A Crash Course in Practical Data Analysis[*]

Sasha Hafner[†]

May 16, 2021

[*] For the latest version, visit https://github.com/sashahafner/CCPDA

[†] sasha@hafnerconsulting.com

Contents

1 Packages and functions

```
source('functions/dfsumm.R')
```

```
library(tidyr)
library(dplyr)
library(ggplot2)
```

2 Problem 1. Inoculum effects on BMP

Koch et al. [2017] studied the effect of inoculum origin on biochemical methane potential (BMP) for four substrates. Data are given in the file BMP_inoc.csv, where the unit of observation is a single BMP bottle. Take a look at the data and answer these questions:

1. Did BMP depend on inoculum type?
2. Did any effect vary by substrate?

The original data are in a intermediate structure, with replicates across columns.

```
bi <- read.csv('data/BMP_inoc.csv')
```

```
bi

#        substrate inoc  BMP1  BMP2  BMP3  BMP4  BMP5  BMP6  BMP7  BMP8
# 1  Sewage Sludge WWTP 293.8 272.8 303.9 260.2 275.7 276.6 309.9 330.1
# 2          Maize WWTP 319.7 320.2 344.5 324.7 328.3 338.6 324.8 351.9
# 3     Food Waste WWTP 453.9 444.5 462.9 451.1 453.9 473.7 423.8 419.5
# 4      Cellulose WWTP 333.3 315.6 341.0 322.8 330.4 338.9 338.9 343.0
# 5  Sewage Sludge  ABP 294.8 294.2 293.9 267.0 269.6 272.5 332.4 319.8
# 6          Maize  ABP 320.1 325.6 348.6 362.5 343.8 412.5 326.6 330.9
# 7     Food Waste  ABP 441.1 432.2 466.2 490.0 398.3 429.3 423.3 432.5
# 8      Cellulose  ABP 344.1 347.7 374.8 348.5 351.3 378.0 354.9 367.5
# 9  Sewage Sludge BWTP 296.6 307.6 307.5 309.1 315.0 319.4 342.3 325.0
# 10         Maize BWTP 328.2 341.6 356.8 339.4 357.3 372.6 336.6 339.5
# 11    Food Waste BWTP 459.0 450.8 484.4 453.2 449.3 483.8 442.9 429.7
# 12     Cellulose BWTP 379.0 389.4 376.8 360.1 357.0 389.0 362.5 369.7
#     BMP9
# 1  328.3
# 2  352.1
# 3  432.0
# 4  350.0
# 5  319.4
# 6  335.5
# 7  439.8
# 8  366.9
# 9  347.1
# 10 356.0
# 11 458.2
# 12 376.7
```

This structure could work well in a spreadsheet analysis. For analysis in R, the structure can be changed to long using the `gather()` function.

```
bil <- gather(bi, key = 'rep', value = 'BMP', contains('BMP'))
head(bil)
```

```
#       substrate inoc  rep   BMP
# 1 Sewage Sludge WWTP BMP1 293.8
# 2         Maize WWTP BMP1 319.7
# 3    Food Waste WWTP BMP1 453.9
# 4     Cellulose WWTP BMP1 333.3
# 5 Sewage Sludge  ABP BMP1 294.8
# 6         Maize  ABP BMP1 320.1
```

```
dim(bil)
```

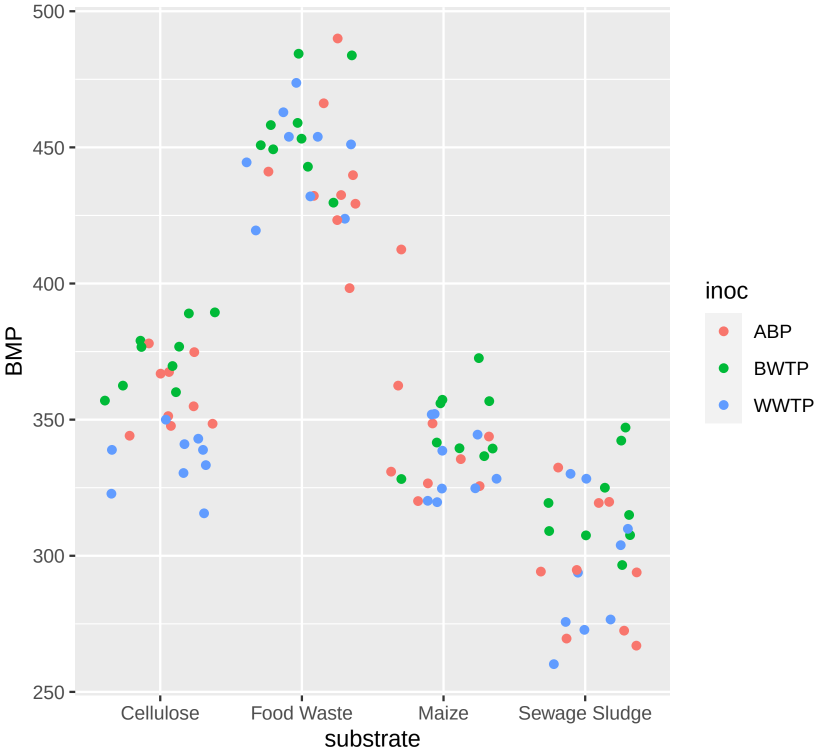
```
# [1] 108   4
```

```
dfsumm(bil)
```

```
#
#  108 rows and 4 columns
#  108 unique rows
#                         substrate   inoc       rep     BMP
# Class                      factor factor character numeric
# Minimum                 Cellulose    ABP      BMP1     260
# Maximum             Sewage Sludge   WWTP      BMP9     490
# Mean                   Food Waste   BWTP      BMP5     362
# Unique (excld. NA)              4      3         9     103
# Missing values                  0      0         0       0
# Sorted                      FALSE  FALSE      TRUE   FALSE
```

Here are the values, with a single point representing a BMP value from a single bottle.

```
ggplot(bil, aes(substrate, BMP, colour = inoc)) +
  geom_jitter(height = 0)
```

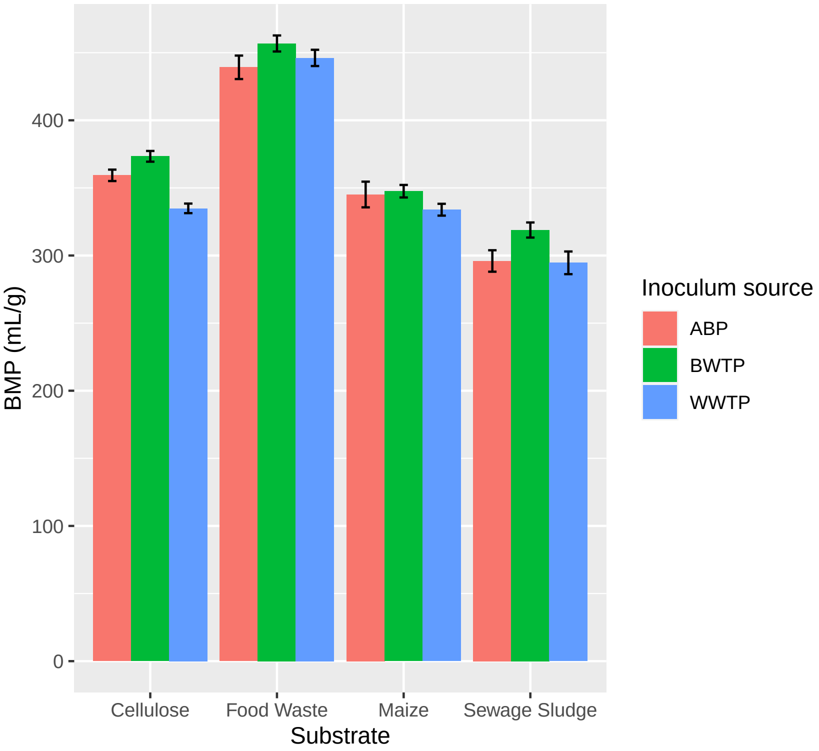


Calculate means and standard deviation.

```
bm <- as.data.frame(summarise(group_by(bil, substrate, inoc), BMP.mn = mean(BMP),
                              BMP.sd = sd(BMP), n = length(BMP)))
bm$BMP.se = bm$BMP.sd / sqrt(bm$n)

bm


#        substrate inoc   BMP.mn   BMP.sd n   BMP.se
# 1      Cellulose  ABP 359.3000 12.65178 9 4.217260
# 2      Cellulose BWTP 373.3556 11.89276 9 3.964254
# 3      Cellulose WWTP 334.8778 10.63329 9 3.544431
# 4     Food Waste  ABP 439.1889 26.05554 9 8.685180
# 5     Food Waste BWTP 456.8111 17.78479 9 5.928262
# 6     Food Waste WWTP 446.1444 18.01694 9 6.005648
# 7          Maize  ABP 345.1222 28.50604 9 9.502014
# 8          Maize BWTP 347.5556 13.87661 9 4.625536
# 9          Maize WWTP 333.8667 13.12355 9 4.374516
# 10 Sewage Sludge  ABP 295.9556 23.81765 9 7.939215
# 11 Sewage Sludge BWTP 318.8444 16.75717 9 5.585724
# 12 Sewage Sludge WWTP 294.5889 25.14202 9 8.380673
```

