Final Project, Tenax Research

Yucheng Shao

Due 12/16/2020

Introduction:

The purpose of this analysis is to find a reasonable model for predicting concentration of some organic chemicals exist in single cell organisms (bioavailibility) using the amount of the chemical found on Tenax (a sorbent material) via single point extraction.
All the measurements with no data for the dependent variable have been removed.
The dependent variable is the concentration of chemical on the organism (corganism).
The potential predictors are:
type of chemical (chemical)
molecular weight (MW)
shape of molecule (planner)
concentration of chemical on Tenax (ctenax)
mass of chemical on Tenax (mtenax)
type of organism (organism)
research origin of data (research)
proportion of organic carbon in sediment (OC)
type of sediment (tsed)
The assumptions are:

The system has reached equilibrium.
The total mass of organism tissue present are equal.
The unknown sediment types are assumed to be the same.
I thought the concentration should be less than 1.

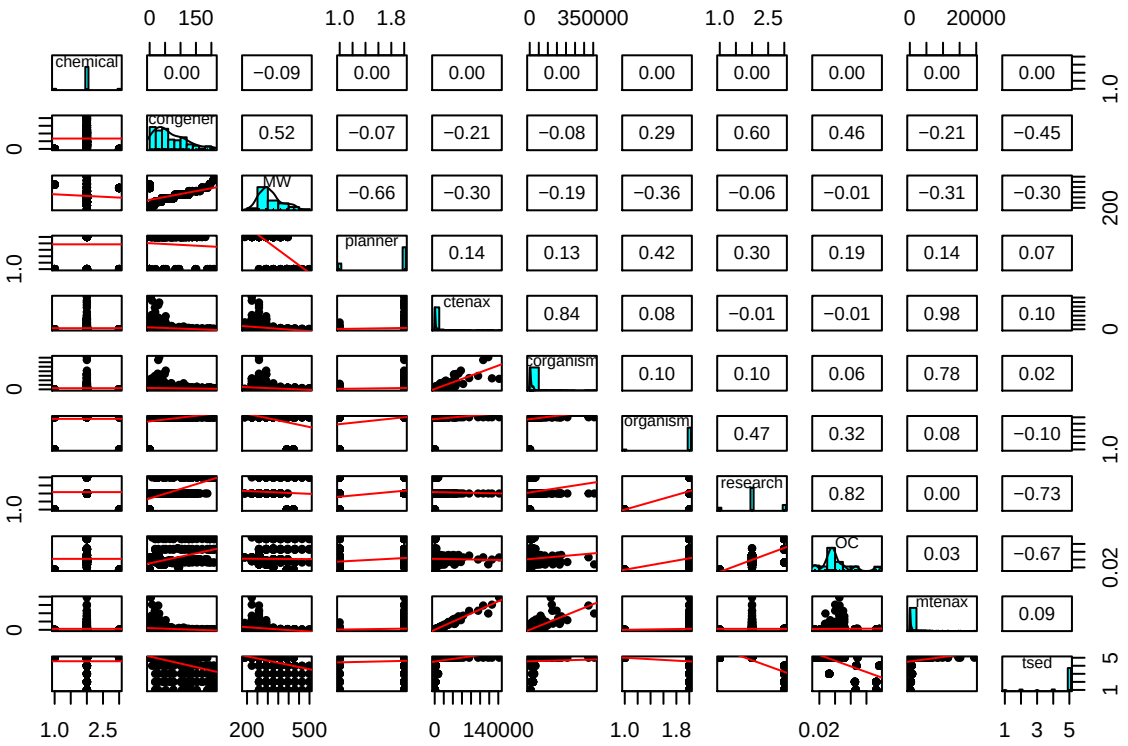
However, the concentration is in total chemical captured (ng) over mass of organic carbon (g), which can well be more than 1.
Therefore, I will not be using logistic regression here.

Mutilple linear regression

First, I will try a linear regression.
Noted that the mass of chemical on Tenax is just mutiplying the concentration of chemical on Tenax by the proportion of organic carbon.
Because the total mass on Tenax may be a better representation of total chemical present.
I will compare models and decide which one to include.

Quick look at the variables and a simple model

First I would like to see possible correlation:



Try no interaction model first.

```
##
## Call:
## lm(formula = corganism ~ chemical + MW + planner + ctenax + organism +
##     research + OC + tsed, data = fp_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -154586   -3268    -756    1193  125499
##
## Coefficients: (2 not defined because of singularities)
##                      Estimate Std. Error t value Pr(>|t|)
```

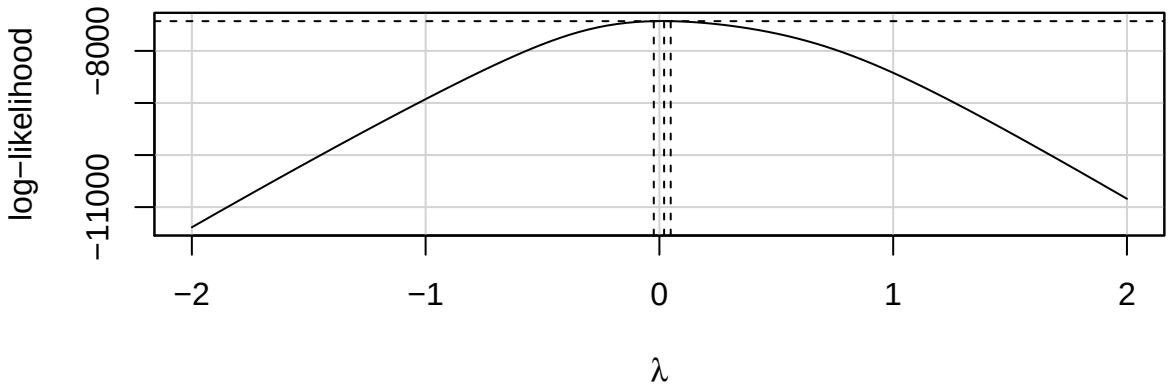
```
## (Intercept)        -2.014e+04  8.020e+03  -2.512  0.01225 *
## chemicalPCB         1.632e+04  6.716e+03   2.430  0.01535 *
## chemicalPermethrin  1.820e+03  3.893e+03   0.468  0.64021
## MW                  5.520e+01  1.679e+01   3.287  0.00107 **
## plannert            3.847e+03  2.091e+03   1.840  0.06628 .
## ctenax              1.925e+00  4.824e-02  39.917  < 2e-16 ***
## organismLV          5.354e+01  3.856e+03   0.014  0.98893
## researchmackenbach -5.211e+03  4.710e+03  -1.106  0.26896
## researchsinche             NA         NA      NA       NA
## OC                 -2.191e+05  7.456e+04  -2.939  0.00341 **
## tsedHumic Acid      1.061e+04  4.083e+03   2.599  0.00957 **
## tsedLPH            -2.355e+03  5.926e+03  -0.397  0.69119
## tsedSaw Dust        1.648e+04  4.054e+03   4.064 5.41e-05 ***
## tsedunknown                NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15430 on 642 degrees of freedom
## Multiple R-squared:  0.7282, Adjusted R-squared:  0.7236
## F-statistic: 156.4 on 11 and 642 DF,  p-value: < 2.2e-16

##
## Call:
## lm(formula = corganism ~ chemical + MW + planner + mtenax + organism +
##     research + OC + tsed, data = fp_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -162415   -3971   -1647     941  207532
##
## Coefficients: (2 not defined because of singularities)
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)        -1.537e+04  9.194e+03  -1.672 0.095006 .
## chemicalPCB         2.577e+04  7.730e+03   3.333 0.000908 ***
## chemicalPermethrin  1.752e+03  4.463e+03   0.392 0.694858
## MW                  4.935e+01  1.930e+01   2.557 0.010774 *
## plannert            3.775e+03  2.398e+03   1.574 0.115969
## mtenax              1.446e+01  4.446e-01  32.533  < 2e-16 ***
## organismLV          5.354e+01  4.422e+03   0.012 0.990342
## researchmackenbach -1.065e+04  5.411e+03  -1.967 0.049571 *
## researchsinche             NA         NA      NA       NA
## OC                 -3.757e+05  8.596e+04  -4.370 1.45e-05 ***
## tsedHumic Acid      9.549e+03  4.681e+03   2.040 0.041777 *
## tsedLPH            -6.528e+03  6.809e+03  -0.959 0.338052
## tsedSaw Dust        1.510e+04  4.648e+03   3.250 0.001216 **
## tsedunknown                NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17690 on 642 degrees of freedom
## Multiple R-squared:  0.6427, Adjusted R-squared:  0.6366
## F-statistic:   105 on 11 and 642 DF,  p-value: < 2.2e-16
```

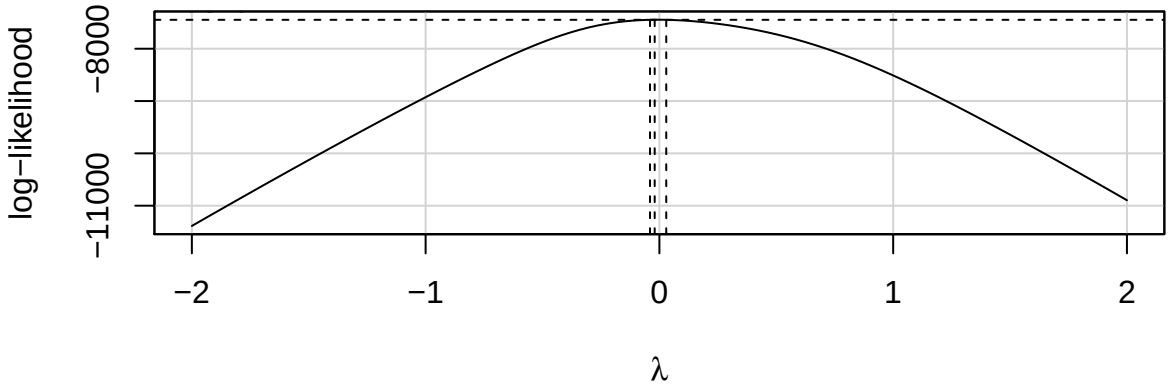
It would appear that concentration on tenax (or mass on tenax) of the chemical, type of chemical, molecular

weight, and type of sediment are likely significant predictors.
The NA's may be singularities where the variables are not linearly independent.
I will look at them and see if the problem persists after the model selection.

