantly lower AIC (–153.94 vs –51.54), a much higher adjusted R squared (0.9986 vs 0.688). Additionally, all 13 terms in the stepwise model were highly statistically significant ($p < 0.01$), compared to only 3–4 in the simplified model. Therefore, the stepwise model was selected as the final model due to its superior statistical performance and predictive accuracy.

1.4 Assessing model assumptions

```
library(ggResidpanel)
resid_panel(step.mod.b, plots=c("resid","qq","ls","cookd","lev"))
```



Residual diagnostics were conducted to assess whether the 5 assumptions of multiple linear regression were met for the final model. The residuals fitted plot shows no clear patterns, indicating that the assumptions of linearity and constant variance are satisfied. The Normal Q-Q plot demonstrates that the residuals were approximately normally distributed. The scale-location plot has a dip in variability toward the center indicative of a U-shape. This gives evidence AGAINST homoscedasticity. The residuals vs leverage plot revealed a cluster of observations with very high leverage, some exceeding 0.9. These points lie close to or beyond the Cook's Distance threshold lines, indicating they may be exerting significant influence our coefficients.

1.5 Model summary

The final multiple linear regression model selected through backward stepwise selection achieving an adjusted R squared of 0.9986 and a highly significant overall p-value ($p < 0.001$). While the model fits the data extremely well, residual diagnostics revealed potential violations of the constant variance assumption and the presence of influential observations. Therefore, while the model demonstrates strong predictive power within the current dataset, caution should be taken when generalising to new data and making predictions.

2 Cadmium Dataset

Rhizospheric soil microbes can have profound effects on plants in Cadmium contaminated soils. Abundance of saprotrophic soil fungi have been found to reduce cadmium accumulation in plant tissues (Cakmak et al., 2023). Wang et al. (2024) conducted an experiment to determine if the soil fungi, Basidiomycota, affected cadmium accumulation in a cadmium hyperaccumulator plant, Black-jack (Bidens Pilosa).

The dataset Cadmium.csv, contains 3 Variables: - Shoot_Cd: Cadmium concentration in the stems and leaves of Bidens Pilosa (mg/kg) - Soil_Cd: Cadmium concentration in the soil (mg/kg) - Basid: Relative abundance of the soil fungus, Basidiomycota (%)

```
cad.df <- read.csv("Cadmium.csv", header=T)
head(cad.df)
```

```
##   Soil_Cd Shoot_Cd Basidiomycota
## 1    2.66     8.64          2.44
## 2    3.53    17.69          1.62
## 3    2.88    13.85         14.74
## 4    4.65     9.87          1.42
## 5    4.71    10.51          1.29
## 6    4.47    12.53          2.14
```

2.1 Fitting second order.

We will fit a second order model and check the summary.

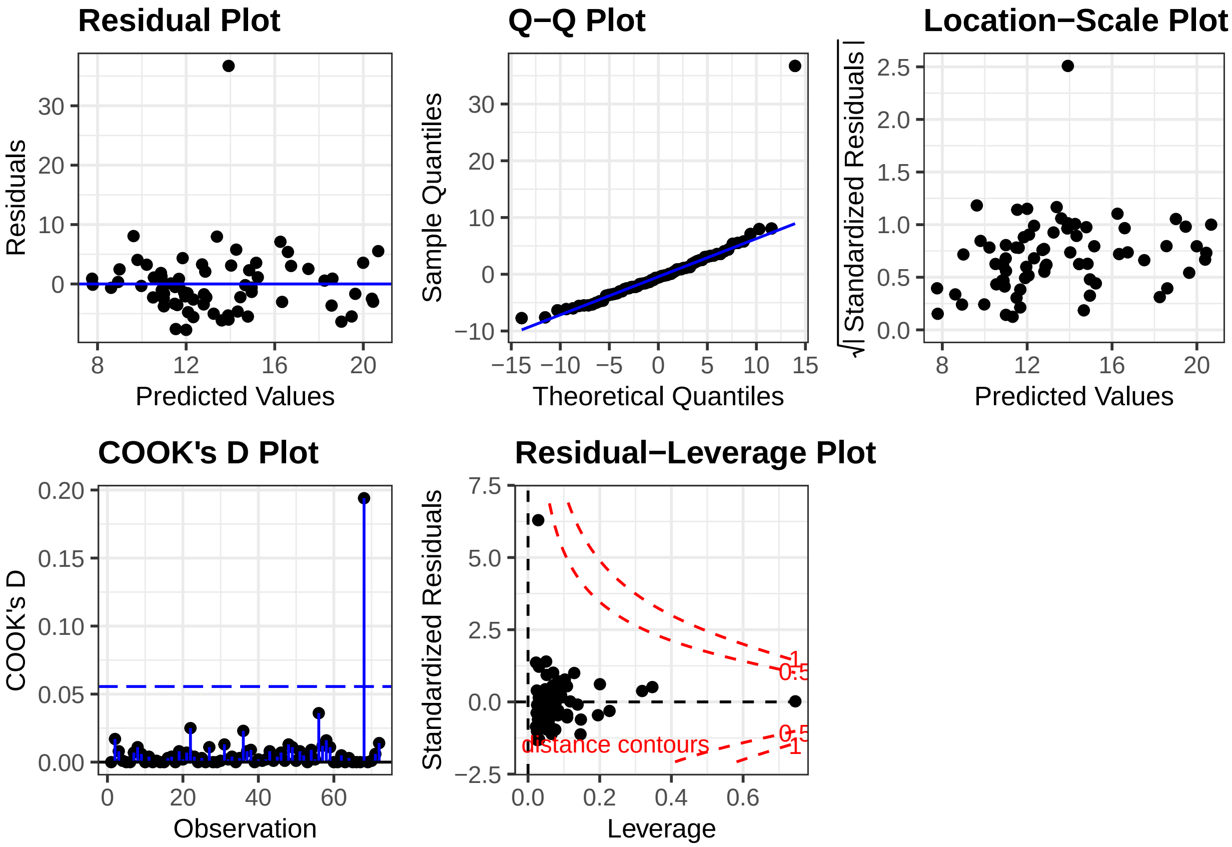
```
mod1.1<-lm(data = cad.df, Shoot_Cd ~ Soil_Cd * Basidiomycota + I(Soil_Cd^2) + I(Basidiomycota^2))
summary(mod1.1)
```

```
##
## Call:
## lm(formula = Shoot_Cd ~ Soil_Cd * Basidiomycota + I(Soil_Cd^2) +
##     I(Basidiomycota^2), data = cad.df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -7.730 -3.001 -0.564  2.132 36.711
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)            0.028931   3.599494   0.008 0.993611
## Soil_Cd                3.026579   0.873059   3.467 0.000932 ***
## Basidiomycota          0.274271   0.202118   1.357 0.179409
## I(Soil_Cd^2)          -0.106252   0.051745  -2.053 0.044003 *
## I(Basidiomycota^2)    -0.002885   0.004078  -0.707 0.481887
## Soil_Cd:Basidiomycota -0.034790   0.019593  -1.776 0.080392 .
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 5.919 on 66 degrees of freedom
## Multiple R-squared:  0.243,  Adjusted R-squared:  0.1856
## F-statistic: 4.237 on 5 and 66 DF,  p-value: 0.002116
```