And plot them.

```
ggplot(bm, aes(substrate, BMP.mn, fill = inoc)) +
  geom_bar(position = position_dodge(), stat = 'identity') +
  geom_errorbar(aes(ymin = BMP.mn - BMP.se, ymax = BMP.mn + BMP.se), position = position_dodge(0.9),
  labs(x = 'Substrate', y = 'BMP (mL/g)', fill = 'Inoculum source')
```



Here is a case where we really do need a statistical analysis to help understand the data.

```
m1 <- lm(BMP ~ substrate * inoc, data = bil)
summary(m1)

#
# Call:
# lm(formula = BMP ~ substrate * inoc, data = bil)
#
# Residuals:
#     Min      1Q  Median      3Q     Max
# -40.889 -11.719  -1.700   9.261  67.378
#
# Coefficients:
#                                  Estimate Std. Error t value Pr(>|t|)
# (Intercept)                       359.300      6.377  56.343  < 2e-16
# substrateFood Waste                79.889      9.018   8.858 4.21e-14
# substrateMaize                    -14.178      9.018  -1.572  0.11922
# substrateSewage Sludge            -63.344      9.018  -7.024 3.10e-10
# inocBWTP                           14.056      9.018   1.559  0.12240
# inocWWTP                          -24.422      9.018  -2.708  0.00801
# substrateFood Waste:inocBWTP        3.567     12.754   0.280  0.78035
# substrateMaize:inocBWTP           -11.622     12.754  -0.911  0.36444
# substrateSewage Sludge:inocBWTP     8.833     12.754   0.693  0.49024
# substrateFood Waste:inocWWTP       31.378     12.754   2.460  0.01567
# substrateMaize:inocWWTP            13.167     12.754   1.032  0.30450
# substrateSewage Sludge:inocWWTP    23.056     12.754   1.808  0.07378
#
# (Intercept)                      ***
# substrateFood Waste             ***
# substrateMaize
# substrateSewage Sludge          ***
# inocBWTP
# inocWWTP                        **
# substrateFood Waste:inocBWTP
# substrateMaize:inocBWTP
# substrateSewage Sludge:inocBWTP
# substrateFood Waste:inocWWTP    *
# substrateMaize:inocWWTP
# substrateSewage Sludge:inocWWTP .
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#
# Residual standard error: 19.13 on 96 degrees of freedom
# Multiple R-squared:  0.8995,Adjusted R-squared:  0.888
# F-statistic: 78.14 on 11 and 96 DF,  p-value: < 2.2e-16
```

```
anova(m1)
```

```
# Analysis of Variance Table
#
# Response: BMP
#                 Df Sum Sq Mean Sq  F value    Pr(>F)
# substrate        3 302030  100677 275.0758 < 2.2e-16 ***
# inoc             2   8804    4402  12.0276 2.181e-05 ***
# substrate:inoc   6   3740     623   1.7031    0.1285
# Residuals       96  35136     366
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

There is clear evidence of an inoculum effect, and a slight suggestion of a possible interaction.

```
m2 <- aov(BMP ~ substrate * inoc, data = bil)
summary(m2)
```

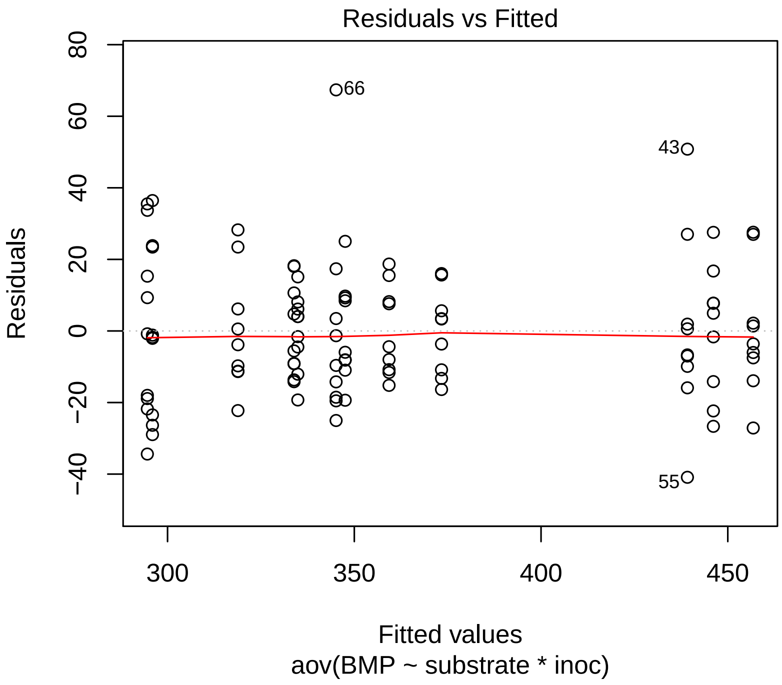
```
#                 Df Sum Sq Mean Sq F value   Pr(>F)
# substrate        3 302030  100677 275.076  < 2e-16 ***
# inoc             2   8804    4402  12.028 2.18e-05 ***
# substrate:inoc   6   3740     623   1.703    0.129
# Residuals       96  35136     366
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