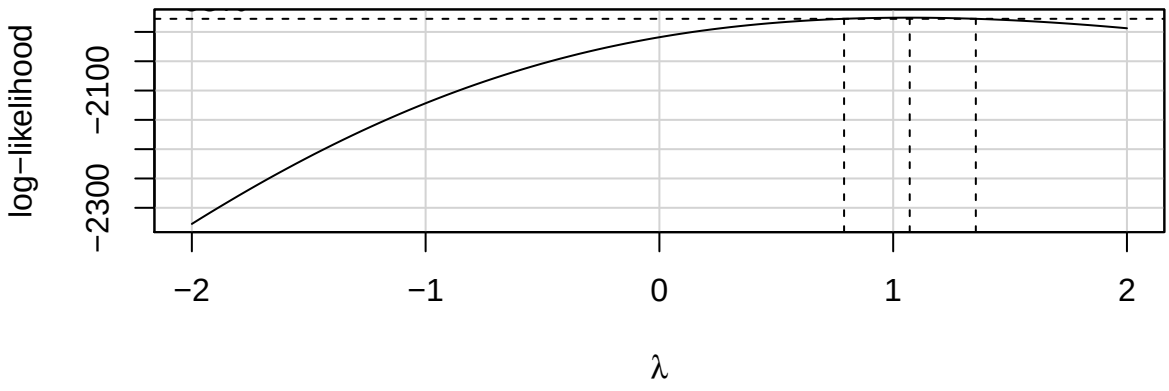
Check the assumptions for linear model

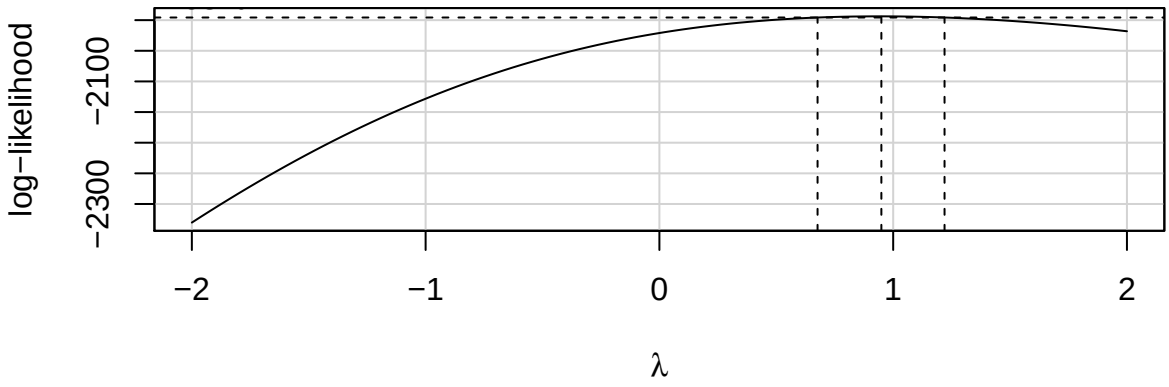
Check the boxCox to decide if a transformation on y is needed:





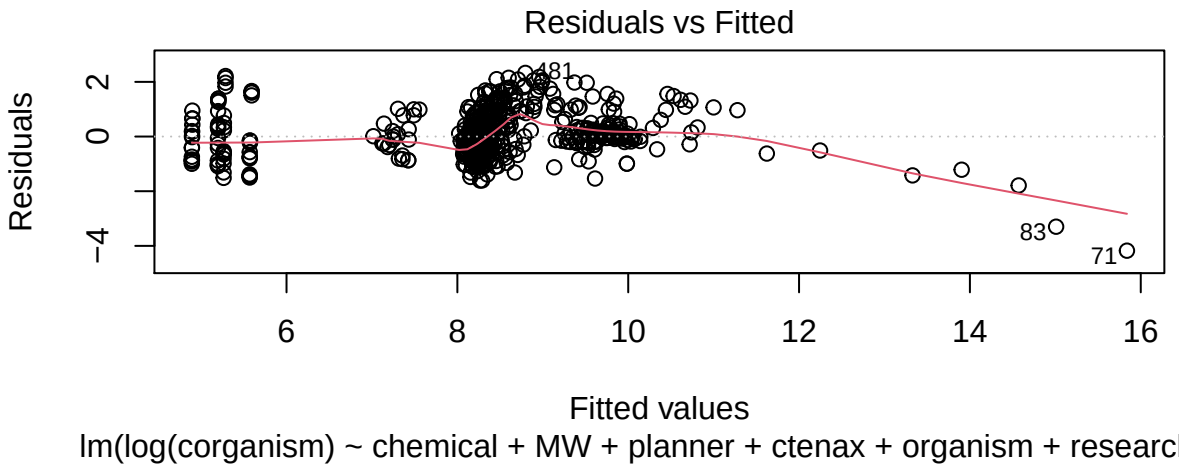
It appears that the variance of the residuals is not constant (1 is not inside the interval).
Take a natural log transformation ($\hat{y} = \ln(y)$):

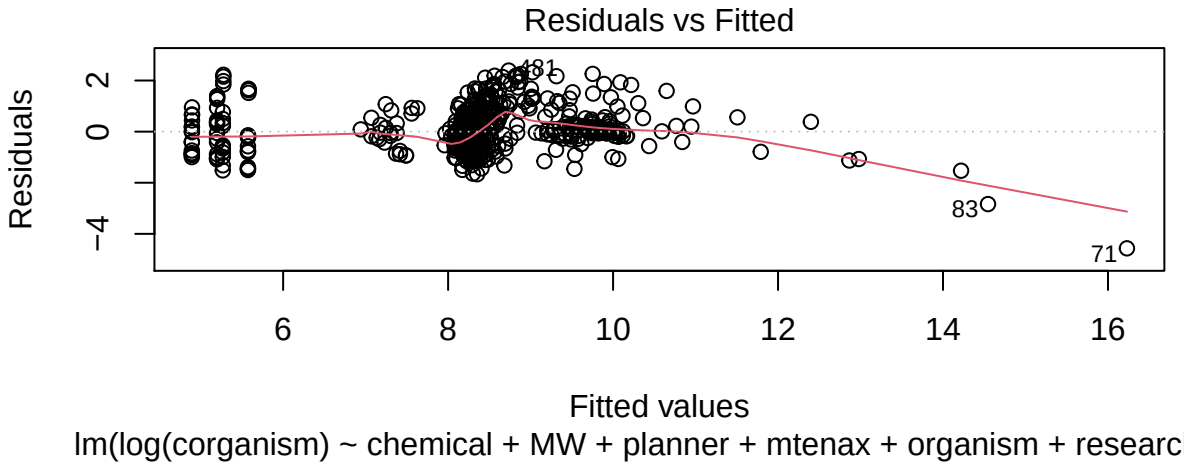




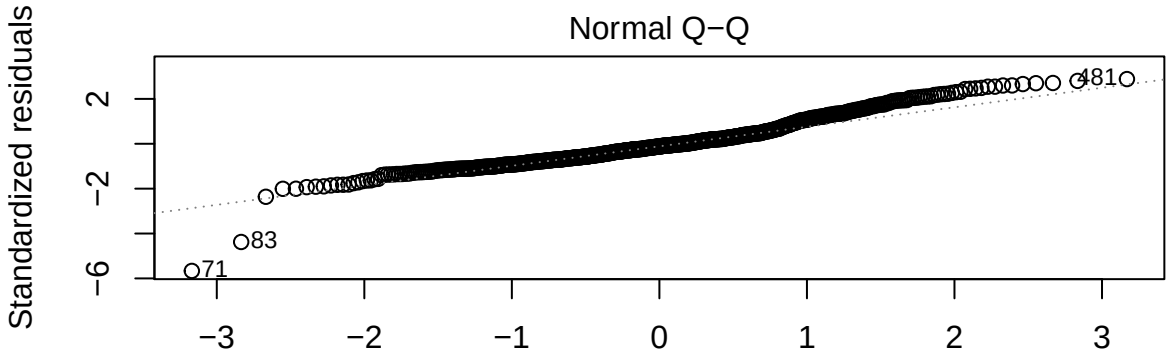
It appears that the transformation does fix the problem.

Now, check the residual plots.

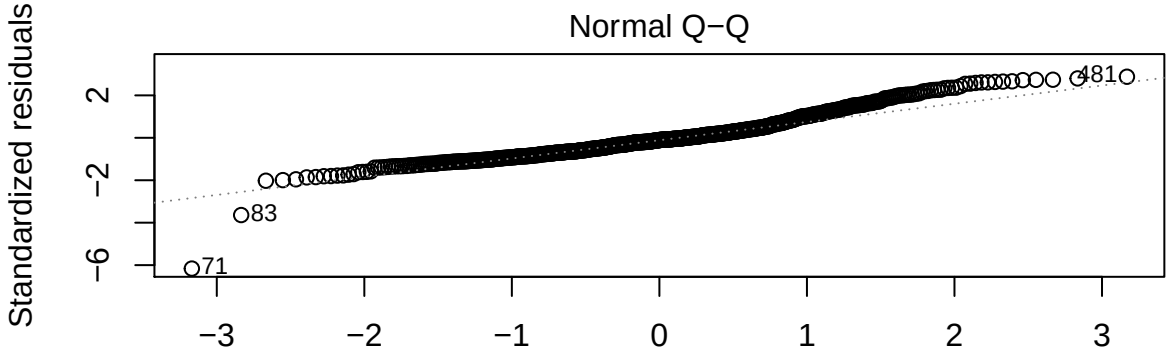




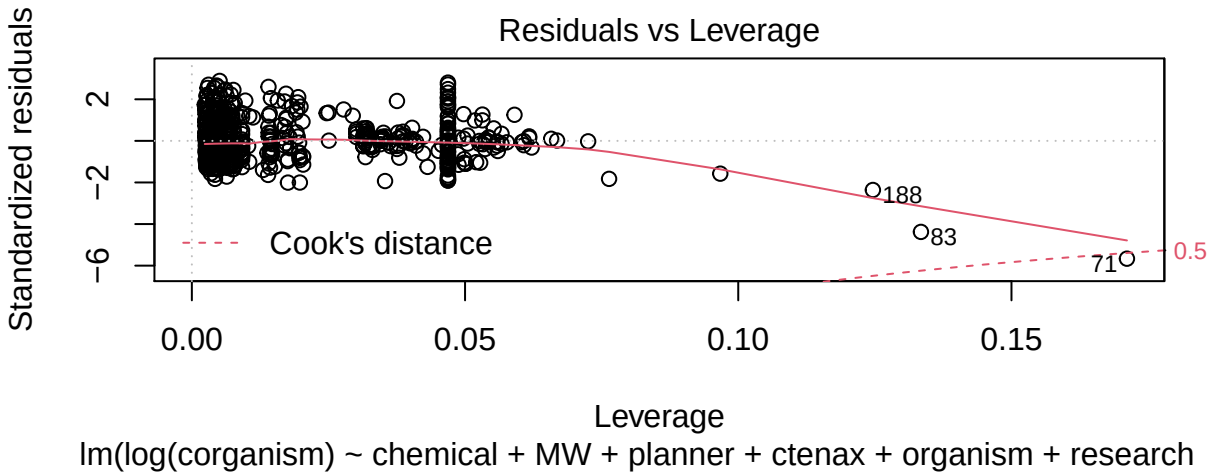
This seems terrible, but I do not know a good way to fix this.
I think there might be something going on with measurement 83 and 71.
Let's check the other plots and the cook's distance.

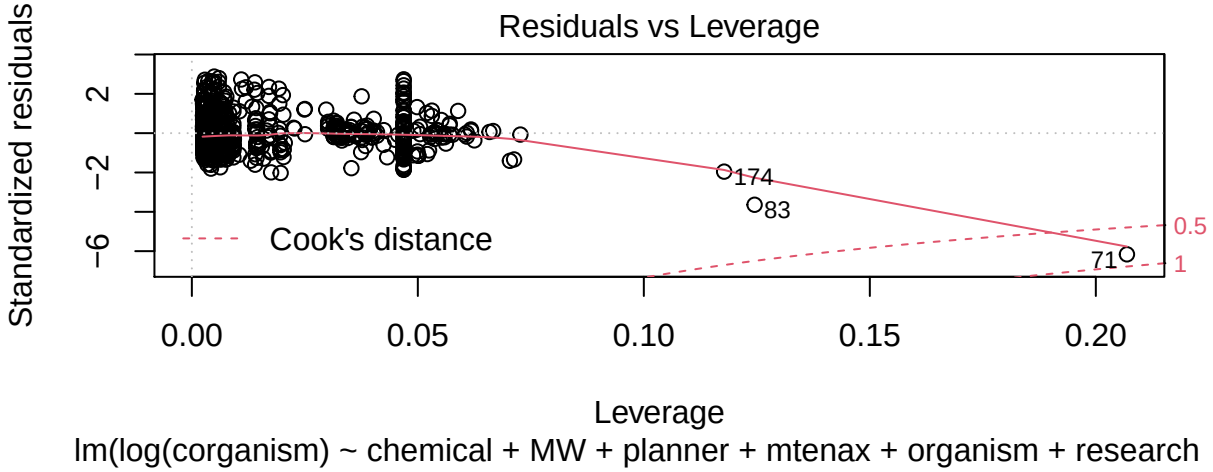


lm(log(corganism) ~ chemical + MW + planner + ctenax + organism + research ...