Here we see that the model is significant but the adjusted R-squared is low meaning our model may not be a strong candidate for prediction. Now we will check the assumptions of our model.

2.2 Assessing the assumptions of linear regression.

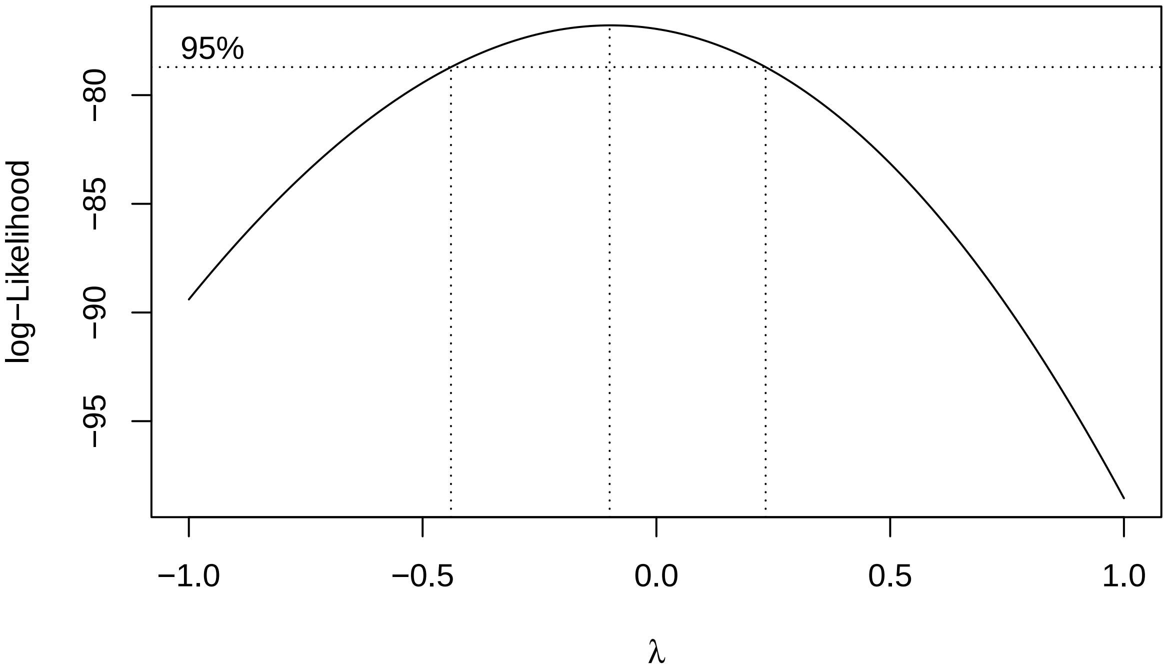
```r
resid_panel(mod1.1, plots=c("resid","qq","ls","cookd","lev"))
```



The residuals fitted plot suggested that the linearity assumption was generally upheld; however, one extreme outlier with a large positive residual was present. The QQ plot revealed significant deviation from the theoretical line in both tails, particularly the upper tail. The outlier also appeared here in the Scale Location plot, impacting the spread and slightly undermining the homoscedasticity assumption. The Cook's Distance plot confirmed that the single observation had a markedly higher Cook's D (~0.2).

2.3 Applying a Box-Cox transformation.

```r
library(MASS)
cad.bc <- boxcox(mod1.1, lambda = seq(-1, 1, 0.01))
```



```
lambda_val <- cad.bc$x[which.max(cad.bc$y)]
lambda_val
```

```
## [1] -0.1
```

Since lambda = -0.1 which is closest to zero we do a log transformation. We do not choose other transformations because they fell outside of the 95% CI of the box cox profile. Applying the wrong transformation would not normalise the distribution of the residuals and create more issues for our assumptions.