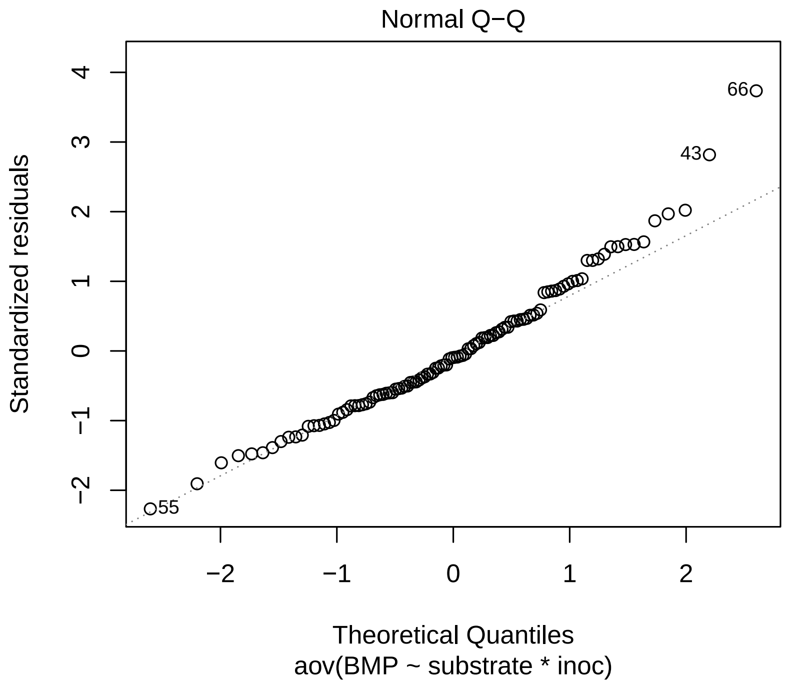
```
TukeyHSD(m2, 'inoc')
```

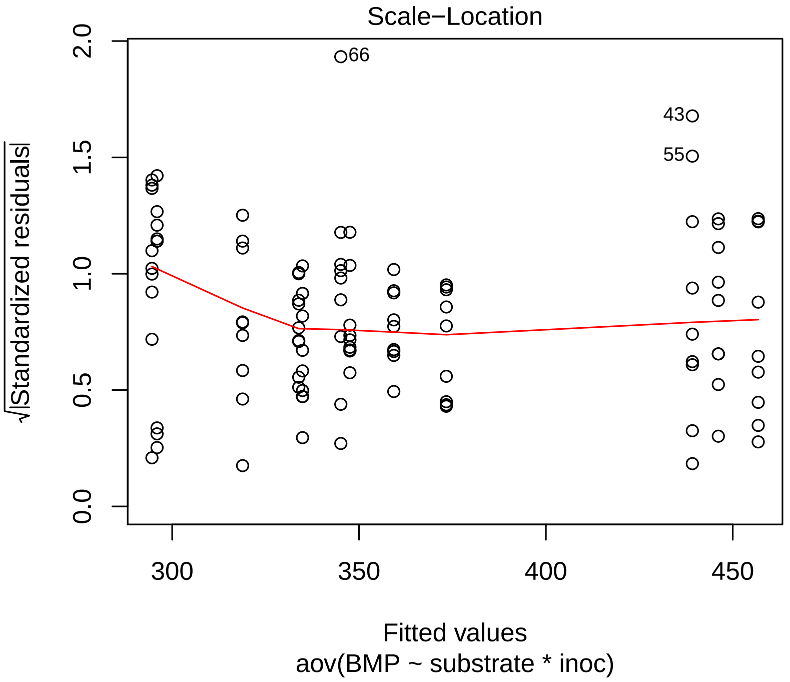
```
#   Tukey multiple comparisons of means
#     95% family-wise confidence level
#
# Fit: aov(formula = BMP ~ substrate * inoc, data = bil)
```

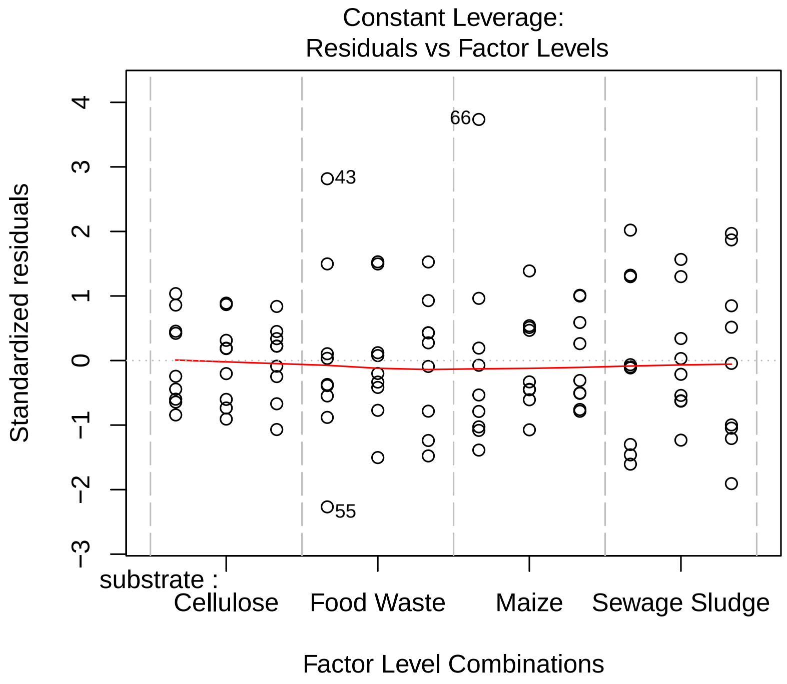
```
#
# $inoc
#                 diff        lwr        upr     p adj
# BWTP-ABP   14.250000   3.515301  24.984699 0.0059271
# WWTP-ABP   -7.522222 -18.256921   3.212477 0.2227058
# WWTP-BWTP -21.772222 -32.506921 -11.037523 0.0000154
```

```
plot(m2, ask = FALSE)
```



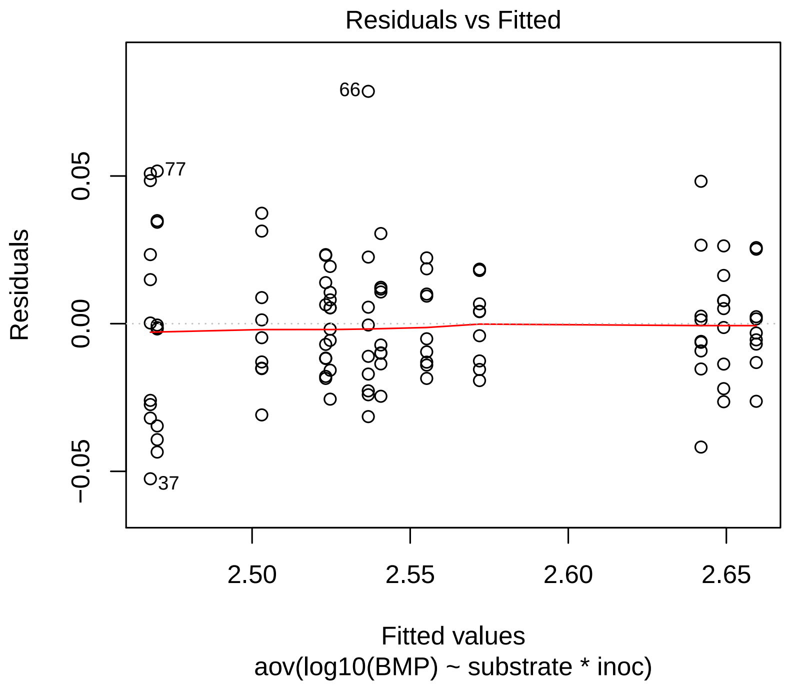

Normal Q−Q
Standardized residuals
4
3
2
1
0
−1
−2
66
43
55
−2
−1
0
1
2
Theoretical Quantiles
aov(BMP ~ substrate * inoc)

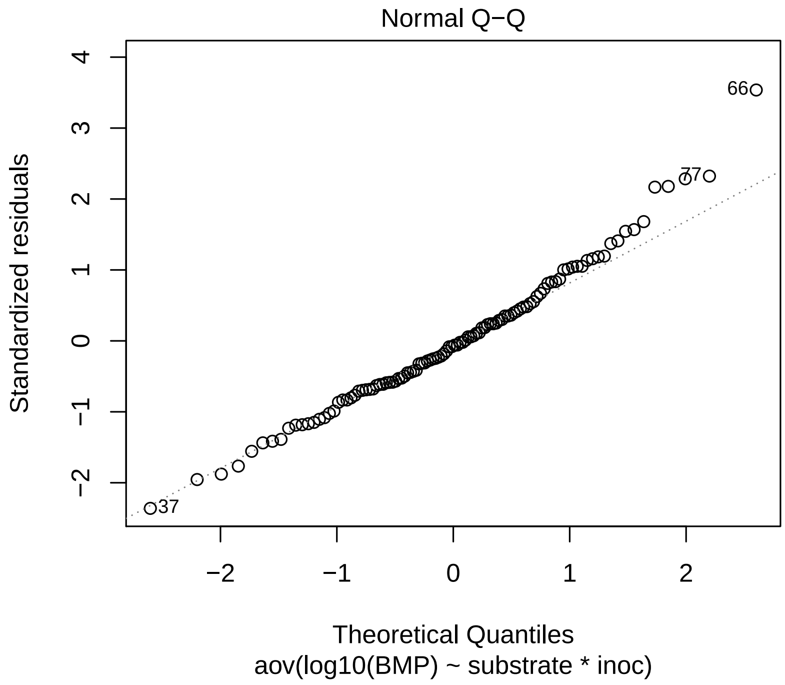

Scale−Location
√|Standardized residuals|
2.0
1.5
1.0
0.5
0.0
66
43
55
300
350
400
450
Fitted values
aov(BMP ~ substrate * inoc)