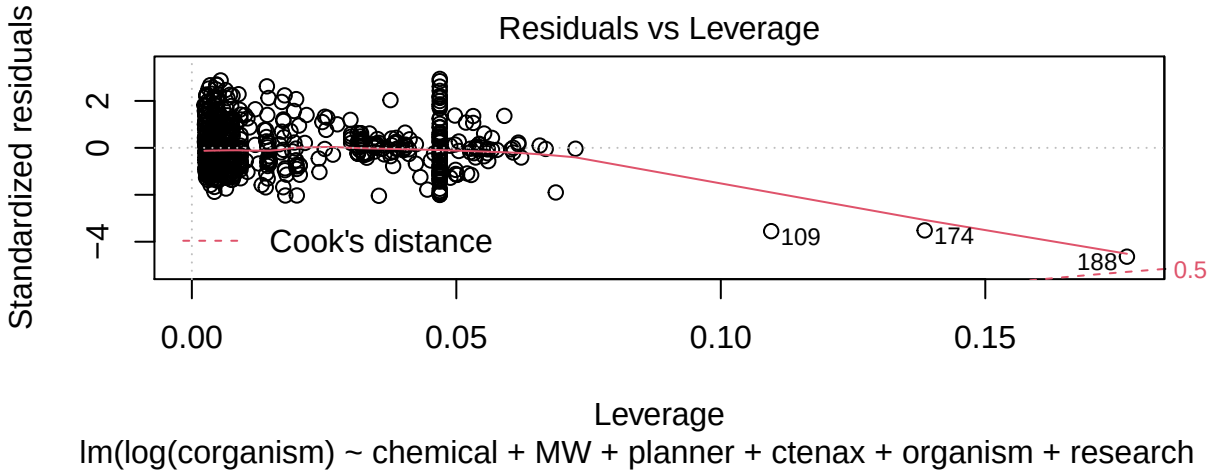


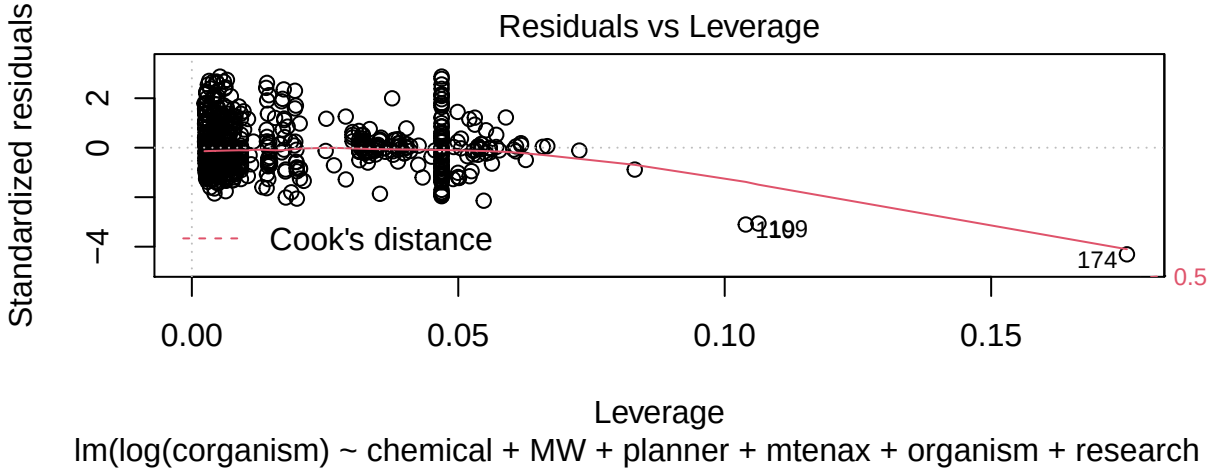
lm(log(corganism) ~ chemical + MW + planner + mtenax + organism + research ...

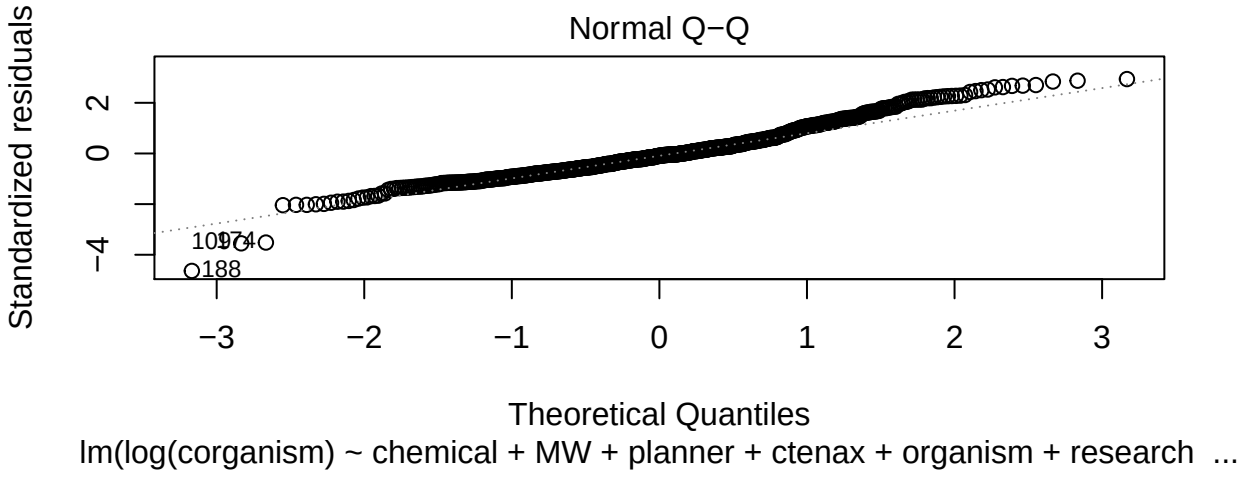


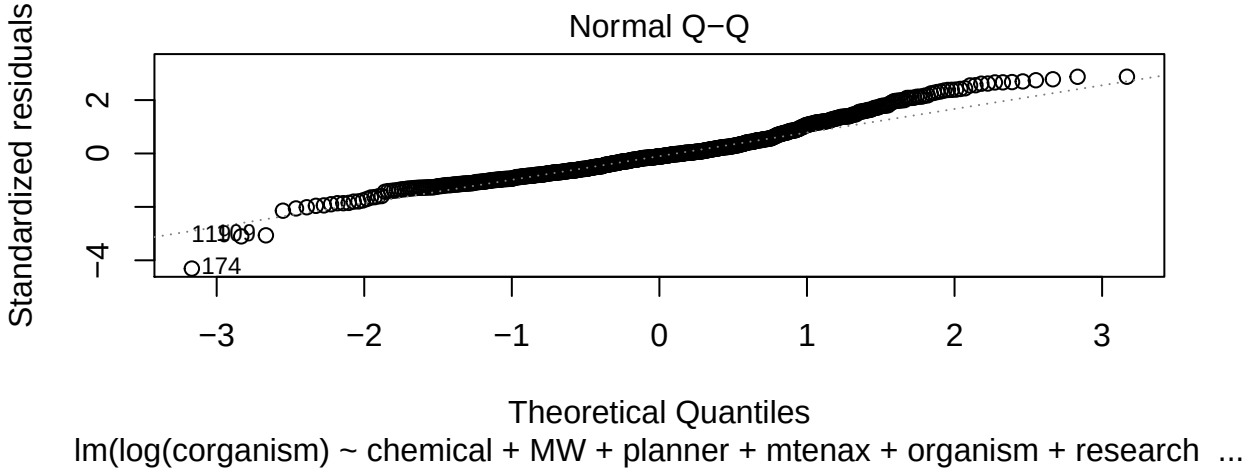


The measurement 71 is definitely bad.
Now, try and remove the measurement.


Residuals vs Leverage
Standardized residuals
2
0
−4
Cook's distance
109
174
188
0.5
0.00
0.05
0.10
0.15
Leverage
lm(log(corganism) ~ chemical + MW + planner + ctenax + organism + research ...


Residuals vs Leverage
Standardized residuals
2
0
−4
Cook's distance
109
174
0.5
0.00
0.05
0.10
0.15
Leverage
lm(log(corganism) ~ chemical + MW + planner + mtenax + organism + research ...

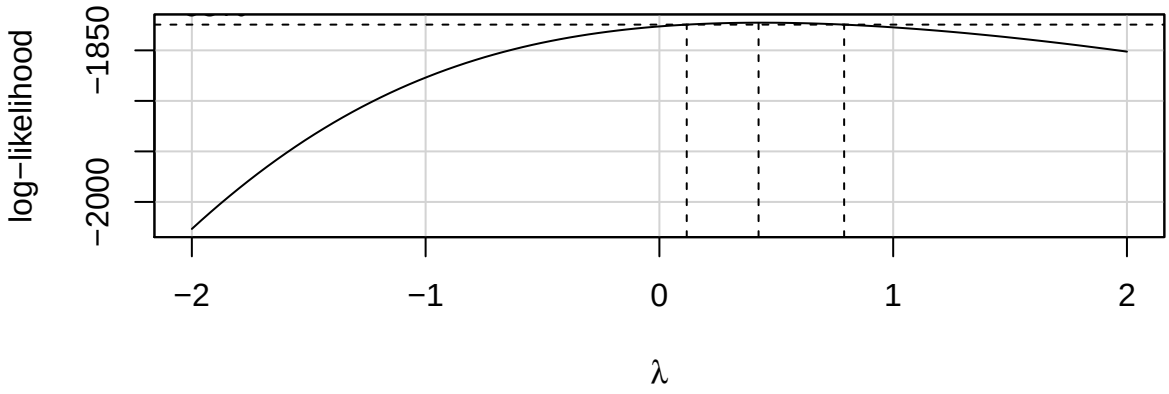


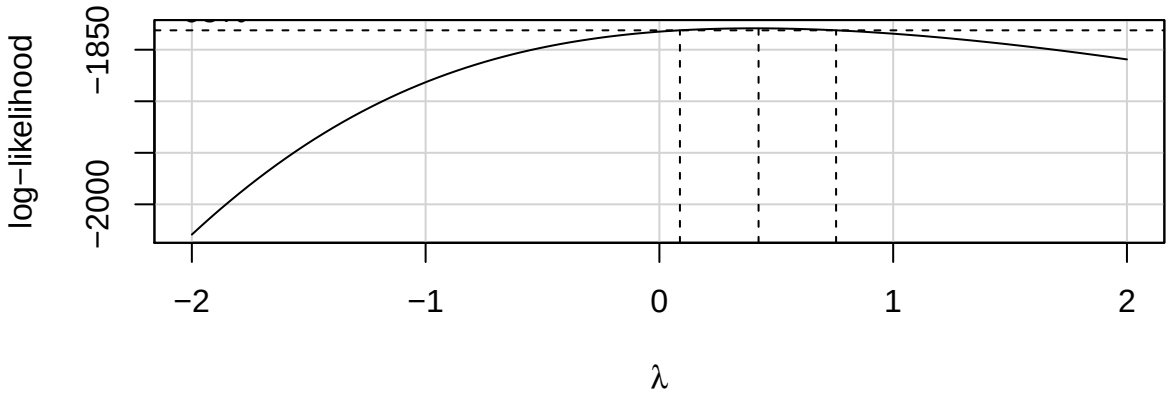


I tried to remove measurement 71, but then measurement 83 get on the 0.5 line.
So I removed it as well and the rest of the data seem to be acceptable.
And the residuals seem to be more normally distributed (QQ plot), which is good.
I checked the two measurements removed.
First I thought that their chemical concentration were too high that they break from the pattern.
However, although they are among the ones with higher chemical concentration, they are not the highest ones.
There are other data having high chemical concentration but not appear to tilt the regression as much.
Nonetheless, the measurements that deviates from the cluster appear to generally have higher chemical concentration, suggesting that the prediction may have much higher error at higher concentration.