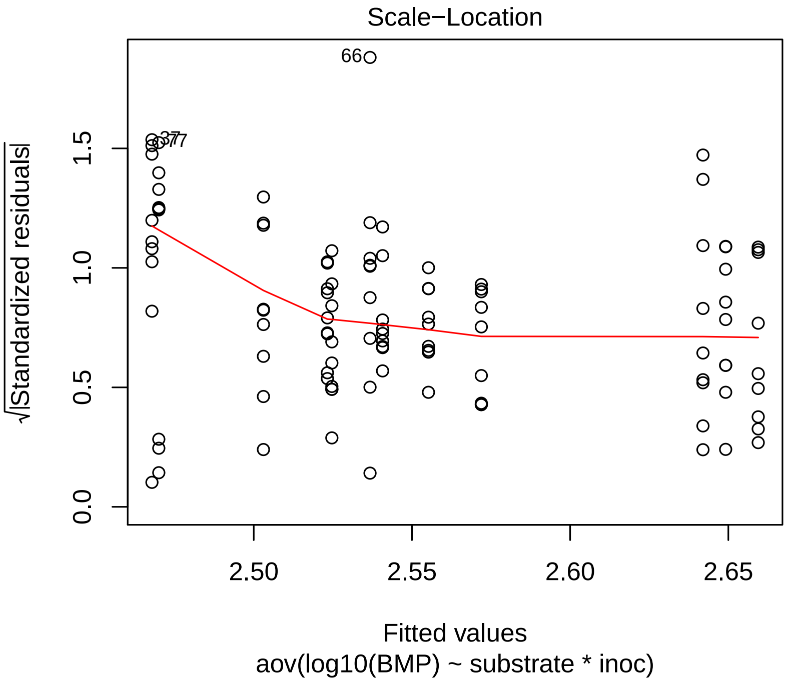


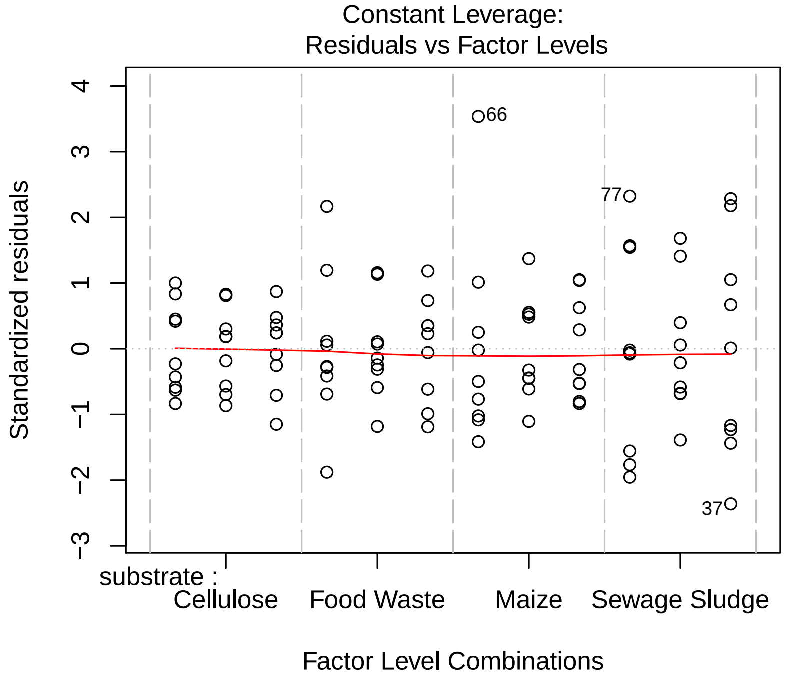
```
m3 <- aov(log10(BMP) ~ substrate * inoc, data = bil)
summary(m3)

#                 Df Sum Sq Mean Sq F value   Pr(>F)    
# substrate        3 0.4081 0.13604 244.417  < 2e-16 ***
# inoc             2 0.0141 0.00703  12.623 1.36e-05 ***
# substrate:inoc   6 0.0062 0.00103   1.853    0.097 .  
# Residuals       96 0.0534 0.00056                     
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(tr <- TukeyHSD(m3, 'inoc'))

#   Tukey multiple comparisons of means
#     95% family-wise confidence level
#
# Fit: aov(formula = log10(BMP) ~ substrate * inoc, data = bil)
#
# $inoc
#                   diff          lwr         upr     p adj
# BWTP-ABP   0.017803269  0.004565351  0.03104119 0.0052233
# WWTP-ABP  -0.009747578 -0.022985495  0.00349034 0.1911260
# WWTP-BWTP -0.027550847 -0.040788764 -0.01431293 0.0000092

100 * (10^tr$inoc[, 'diff'] - 1)

#  BWTP-ABP  WWTP-ABP WWTP-BWTP 
#  4.184538 -2.219462 -6.146785
```

```
plot(m3, ask = FALSE)
```









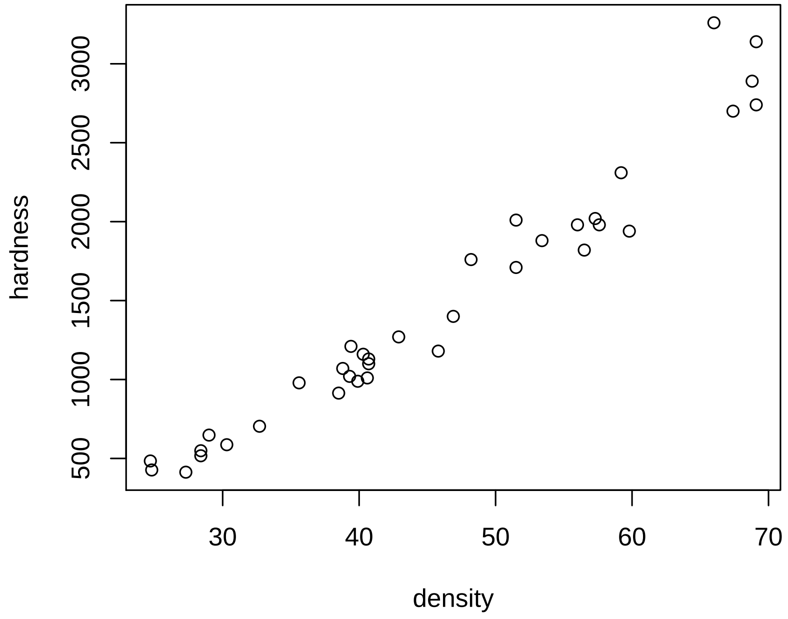
We can conclude that the BWTP inoculum resulted in BMP values about 4-6% higher than the other two.

3 Problem 2. Wood hardness and density

```
hard <- read.csv("data/janka.csv")
dfsumm(hard)


#
#  36 rows and 2 columns
#  36 unique rows
#                    density hardness
# Class              numeric  integer
# Minimum               24.7      413
# Maximum               69.1     3260
# Mean                  45.7     1180
# Unique (excld. NA)      32       35
# Missing values           0        0
# Sorted                TRUE    FALSE
```

Let's start out by seeing what the data look like.

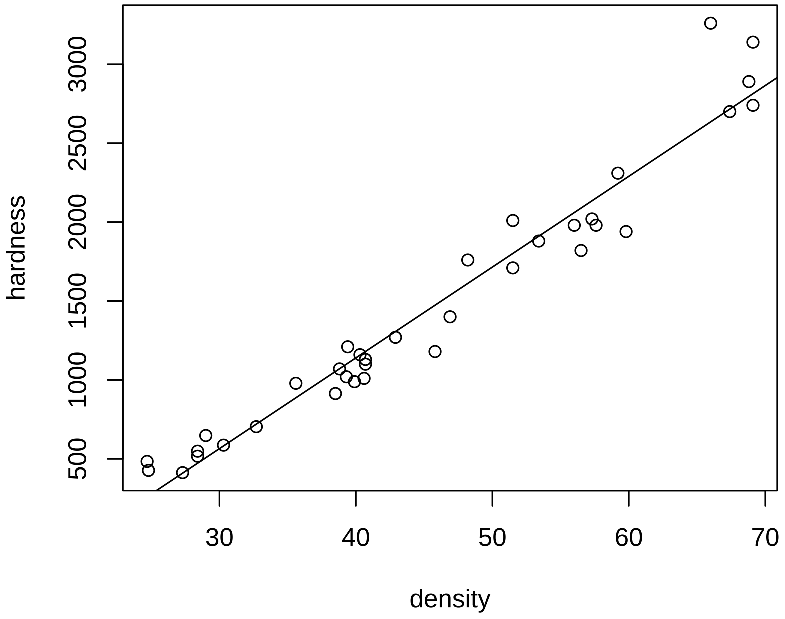
```
plot(hardness ~ density, data = hard)
```