Try a two way interaction model

Check for y transformation:

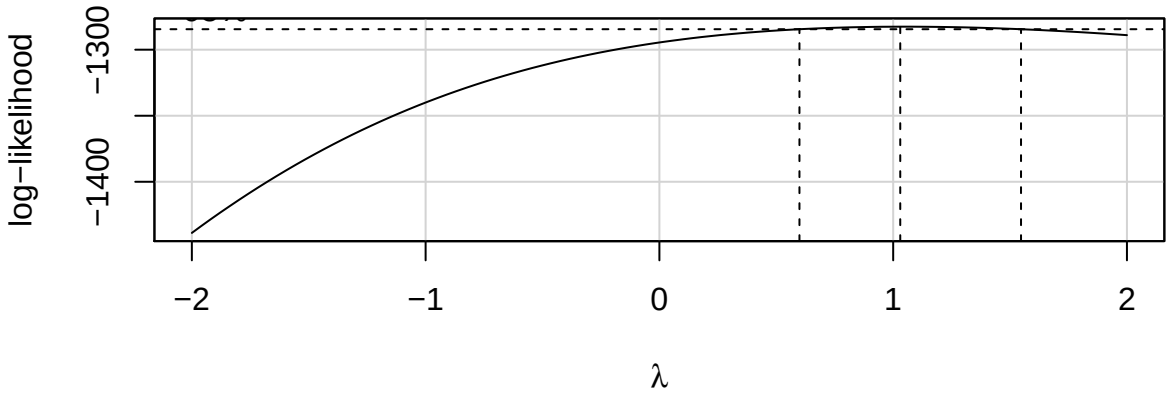


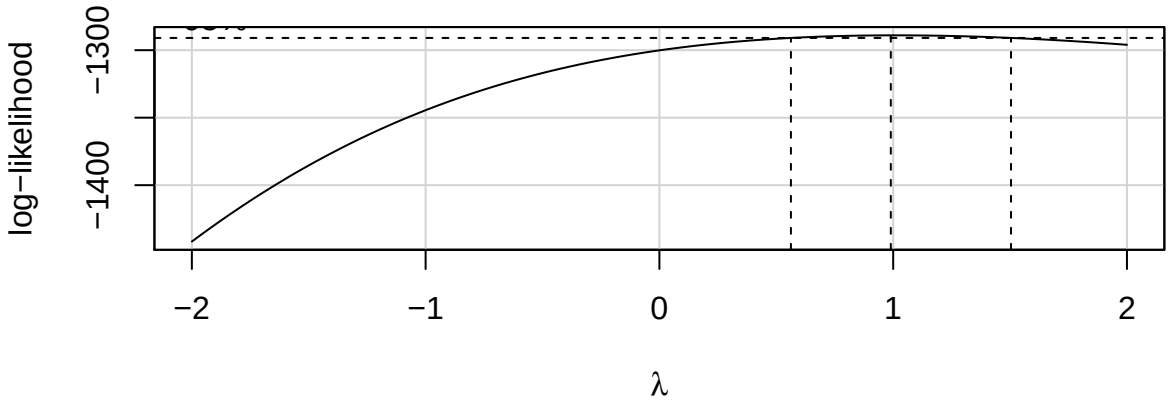



```
## [1] 0.4242424
```

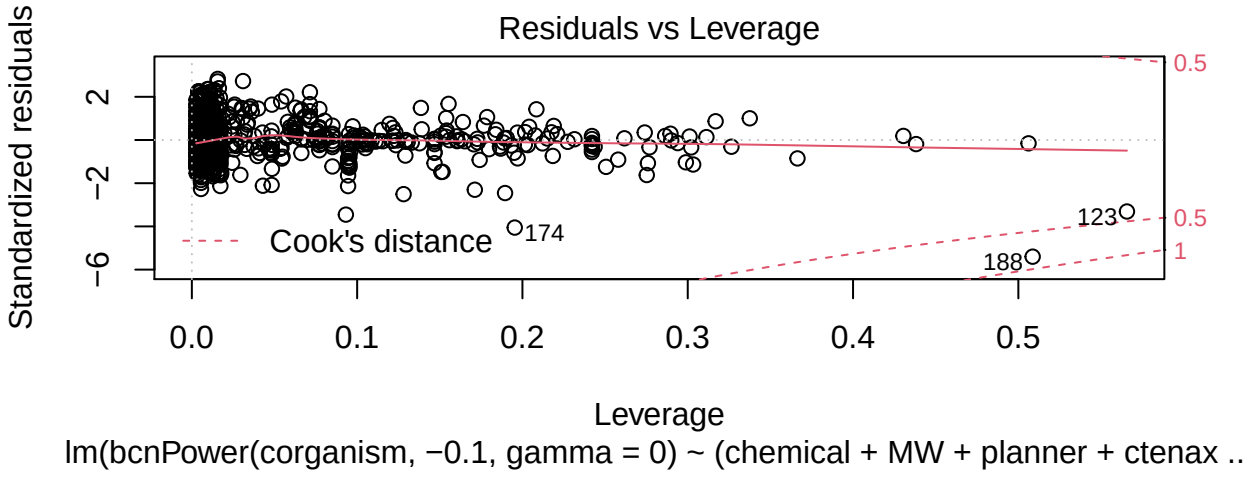
```
## [1] 0.4242424
```

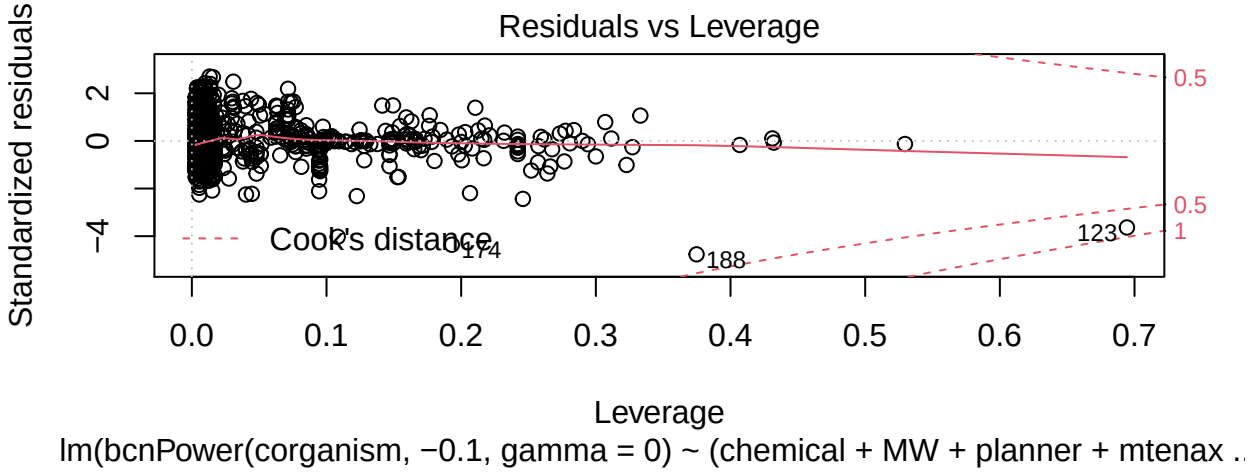
The two way interaction requires a different y transformation where $\hat{y} = (y^{\lambda} - 1) \div \lambda$ with $\lambda = 0.4242$.



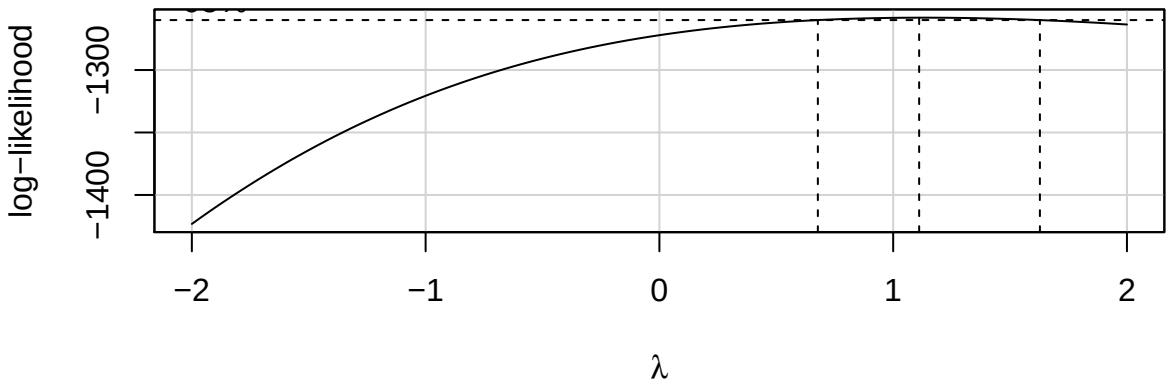


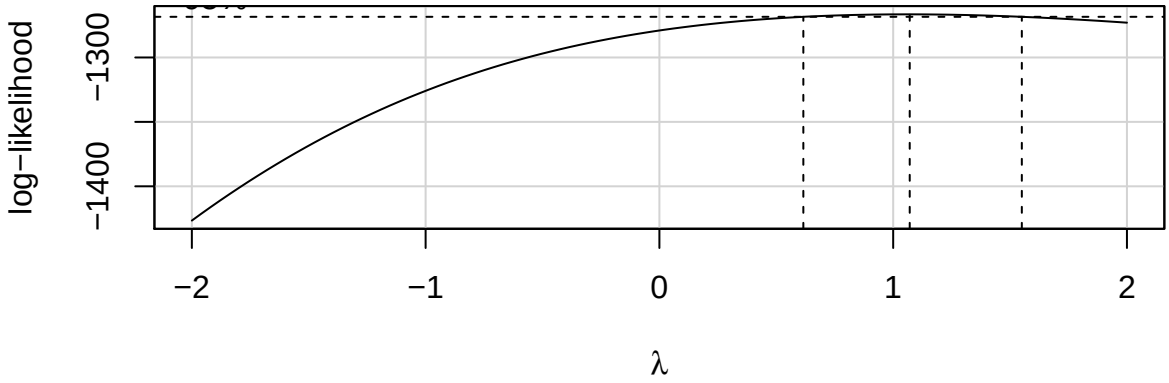
I tried with $\lambda = 0.4242$, but the transformation did not fix the problem.
Instead, it worked with λ = -0.1, which is displayed in the graphs above.
I will be using that for the two way interaction model.
Now check the cook's distance.

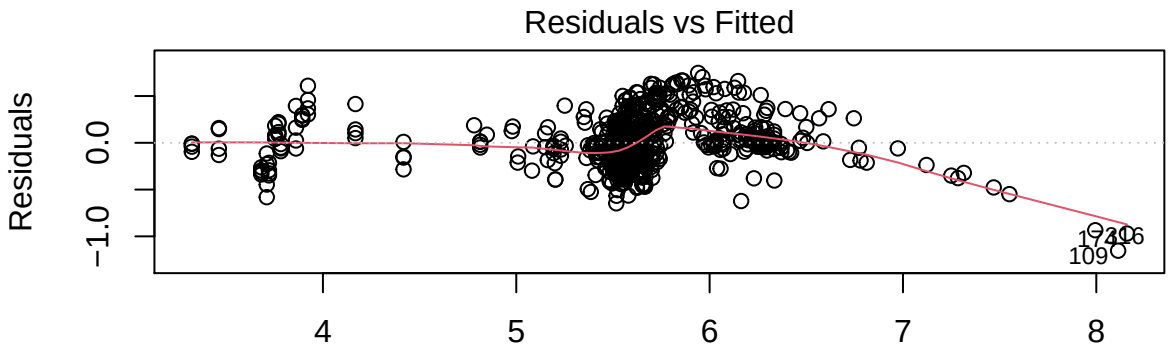




It appears that I should remove the measurements 123 and 188.
These are different points than the no interaction model.
This means that I cannot compare the resulting best models from the two groups.
Now, I want to check if removing the two points solve the problem, or cause any new problems.