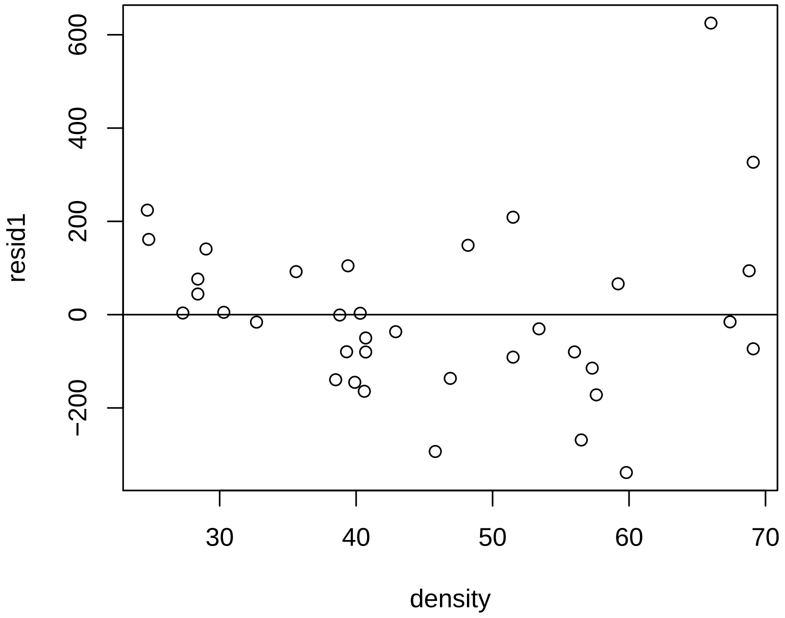
We might be interested in doing two things with these data: determining if wood hardness (difficult to measure) is related to wood density (easy to measure), and, if so, predicting hardness from the density. Are these data experimental or observational? Try to fit an appropriate regression model to these data, and take a look at the residuals to check the structure. Can you improve it?

```
m1 <- lm(hardness ~ density, data = hard)
summary(m1)


#
# Call:
# lm(formula = hardness ~ density, data = hard)
#
# Residuals:
#     Min      1Q  Median      3Q     Max
# -338.40  -96.98  -15.71   92.71  625.06
#
# Coefficients:
#              Estimate Std. Error t value Pr(>|t|)
# (Intercept) -1160.500    108.580  -10.69 2.07e-12 ***
# density        57.507      2.279   25.24  < 2e-16 ***
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#
# Residual standard error: 183.1 on 34 degrees of freedom
# Multiple R-squared:  0.9493,Adjusted R-squared:  0.9478
# F-statistic:   637 on 1 and 34 DF,  p-value: < 2.2e-16


hard$pred1 <- predict(m1)
hard$resid1 <- resid(m1)
```

```
plot(hardness ~ density, data = hard)
abline(m1)
```

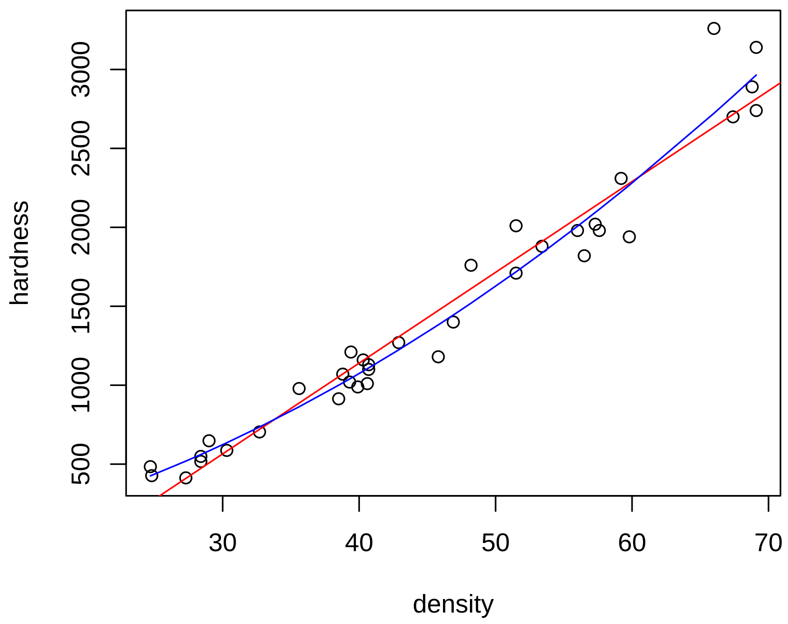


```
plot(resid1 ~ density, data = hard)
abline(h = 0)
```

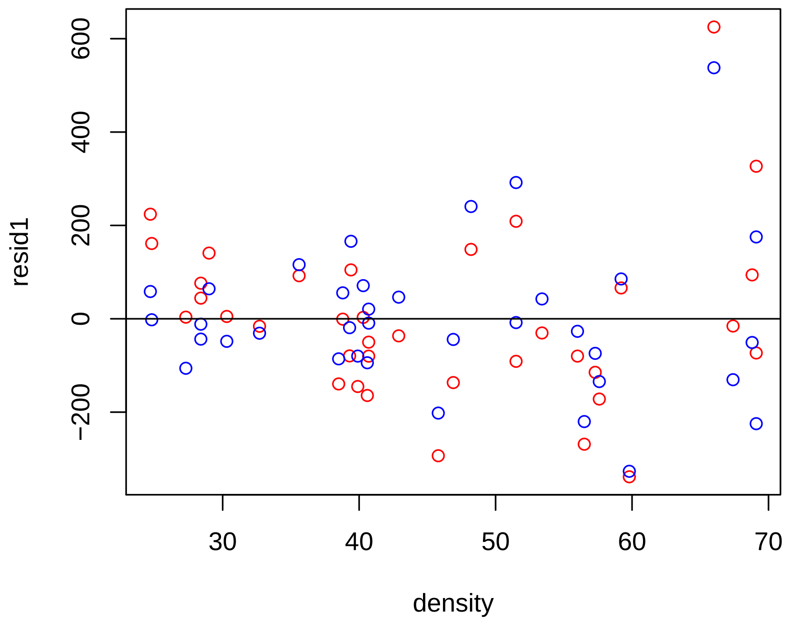


```
m2 <- lm(hardness ~ poly(density, 3), data = hard)
summary(m2)


#
# Call:
# lm(formula = hardness ~ poly(density, 3), data = hard)
#
# Residuals:
#     Min      1Q  Median      3Q     Max
# -310.98  -92.52  -14.94   61.41  537.92
#
# Coefficients:
#                    Estimate Std. Error t value Pr(>|t|)
# (Intercept)         1469.47      27.29  53.841  < 2e-16 ***
# poly(density, 3)1   4620.14     163.76  28.213  < 2e-16 ***
# poly(density, 3)2    525.40     163.76   3.208  0.00303 **
# poly(density, 3)3     72.14     163.76   0.441  0.66252
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#
# Residual standard error: 163.8 on 32 degrees of freedom
# Multiple R-squared:  0.9618,Adjusted R-squared:  0.9583
# F-statistic: 268.8 on 3 and 32 DF,  p-value: < 2.2e-16


m2 <- lm(hardness ~ poly(density, 2), data = hard)
hard$pred2 <- predict(m2)
hard$resid2 <- resid(m2)
```

```
plot(hardness ~ density, data = hard)
abline(m1, col = 'red')
lines(pred2 ~ density, data = hard, col = 'blue')
```

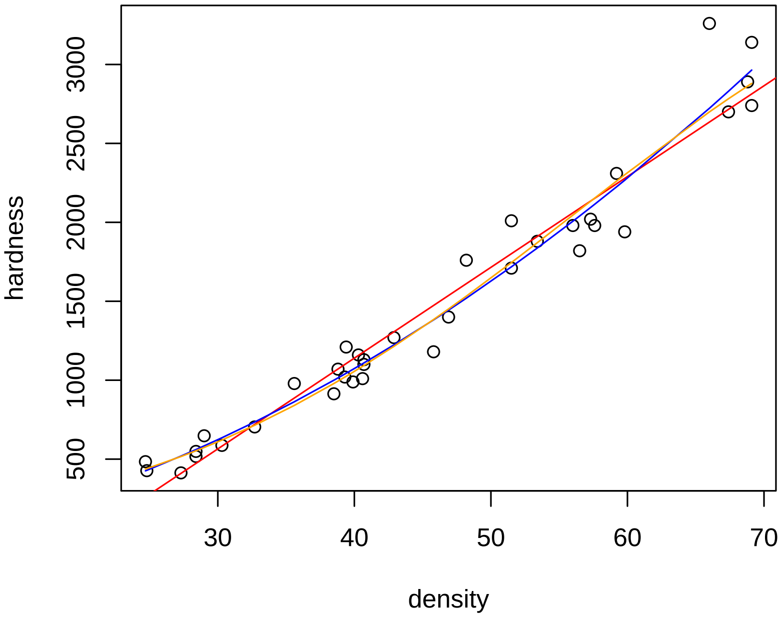


```
plot(resid1 ~ density, data = hard, col = 'red')
points(resid2 ~ density, data = hard, col = 'blue')
abline(h = 0)
```