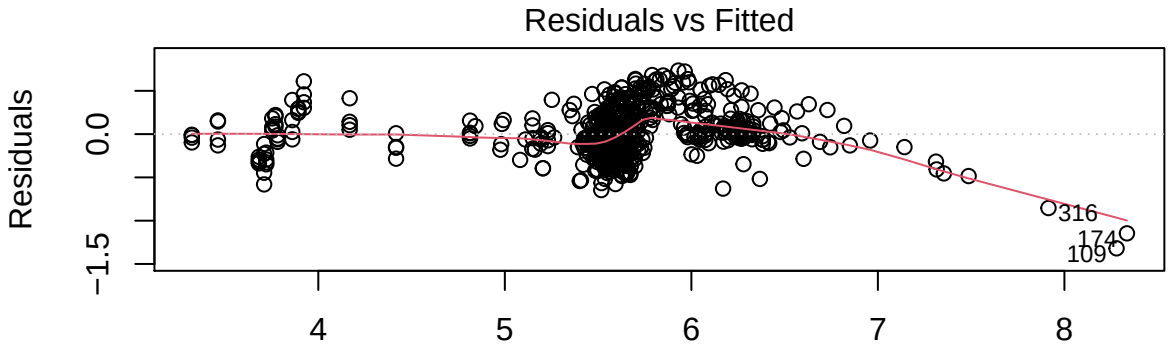

log−likelihood
−1300
−1400
−2
−1
0
1
2
λ


log−likelihood
−1300
−1400
−2
−1
0
1
2
λ



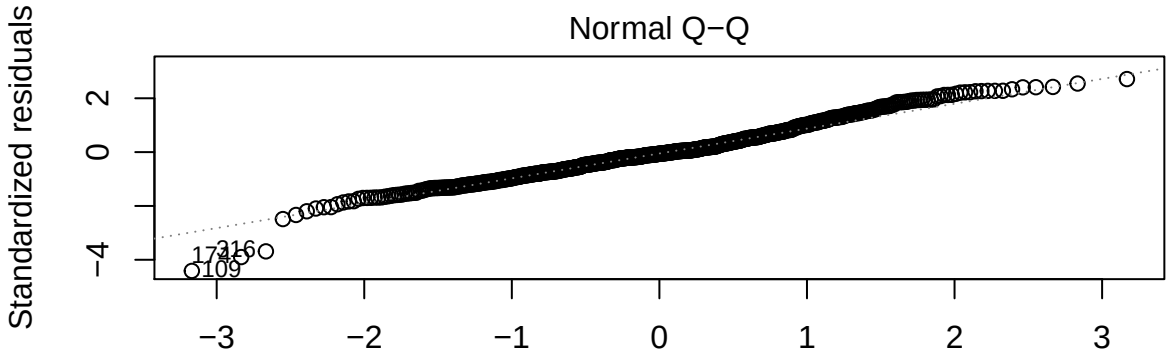
Fitted values

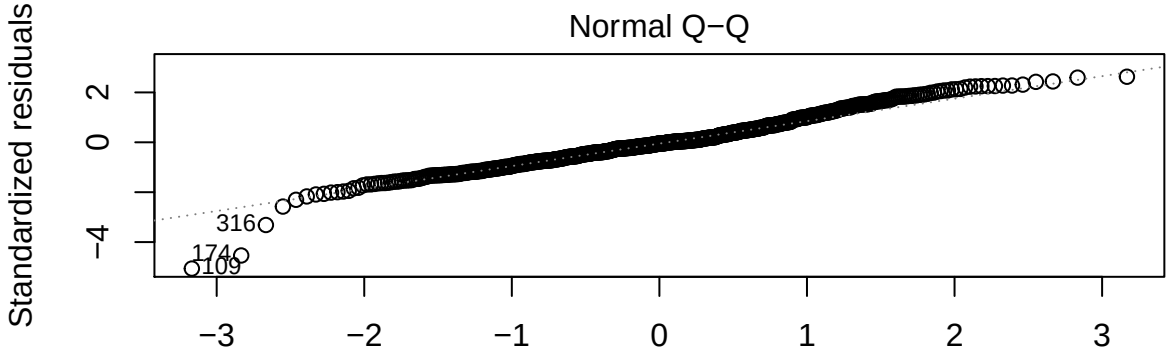
lm(bcnPower(corganism, −0.1, gamma = 0) ~ (chemical + MW + planner + ctenax ..

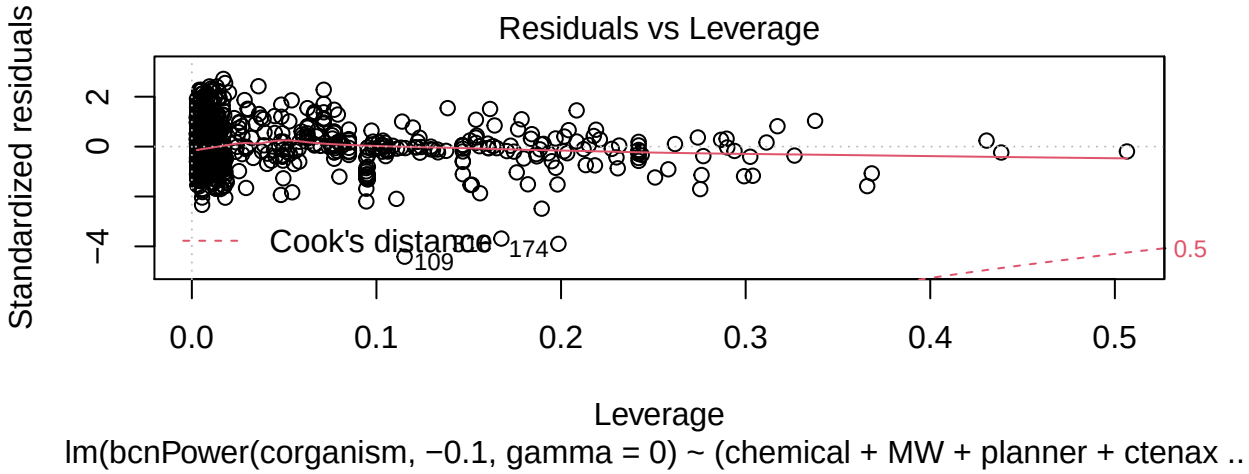


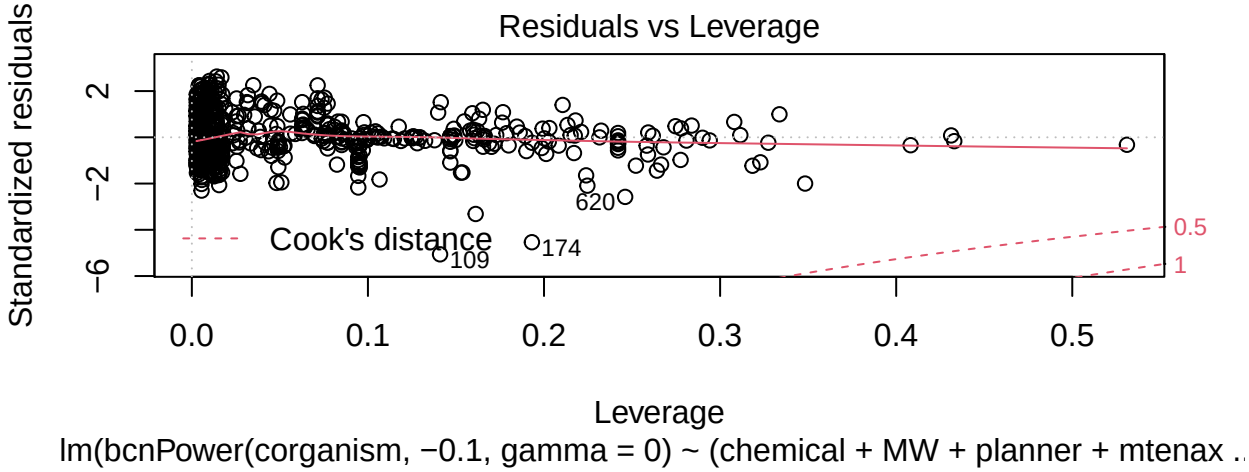
Fitted values

lm(bcnPower(corganism, −0.1, gamma = 0) ~ (chemical + MW + planner + mtenax ..


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(bcnPower(corganism, −0.1, gamma = 0) ~ (chemical + MW + planner + ctenax ..


Normal Q−Q
Standardized residuals
316
174
109
Theoretical Quantiles
lm(bcnPower(corganism, −0.1, gamma = 0) ~ (chemical + MW + planner + mtenax ..





It appears that the cook's distance plot seems fine now. No one data point is having too much influence.
The residual plot still seems bad, but I do not know how to fix that.
I could remove more points, but that is not very reasonable to do.
The QQ plot appears that the residuals are normally distributed.
Again, the points 188 and 123 are also among the highest chemical concentration.

Model selection

First look at the no interaction models

Models using concentration of chemical on Tenax:

n	predictors	aic	sbic
1	research	2026.093	171.0491
2	ctenax research	1664.100	-189.4582

n	predictors	aic	sbic
3	chemical ctenax tsed	1523.114	-335.1132
4	chemical MW ctenax tsed	1522.478	-335.6912
5	chemical MW ctenax organism tsed	1521.933	-336.1636
6	chemical MW ctenax organism OC tsed	1522.605	-335.4297
7	chemical MW planner ctenax organism OC tsed	1524.593	-333.4034
8	chemical MW planner ctenax organism research OC tsed	1528.593	-331.3659

It would appear the best model based on best subset is the fifth model:
corganims = chemical + MW + ctenax + organism + tsed
(since they have smallest aic and bic value)
Let's see if the forward step wise selection (using p-value):

```
##
##                              Selection Summary
## ------------------------------------------------------------------------------
##         Variable                 Adj.
## Step    Entered     R-Square    R-Square      C(p)          AIC         RMSE
## ------------------------------------------------------------------------------
##    1    research      0.4627      0.4610    771.7655    2026.0929    1.1400
##    2    ctenax        0.6926      0.6911    166.3863    1664.1002    0.8630
##    3    tsed          0.7540      0.7517      6.1319    1526.8341    0.7738
## ------------------------------------------------------------------------------
```

The model selected is:
corganims = research + ctenax + tsed
What about using backward step wise selection?

```
##
##
##                             Elimination Summary
## ------------------------------------------------------------------------------
##         Variable                 Adj.
## Step    Removed     R-Square    R-Square     C(p)         AIC         RMSE
## ------------------------------------------------------------------------------
##    1    planner       0.7578       0.754    4.0113    1526.6046    0.7701
##    2    OC            0.7573      0.7539    3.3165    1525.9329    0.7703
##    3    organism      0.7564      0.7533    3.8247    1526.4779    0.7712
##    4    MW            0.7554      0.7527    4.4329    1527.1139    0.7722
##    5    chemical       0.754      0.7517    6.1319    1526.8341    0.7738
## ------------------------------------------------------------------------------
```

The resulting model is the same as suggested by the forward step wise selection.