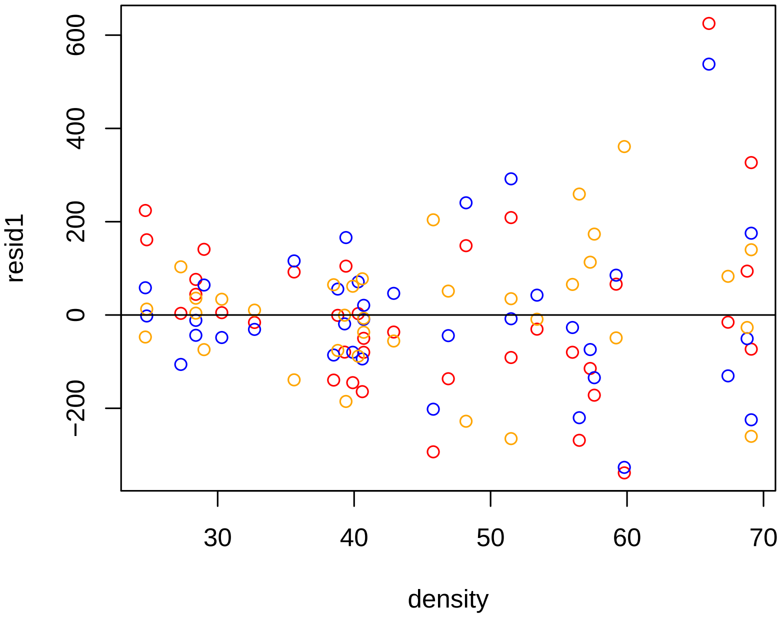


```
m3 <- lm(log10(hardness) ~ poly(density, 2), data = hard)
summary(m3)


#
# Call:
# lm(formula = log10(hardness) ~ poly(density, 2), data = hard)
#
# Residuals:
#       Min        1Q    Median        3Q       Max
# -0.096983 -0.024792 -0.004795  0.032573  0.081955
#
# Coefficients:
#                     Estimate Std. Error t value Pr(>|t|)
# (Intercept)         3.099195   0.007294 424.896  < 2e-16 ***
# poly(density, 2)1  1.470617   0.043764  33.603  < 2e-16 ***
# poly(density, 2)2 -0.234322   0.043764  -5.354 6.49e-06 ***
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#
# Residual standard error: 0.04376 on 33 degrees of freedom
# Multiple R-squared:  0.9723,Adjusted R-squared:  0.9706
# F-statistic: 578.9 on 2 and 33 DF,  p-value: < 2.2e-16


hard$pred3 <- 10^predict(m3)
hard$resid3 <- hard$pred3 - hard$hardness
```

```
plot(hardness ~ density, data = hard)
abline(m1, col = 'red')
lines(pred2 ~ density, data = hard, col = 'blue')
lines(pred3 ~ density, data = hard, col = 'orange')
```



```
plot(resid1 ~ density, data = hard, col = 'red')
points(resid2 ~ density, data = hard, col = 'blue')
points(resid3 ~ density, data = hard, col = 'orange')
abline(h = 0)
```



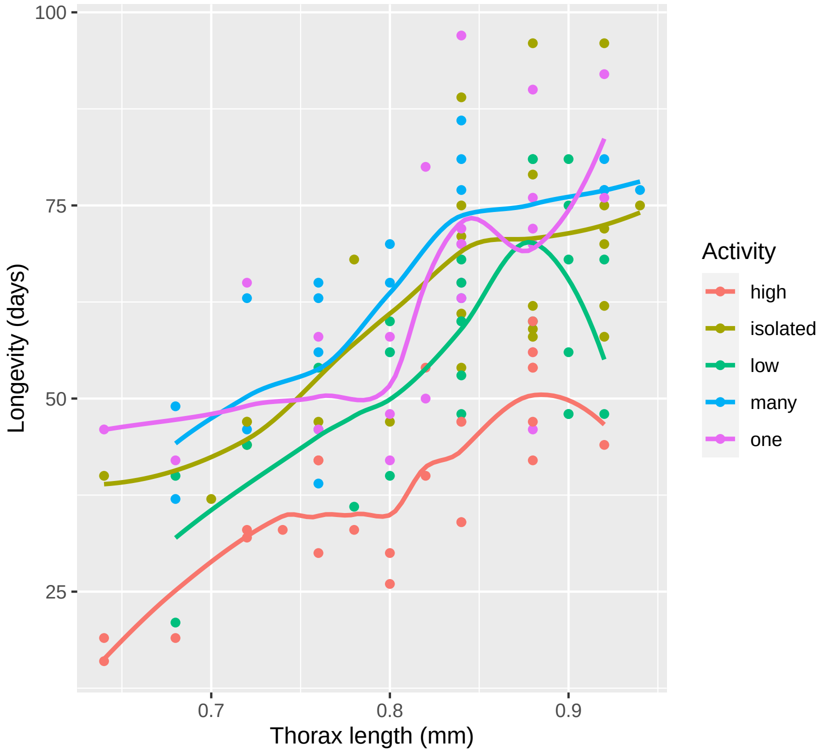
4 Problem 3. Fruit fly longevity and sexual activity

The data in the file fruitfly.csv are from an experiment on fruitfly longevity and are also from Faraway [2005]. The original objective of this famous experiment was to assess the effect of sexual activity (manipulated by controlling the number of females placed with a single male, `activity` column) on fruitfly longevity (how long the flies live, `longevity` column). But longevity is known to be correlated with thorax length (`thorax` column.

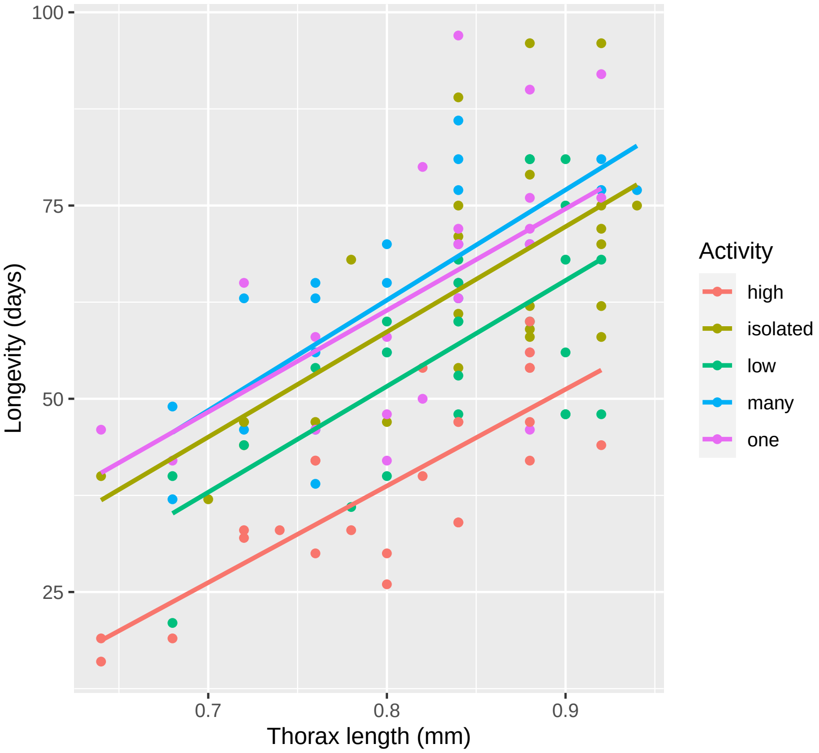
```
ff <- read.csv('data/fruitfly.csv')
head(ff)


#   thorax longevity activity
# 1   0.68        37     many
# 2   0.68        49     many
# 3   0.72        46     many
# 4   0.72        63     many
# 5   0.76        39     many
# 6   0.76        46     many
```

1. How might you plot these data to assess the effect of activity?
2. How can you fit a statistical model that utilizes the correlation with thorax length to increase power?
3. What approach should you use to compare the levels of `activity` to each other?