Now, check the two selected model.
Since they are not reduced version of each other, I would better use cross validation (leave one out).

```
## Warning: package 'cvTools' was built under R version 4.0.3
```

```
## Loading required package: lattice
```

```
## Loading required package: robustbase
```

```
## Warning: package 'robustbase' was built under R version 4.0.3
```

```
## Leave-one-out CV results:
##        CV
## 0.7835531
```

The second model (from step wise selection) suffers from rank deficiency and the CV did not work on it.

Let's check the AIC value of both model:

```
## [1] 1521.933
```

```
## [1] 1524.834
```

It appears that the model from the best subset have the smallest AIC value.
This is kind of unfair because the best subset selection used AIC in the first place.
Nonetheless, I will go with the model selected by best subset.
(corganims = chemical + MW + ctenax + organism + tsed)

Models using mass of chemical on Tenax:

n	predictors	aic	sbic
1	research	2026.093	171.2079
2	mtenax research	1693.639	-159.8664
3	chemical mtenax tsed	1563.742	-294.5439
4	chemical mtenax OC tsed	1559.539	-298.6486
5	chemical MW mtenax OC tsed	1558.889	-299.2272
6	chemical MW mtenax organism OC tsed	1558.475	-299.5587
7	chemical MW planner mtenax organism OC tsed	1560.474	-297.5224
8	chemical MW planner mtenax organism research OC tsed	1564.474	-295.4849

Similarly, the best model based on best subset is the sixth model:
corganims = chemical + MW + mtenax + organism + OC + tsed
Notice that compared to the result from previous selection, the proportion of organic carbon is included.
This is likely because the ctenax is just mtenax ÷ OC.

Now, try the step wise selections:

```
##
##                              Selection Summary
## ------------------------------------------------------------------------
##         Variable                 Adj.
## Step    Entered     R-Square    R-Square      C(p)         AIC        RMSE
## ------------------------------------------------------------------------
##    1    research      0.4627      0.4610    695.7437    2026.0929    1.1400
##    2    mtenax        0.6783      0.6768    158.5221    1693.6388    0.8827
##    3    tsed          0.7382      0.7358     10.7326    1567.3129    0.7982
##    4    OC            0.7407      0.7379      6.5665    1563.1434    0.7950
## ------------------------------------------------------------------------
```

```
##
##
##                            Elimination Summary
## ------------------------------------------------------------------------
##         Variable                  Adj.
## Step    Removed     R-Square    R-Square     C(p)        AIC        RMSE
## ------------------------------------------------------------------------
##    1    planner       0.7441      0.7401    4.0012    1562.4754    0.7916
##    2    organism      0.7431      0.7395    4.3751    1562.8893    0.7925
##    3    MW            0.7421      0.7389    4.9908    1563.5388    0.7935
##    4    chemical      0.7407      0.7379    6.5665    1563.1434    0.7950
## ------------------------------------------------------------------------
```

Similarly, the models selected by the backward and forward step wise selection are the same.
Also, the variable "research" is used replacing "organism", "MW", and "chemical".
The reason, I suppose is that different types of molecules, organisms, and chemicals were used for different researches.
So in a way, the information of these variable may be partially represented by the "research".

Anyway, let's check the two models from best subset and step wise selection:

```
## Leave-one-out CV results:
##        CV
## 0.8033963
```

The model using "research" (step wise) again suffers from rank deficiency.

Let's check the AIC value of both model:

```
## [1] 1558.475
```

```
## [1] 1561.143
```

Again, the best subset model is better.

Now compare the two model using ctenax and mtenax.
Since they are using the same data set, I can try to use cross validation.

```
## Leave-one-out CV results:
##        CV
## 0.7835531
```

```
## Leave-one-out CV results:
##        CV
## 0.8033963
```

It appears that the model with ctenax, concentraion of chemical on tenax, is better than the model using mtenax and OC for, essentially, the same portion of information.
The combined version of the information is good enough and it is not necessary to use the two variables for that.
So the resulting best model is:
corganims = chemical + MW + ctenax + organism + tsed

Let's take a look at the model:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	4.616	0.4302	10.73	8.148e-25
chemicalPCB	2.802	0.1945	14.41	5.518e-41
chemicalPermethrin	0.3331	0.194	1.717	0.08645
MW	-0.00119	0.0007367	-1.616	0.1066
ctenax	6.579e-05	2.857e-06	23.03	5.314e-86
organismLV	0.3052	0.1926	1.585	0.1136
tsedHumic Acid	2.116	0.2037	10.39	1.781e-23
tsedLPH	2.204	0.2393	9.21	4.567e-19
tsedSaw Dust	2.39	0.2018	11.84	2.152e-29
tsedunknown	0.7779	0.1723	4.516	7.508e-06

Table 4: Fitting linear model: log(corganism) ~ chemical + MW + ctenax + organism + tsed

Observations	Residual Std. Error	R^2	Adjusted R^2
652	0.7703	0.7573	0.7539

The model is:
$ln(corganism) = 4.616 - 0.00119 \times MW + 6.579 \times 10^{-5} \times ctenax$ The above add:
2.802 if chemical is PCB
.3331 if chemical is Permethrin
.3052 if organism is LV
2.116 if sediment is humic acid
2.204 if sediment is LPH
2.390 if sediment is saw dust
.7779 if sediment is unknown
So the concentration of chemical captured by the organism is negatively correlated with the molecular weight of the chemical and positively correlated with the concentration of chemical captured by the Tenax.
This is expected since the heavier the molecule, the slower and harder it is expected to move across phases. The more chemical captured by Tenax indicates a higher level of chemical present in the system and thus the chemical captured by organism in the same system is expected to be higher.
For the type of chemical, the Permethrin and Bifenthrin are less different in their prediction, only different in a multiple of 1.4 .
The PCBs have a 16.5 fold increase in the chemical concentration on organism.

Can I use the information from molecular weight to cover the information from type of chemical?

```
## Analysis of Variance Table
##
## Model 1: log(corganism) ~ MW + ctenax + organism + tsed
## Model 2: log(corganism) ~ chemical + MW + ctenax + organism + tsed
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    644 536.61
## 2    642 380.95  2    155.65 131.16 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p-value is smaller than 0.05, suggesting that there is enough evidence that the full model is better.
I cannot remove the variable "chemical".
The type of chemical is important even when the molecular weight is considered.

I would like to see if the linear dependent problem of the variables is resolved.

```
##              GVIF Df GVIF^(1/(2*Df))
## chemical 2.610057  2        1.271050
## MW       1.969709  1        1.403463
## ctenax   1.112249  1        1.054632
## organism 1.901840  1        1.379072
## tsed     1.347268  4        1.037963
```

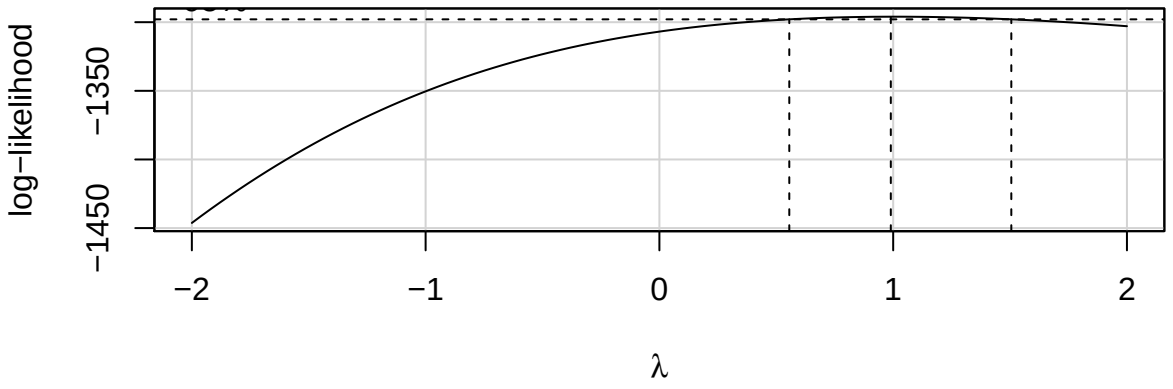
The VIF value is less than 5, which is suggested to be acceptable.
The linear dependent problem is resolved.

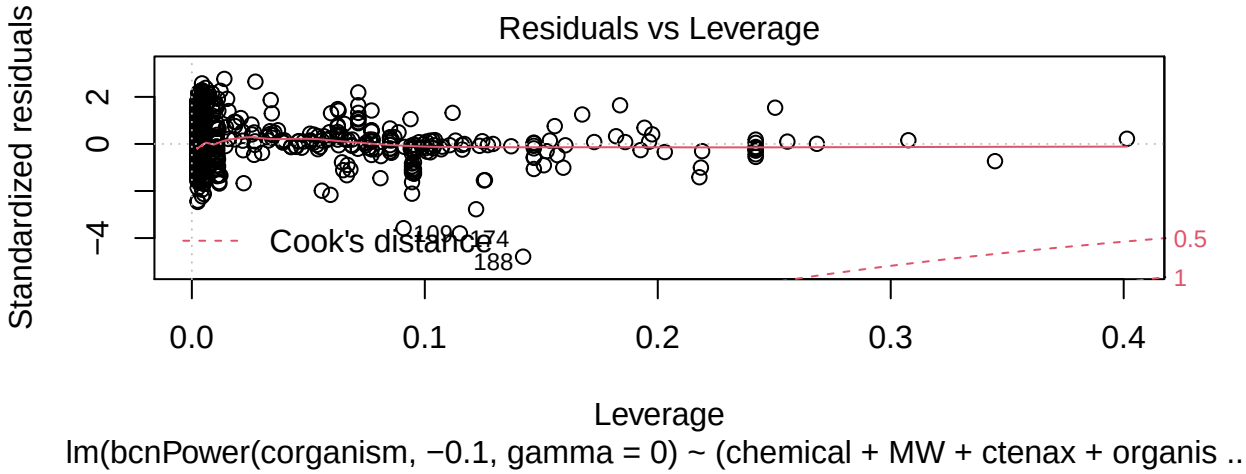
Now check the two way interaction model models

Models using concentration of chemical on Tenax:

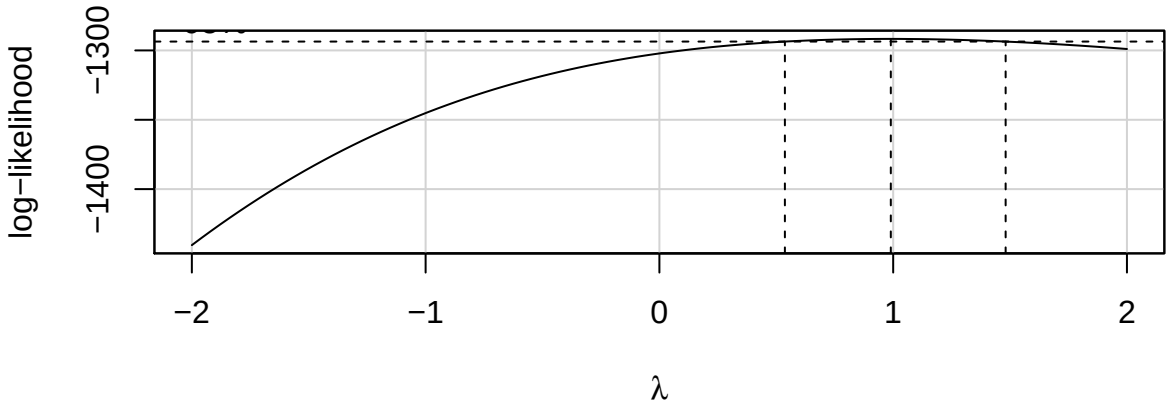
The full model was too big for both cross validation and best subset selection.
Therefore, I reduced the model to only include the variables determined to be significant in the no interaction model.

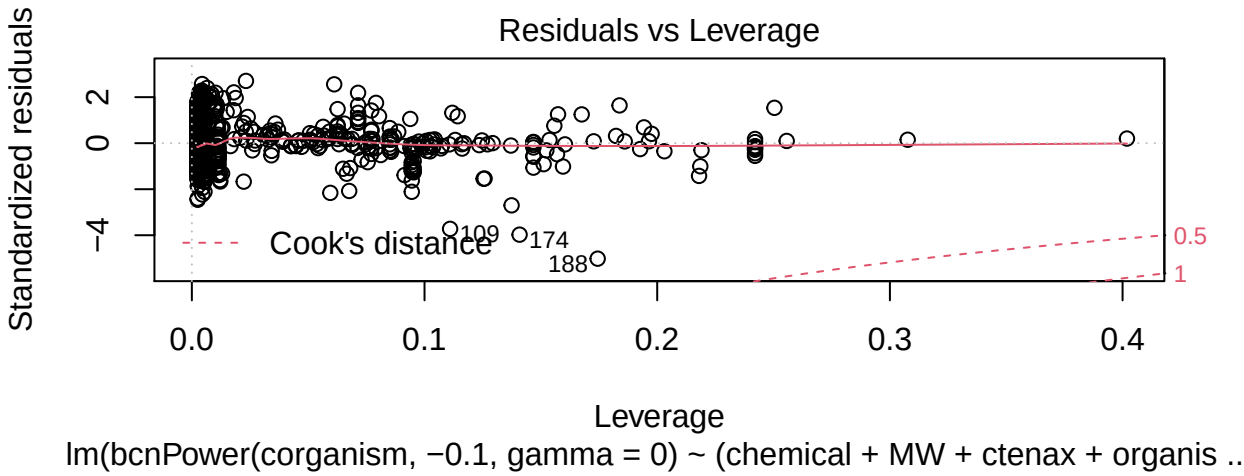
I will have to recheck the assumptions:





It appears that the $\lambda = -0.1$ transformation still works.
Also there is no data need to be removed based on cook's distance.
However, for the sake of comparison, I would like to remove data 71 and 83 so that I may be able to compare the resulting model with the no interaction model.





(the assumptions still hold after removing 71 and 83)

n	predictors	aic	sbic
1	chemical:tsed	832.3334	-4681.382
2	chemical:ctenax chemical:tsed	340.2799	-18629.212
3	chemical:ctenax chemical:tsed MW:ctenax	287.5180	-21693.036
4	chemical:MW chemical:ctenax MW:ctenax organism:tsed	262.2278	-23314.729
5	chemical:MW chemical:ctenax chemical:organism chemical:tsed MW:ctenax	263.2450	-23732.469
6	chemical:MW chemical:ctenax chemical:organism MW:ctenax MW:tsed ctenax:tsed	250.5284	-23852.030
7	chemical:MW chemical:ctenax chemical:organism chemical:tsed MW:ctenax MW:tsed ctenax:tsed	272.7144	-23762.745
8	chemical:MW chemical:ctenax chemical:organism chemical:tsed MW:ctenax MW:tsed ctenax:organism ctenax:tsed	274.7128	-23542.285
9	MW ctenax chemical:ctenax chemical:organism chemical:tsed MW:ctenax MW:tsed ctenax:organism ctenax:tsed	270.7128	-23321.708
10	MW ctenax chemical:ctenax chemical:tsed MW:ctenax MW:organism MW:tsed ctenax:organism ctenax:tsed organism:tsed	274.7128	-23101.131
11	MW ctenax organism tsed chemical:MW chemical:ctenax MW:ctenax MW:organism MW:tsed ctenax:organism ctenax:tsed	256.7128	-22880.555
12	MW ctenax organism tsed chemical:MW chemical:ctenax chemical:tsed MW:ctenax MW:organism MW:tsed ctenax:organism ctenax:tsed	272.7128	-22880.555
13	MW ctenax organism tsed chemical:MW chemical:ctenax chemical:tsed MW:ctenax MW:organism MW:tsed ctenax:organism ctenax:tsed organism:tsed	280.7128	-22880.555
14	MW ctenax organism tsed chemical:MW chemical:ctenax chemical:organism chemical:tsed MW:ctenax MW:organism MW:tsed ctenax:organism ctenax:tsed organism:tsed	284.7128	-22880.555

n	predictors	aic	sbic
15	chemical MW ctenax organism tsed chemical:MW chemical:ctenax chemical:organism chemical:tsed MW:ctenax MW:organism MW:tsed ctenax:organism ctenax:tsed organism:tsed	288.7128	-22880.555

The model selected:
(the ":" in equation means multiply, interaction of the two varibles without linear teams)
corganism = chemical:MW+chemical:ctenax+chemical:organism+MW:ctenax+MW:tsed+ctenax:tsed
Problem with this model is that the interaction terms does not have the corresponding linear terms in the model.
However, I will still take it and compare it to the models selected by step wise selections.

Now, try some step wise selections:

```
##
##                                Selection Summary
## ------------------------------------------------------------------------------------
##         Variable                          Adj.
## Step        Entered        R-Square    R-Square      C(p)        AIC        RMSE
## ------------------------------------------------------------------------------------
##    1    chemical:tsed        0.6066      0.6030    905.7978    832.3334    0.4487
##    2    ctenax               0.7612      0.7586    297.2720    508.9185    0.3499
##    3    MW                   0.7623      0.7594    294.7796    507.8059    0.3493
##    4    organism             0.7640      0.7607    290.3420    505.3194    0.3484
##    5    tsed                 0.7640      0.7607    292.3420    503.3194    0.3484
##    6    chemical:ctenax      0.8178      0.8147     81.5464    338.4001    0.3066
##    7    chemical             0.8178      0.8147     81.5464    338.4001    0.3066
##    8    MW:ctenax            0.8369      0.8339      8.1472    268.2278    0.2903
##    9    MW:organism          0.8392      0.8359      1.3335    261.2450    0.2885
## ------------------------------------------------------------------------------------
```

```
##
##
##                               Elimination Summary
## ------------------------------------------------------------------------------------
##         Variable                          Adj.
## Step        Removed        R-Square    R-Square     C(p)        AIC        RMSE
## ------------------------------------------------------------------------------------
##    1    ctenax:organism      0.8408      0.8355    -0.9984    286.7144    0.2889
##    2    ctenax:tsed          0.8398      0.8355     0.6343    282.4691    0.2888
##    3    MW:tsed              0.8392      0.8359     1.3335    277.2450    0.2885
## ------------------------------------------------------------------------------------
```

The model for forward step wise selection:
corganism = chemical+ctenax+organism+MW+tsed+chemical:tsed+chemical:ctenax+MW:ctenax+MW:organism
The model for backward step wise selection:
corganism = chemical+ctenax+organism+MW+tsed+chemical:ctenax+chemical:organism+chemical:MW+chemical:tsed+ctenax:MW+organism:MW+organism:tsed

Now, compare the three models selected:

```
## Leave-one-out CV results:
##        CV
## 0.2939766
```

```
## Leave-one-out CV results:
##        CV
## 0.2938397
```

```
## Leave-one-out CV results:
##        CV
## 0.2938397
```

The rank deficiency warning still pops up.
However, this time cross validation does produce a result.
It suggests that the model suggested by the best subset is not better than the onces suggested by the step wise selection.
The two step wise selection models have equal CV values.

Let's look at AIC of the two remaining models:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.19	1.903	4.829	1.72e-06
chemicalPCB	1.799	0.1067	16.87	4.446e-53
chemicalPermethrin	-0.007123	0.13	-0.05481	0.9563
ctenax	0.00978	0.001205	8.117	2.475e-15
organismLV	-5.345	1.86	-2.874	0.004186
MW	-0.01491	0.004574	-3.26	0.001175
tsedHumic Acid	0.8975	0.07636	11.75	5.198e-29
tsedLPH	0.7918	0.0917	8.634	4.708e-17
tsedSaw Dust	0.9902	0.07571	13.08	8.477e-35
tsedunknown	0.2979	0.06487	4.592	5.28e-06
chemicalPCB:ctenax	-0.009858	0.001205	-8.182	1.513e-15
chemicalPermethrin:ctenax	-0.006389	0.00124	-5.154	3.396e-07
ctenax:MW	4.001e-07	4.598e-08	8.701	2.781e-17
organismLV:MW	0.01358	0.004565	2.975	0.003041

Table 7: Fitting linear model: bcnPower(corganism, -0.1, gamma = 0) ~ chemical + ctenax + organism + MW + tsed + chemical:tsed + chemical:ctenax + MW:ctenax + MW:organism

Observations	Residual Std. Error	R^2	Adjusted R^2
652	0.2885	0.8392	0.8359

I checked the output of the two linear model and find that they are actually the same, despite the inputs are different.
Somehow, some of the variables I put in are lost after fitting the model.
Regardless, I will select this model to be the best model from two way selection:

$(corganism^{-0.1} - 1) \div (-0.1) = 9.19 - 0.01491 \times MW + 0.00978 \times ctenax + 4.001 \times 10^{-7} \times MW \times ctenax$
The above add:

1.799 if chemical is PCB
-0.007123 if chemical is Permethrin
-5.345 if organism is LV
0.8975 if sediment is humic acid
0.7918 if sediment is LPH
0.9902 if sediment is saw dust
0.2979 if sediment is unknown
The slope of ctenax term add:
-0.009858 if the chemical is PCB -0.006389 if the chemical is Permethrin The slope of MW term add:
0.01358 if organism is LV

Again, we see a positive correlation between corganism and ctenax and a negative correlation between corganism and MW.
The interaction term is just a correction of the effect of ctenax or MW on corganism when the other variable is constant.
The categorical variables are correction of the y-intercept (after the transformation).
The type of chemical is used to correct the slope of ctenax and the type of organism is used to correct the slope of MW.
This is likely due to the tenax having different affinity towards different types of chemical, even the molecular weight is considered.
Also, the type of organism may be more different in respond to the molecular weight of the chemical when picking the chemical up. Whereas the Tenax is more concerned with the hydrophobicity, and thus the type of chemical.

Compare the no interaction model with the two way interaction model

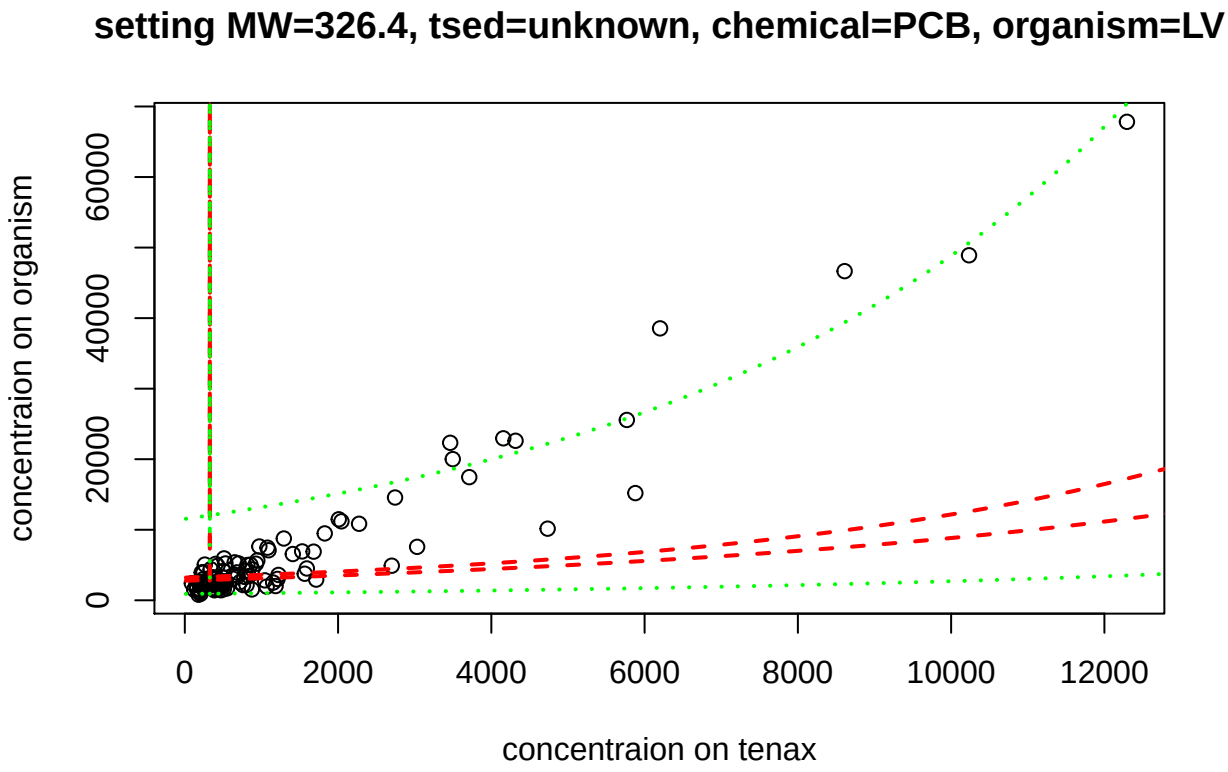
Since I have set the two model to use the same data set, I can try to use cross validation:
(notice that I change the expression of the two way interaction model to the one provided after the fitting)

```
## Leave-one-out CV results:
##        CV
## 0.7835531
```

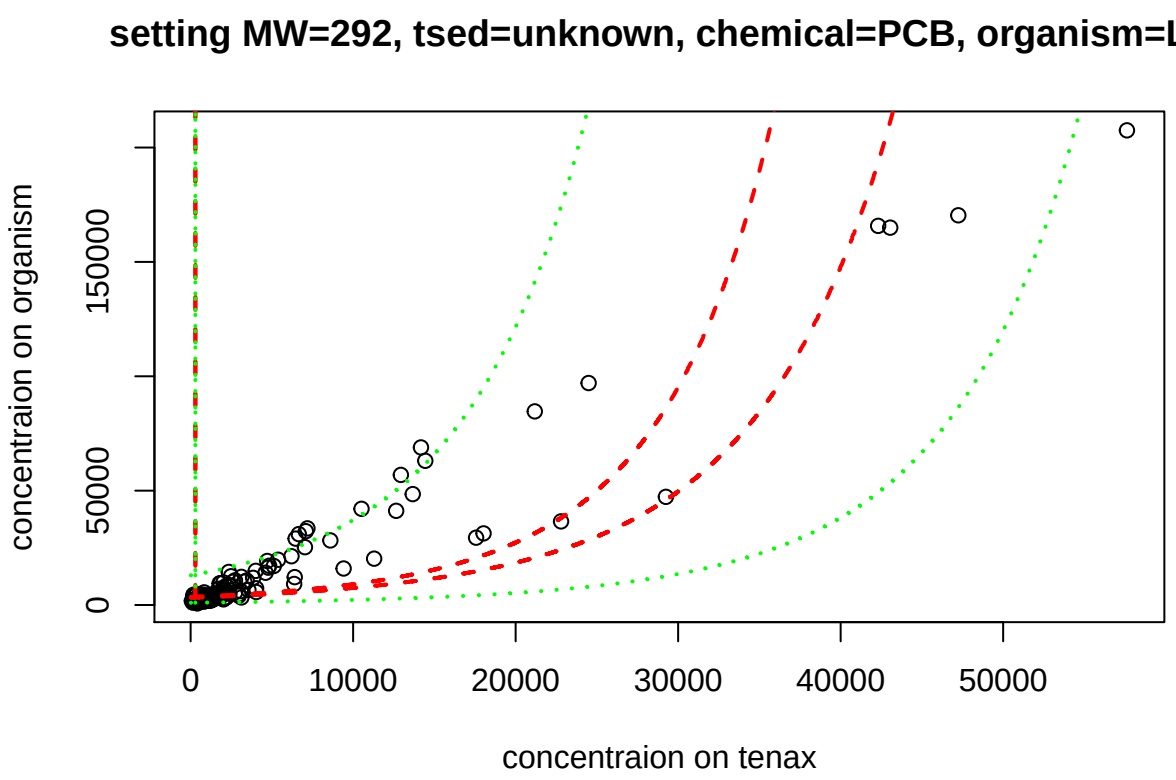
```
## Leave-one-out CV results:
##        CV
## 0.2938397
```

The cross validation suggests that the full model fits better than the reduced model.
Thus the best model is the selected model from the 2 way interaction models.

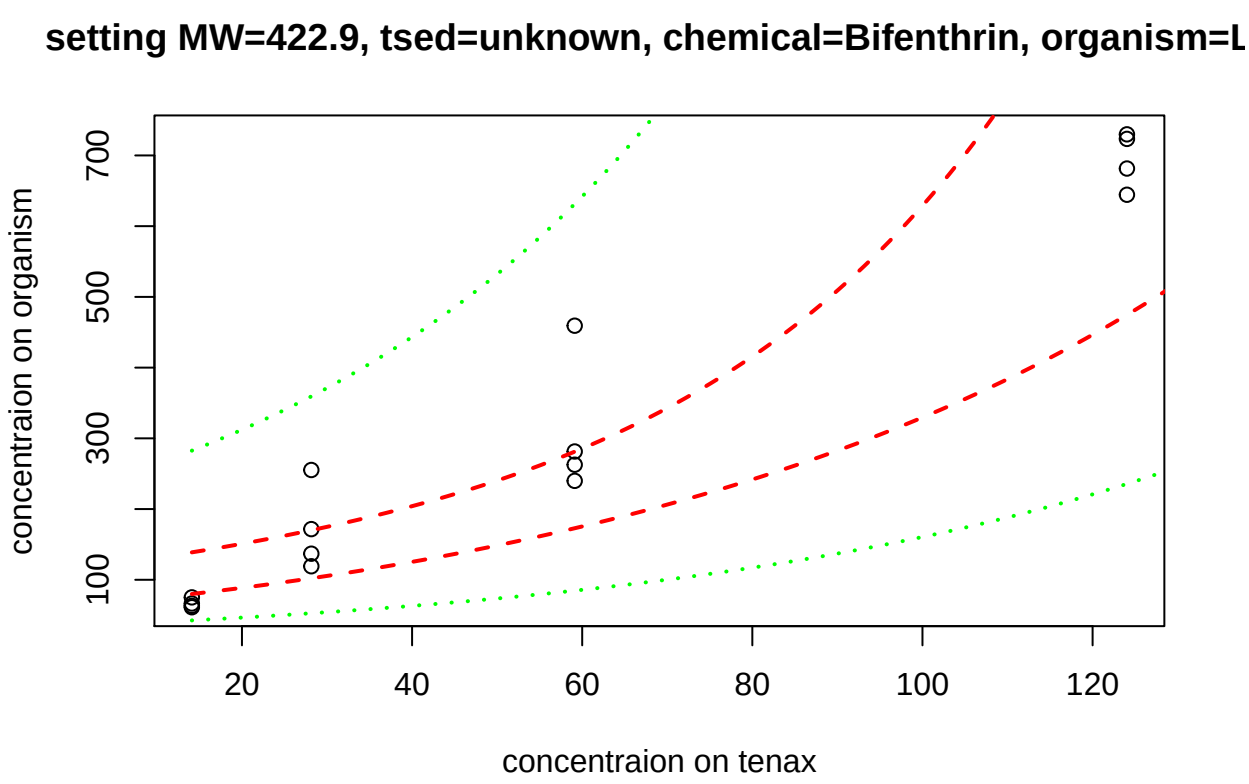
Look at the confidence intervals



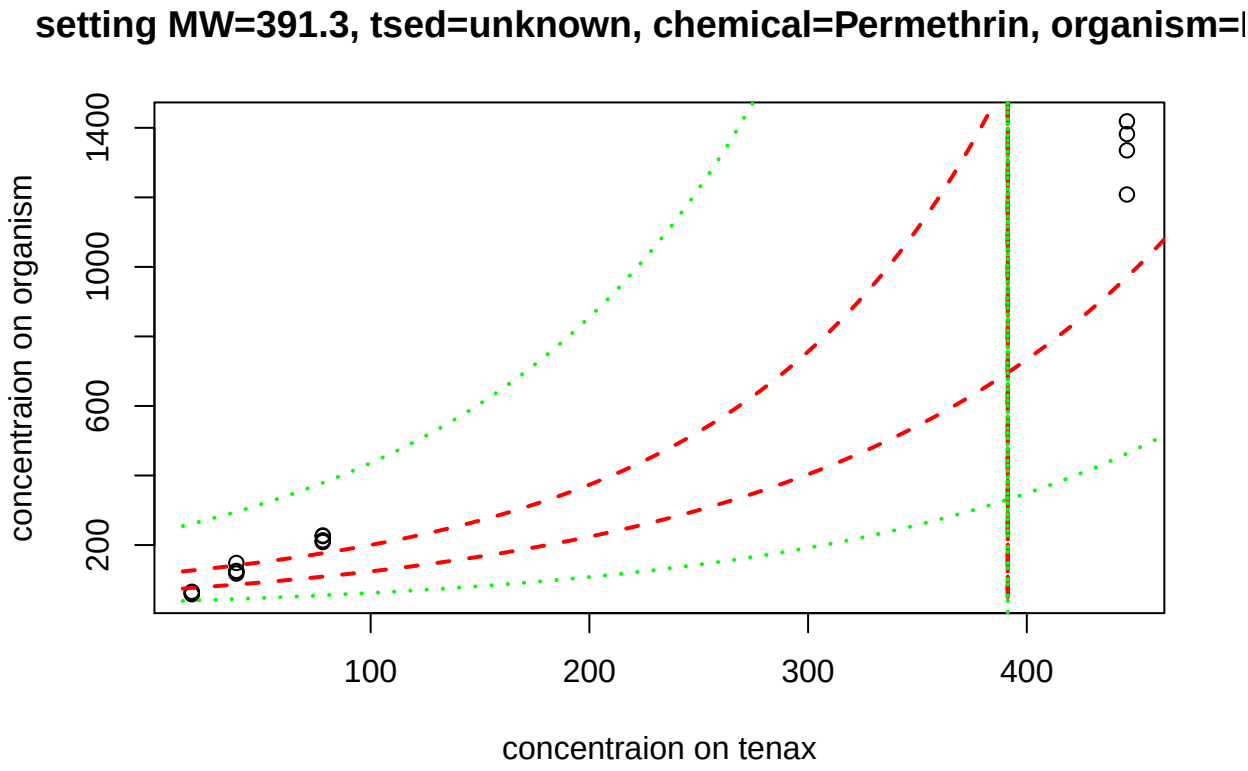
It appears that the prediction interval seems more reasonable.
The confidence interval is way off.



This does not look good at all.
It appears that the prediction interval is working somewhat at lower concentration but blow off once the concentration on Tenax goes over 15000 or so.



I think the model predicts heavier molecules better.
The prediction interval seems reasonable here.
Additionally, the set of points that is above the confidence interval are the ones with least age time.
Age time is a variable I did not include, because the other researches do not have this variable.
The small age time might mean that the organic carbon did not fully absorb the chemical (not settled) before the sorbent (organism) was introduced, resulting in a higher concentration of chemical.
Thus I think the confidence interval here is probably reasonable.



Similar result for Permethrin, however, there is a huge gap between the sets of data.
I would propose more experiments in between the concentraion to fix that.