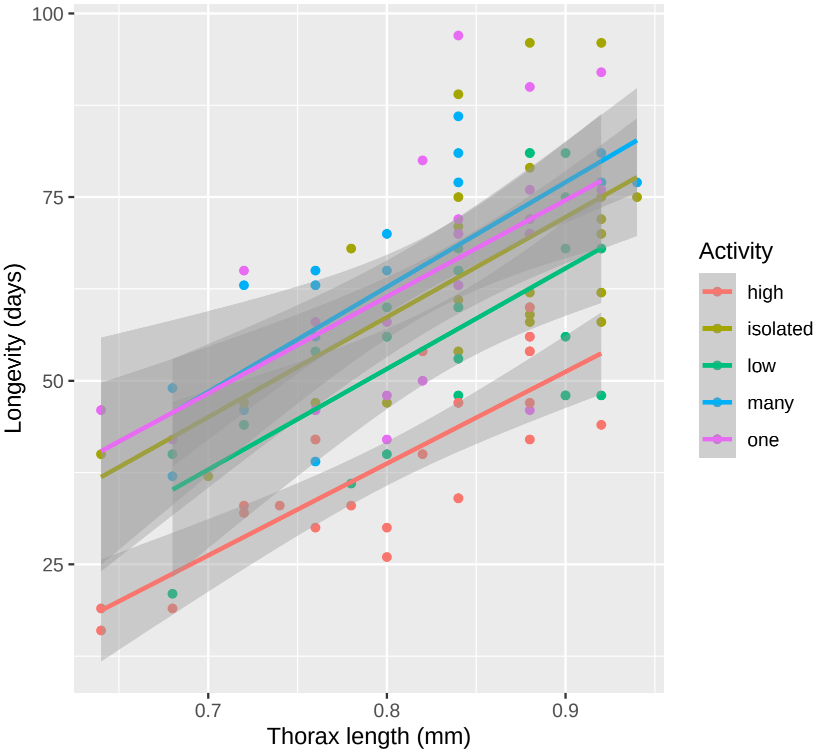
```
ggplot(ff, aes(thorax, longevity, colour = activity)) +
  geom_point() +
  geom_smooth(se = FALSE) +
  labs(x = 'Thorax length (mm)', y = 'Longevity (days)', colour = 'Activity')


# ‘geom_smooth()‘ using method = ’loess’ and formula ’y ~ x’

# Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric, :
pseudoinverse used at 0.84

# Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric, :
neighborhood radius 0.04

# Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric, :
reciprocal condition number  8.6863e-22
```



```
ggplot(ff, aes(thorax, longevity, colour = activity)) +
  geom_point() +
  geom_smooth(method = lm, se = FALSE) +
  labs(x = 'Thorax length (mm)', y = 'Longevity (days)', colour = 'Activity')


# ‘geom_smooth()‘ using formula ’y ~ x’
```



```
ggplot(ff, aes(thorax, longevity, colour = activity)) +
  geom_point() +
  geom_smooth(method = lm) +
  labs(x = 'Thorax length (mm)', y = 'Longevity (days)', colour = 'Activity')


# `geom_smooth()` using formula 'y ~ x'
```



```
levels(ff$activity)


# [1] "high"     "isolated" "low"      "many"     "one"
```

First level will be reference. Let's change it to isolated.

```
ff$activity <- relevel(ff$activity, ref= 'isolated')
```

```
m1 <- lm(longevity ~ activity * thorax, data = ff)
anova(m1)


# Analysis of Variance Table
#
# Response: longevity
#                    Df   Sum Sq Mean Sq F value  Pr(>F)
# activity            4 12269.5  3067.4  26.728 1.2e-15 ***
# thorax              1 12368.4 12368.4 107.774 < 2e-16 ***
# activity:thorax   4    24.3      6.1   0.053  0.9947
# Residuals        114 13083.0   114.8
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
m2 <- lm(longevity ~ activity + thorax, data = ff)
anova(m2)
```

```
# Analysis of Variance Table
#
# Response: longevity
#            Df Sum Sq Mean Sq F value    Pr(>F)
# activity    4  12270  3067.4  27.614 3.481e-16 ***
# thorax      1  12368 12368.4 111.348 < 2.2e-16 ***
# Residuals 118  13107   111.1
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(m2)
```

```
#
# Call:
# lm(formula = longevity ˜ activity + thorax, data = ff)
#
# Residuals:
#     Min      1Q  Median      3Q     Max
# -26.108  -7.014  -1.101   6.234  30.265
#
# Coefficients:
#              Estimate Std. Error t value Pr(>|t|)
# (Intercept)   -48.749     10.850  -4.493 1.65e-05 ***
# activityhigh  -20.004      3.016  -6.632 1.05e-09 ***
# activitylow    -7.015      2.981  -2.353   0.0203 *
# activitymany    4.139      3.027   1.367   0.1741
# activityone     2.637      2.984   0.884   0.3786
# thorax        134.341     12.731  10.552  < 2e-16 ***
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#
# Residual standard error: 10.54 on 118 degrees of freedom
# Multiple R-squared:  0.6527,Adjusted R-squared:  0.638
# F-statistic: 44.36 on 5 and 118 DF,  p-value: < 2.2e-16
```

We can use Bonferroni adjustment, 0.05 / 5 = 0.01. So only `high` level is clearly different–20 days shorter longevity, which is a lot!

```
confint(m2)
```

```
#                   2.5 %     97.5 %
# (Intercept)  -70.235303 -27.263477
# activityhigh -25.976247 -14.031174
# activitylow  -12.918256  -1.111636
# activitymany  -1.855011  10.132389
# activityone   -3.271842   8.546143
# thorax       109.130197 159.552553
```

Strange that “many” level is so different from others.

```
confint(m2)
```

```
#                   2.5 %     97.5 %
# (Intercept)  -70.235303 -27.263477
# activityhigh -25.976247 -14.031174
# activitylow  -12.918256  -1.111636
# activitymany  -1.855011  10.132389
# activityone   -3.271842   8.546143
# thorax       109.130197 159.552553
```

5 Problem 4: Growth and nitrate accumulation by *Lemna minor*

Duckweeds are very tiny floating plants that can be used for wastewater treatment and recovery of nitrogen. Harvested material can be used as an animal feed. Devlamynck et al. [2020] measured biomass production and nitrate accumulation in a duckweed species *Lemna minor*. The data are in lemna.csv. Use them to explore the following questions.

1. Did medium affect growth (`grow`)?
2. Did medium affect NO_3^- accumulation (`NO3.accum`)?
3. Is NO_3^- accumulation related to NO_3^- concentration in the medium (`NO3.med`)?

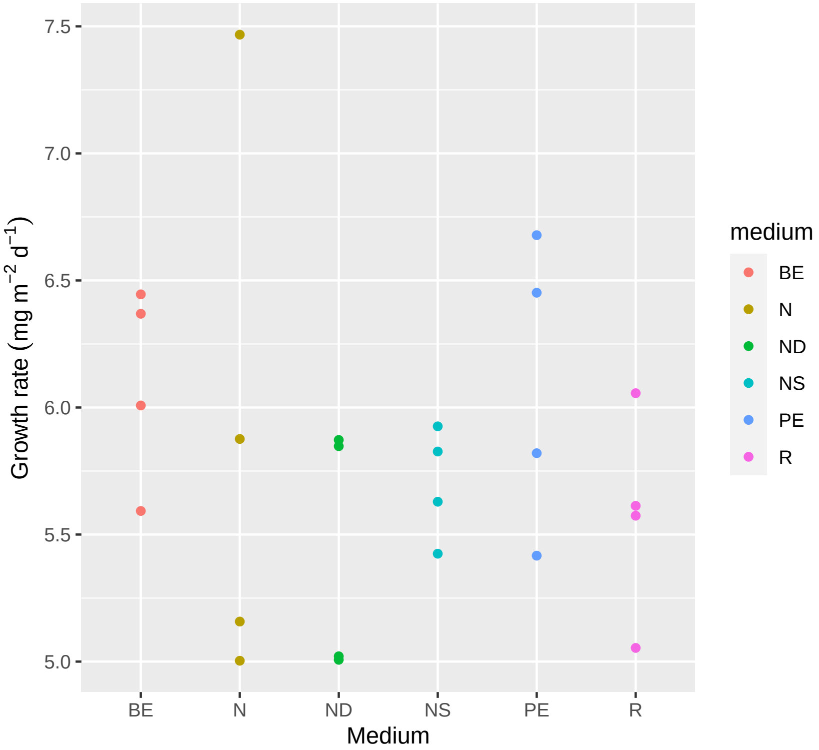
```
lem <- read.csv('data/lemna.csv')
```

```
summary(lem)
```

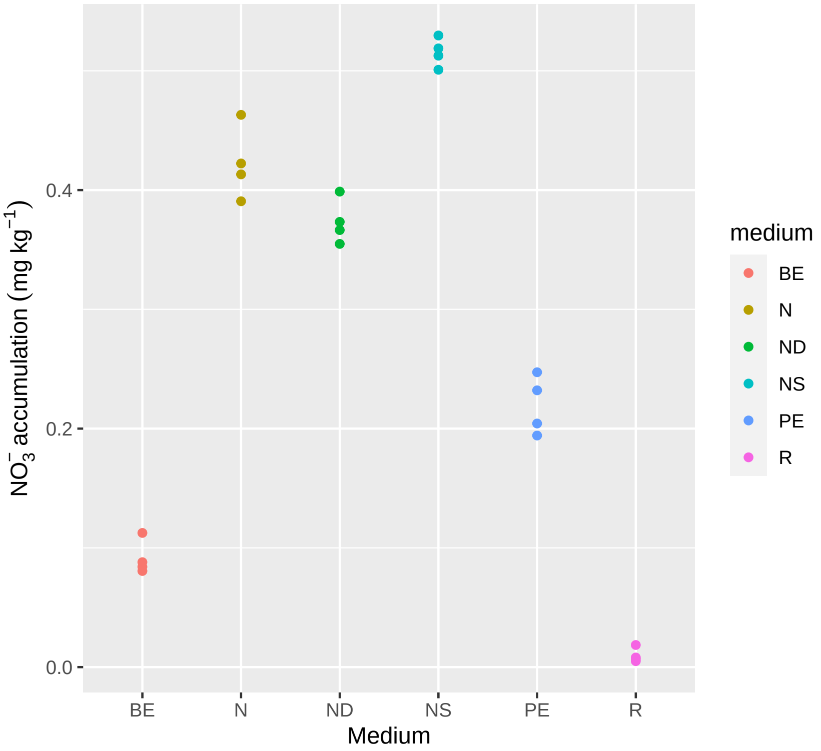
```
#                    med.descrip medium      grow
#  Concentrated C medium   :4    BE:4   Min.   :5.003
#  Diluted N medium        :4    N :4   1st Qu.:5.423
#  Fish wastewater effluent:4    ND:4   Median :5.823
#  Pig wastewater effluent :4    NS:4   Mean   :5.797
#  Rainwater               :4    PE:4   3rd Qu.:6.020
#  Synthetic N medium      :4    R :4   Max.   :7.467
#    NO3.accum            pH.med         NO3.med
#  Min.   :0.005025   Min.   :5.760   Min.   : 0.009594
#  1st Qu.:0.087042   1st Qu.:6.388   1st Qu.: 2.215323
#  Median :0.301076   Median :7.390   Median : 4.138554
#  Mean   :0.271930   Mean   :7.461   Mean   : 7.795348
#  3rd Qu.:0.415473   3rd Qu.:8.525   3rd Qu.: 9.410879
#  Max.   :0.529639   Max.   :9.632   Max.   :27.129694
```

```
library(ggplot2)
```

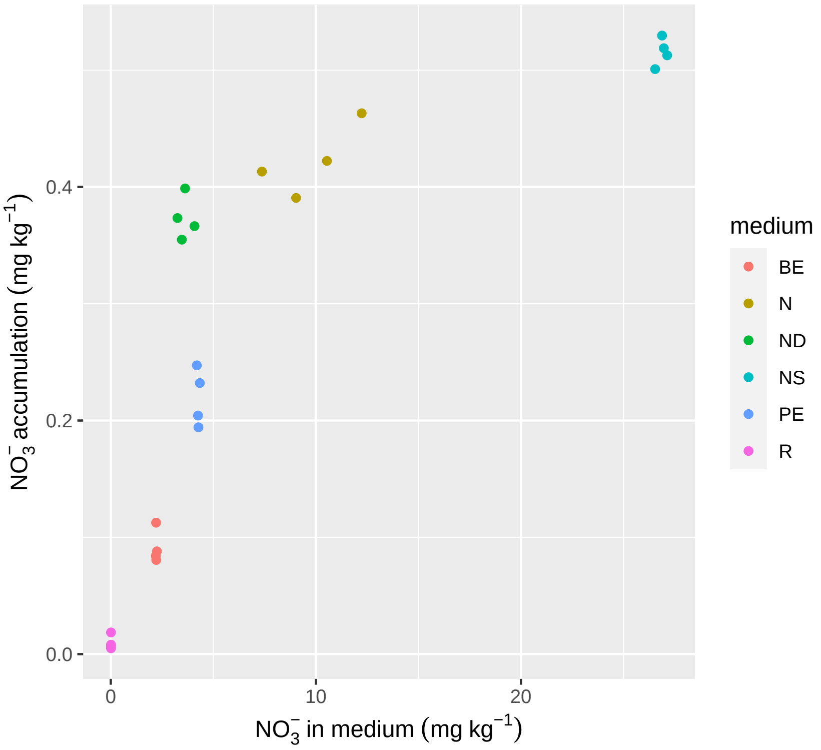
```
ggplot(lem, aes(medium, grow, colour = medium)) +
  geom_point() +
  labs(x = 'Medium', y = expression('Growth rate'~(mg~m^'-2'~d^'-1')))
```



```
ggplot(lem, aes(medium, NO3.accum, colour = medium)) +
  geom_point() +
  labs(x = 'Medium', y = expression(NO[3]^'-'~'accumulation'~(mg~kg^'-1')))
```



```
ggplot(lem, aes(NO3.med, NO3.accum, colour = medium)) +
  geom_point() +
  labs(x = expression(NO[3]^'-'~'in medium'~(mg~kg^'-1')), y = expression(NO[3]^'-'~'accumulation'~(mg
```



First growth. Check plot–no clear effect, no stats needed. We can calculate average and sd at least.

```
lemsum <- as.data.frame(summarise(group_by(lem, medium),
                                  grow.mean = mean(grow), grow.sd = sd(grow),
                                  NO3.med.mean = mean(NO3.med), NO3.med.sd = sd(NO3.med),
                                  NO3.accum.mean = mean(NO3.accum),
                                  NO3.accum.sd = sd(NO3.accum)))

lemsum


#   medium grow.mean   grow.sd NO3.med.mean   NO3.med.sd NO3.accum.mean
# 1     BE  6.103780 0.3904054  2.217319070 0.0227861966    0.091324595
# 2      N  5.876119 1.1268901  9.794380141 2.0779095490    0.422308931
# 3     ND  5.437048 0.4885779  3.604850206 0.3524655526    0.373366822
# 4     NS  5.701615 0.2220019 26.879508387 0.2478468075    0.515520870
# 5     PE  6.091729 0.5780991  4.266298899 0.0601486883    0.219471297
# 6      R  5.574388 0.4101557  0.009733427 0.0001607362    0.009586386
#   NO3.accum.sd
# 1  0.014458101
# 2  0.030275823
# 3  0.018557293
# 4  0.011995311
# 5  0.024499245
# 6  0.006123297
```

For nitrate accumulation, there seem to be effects.

```
levels(lem$medium)


# [1] "BE" "N"  "ND" "NS" "PE" "R"


lem$medium <- relevel(lem$medium, ref= 'R')
```

```
m1 <- lm(NO3.accum ~ medium, data = lem)
anova(m1)


# Analysis of Variance Table
#
# Response: NO3.accum
#           Df  Sum Sq  Mean Sq F value    Pr(>F)
# medium     5 0.78574 0.157147  418.76 < 2.2e-16 ***
# Residuals 18 0.00675 0.000375
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1


summary(m1)


#
# Call:
```

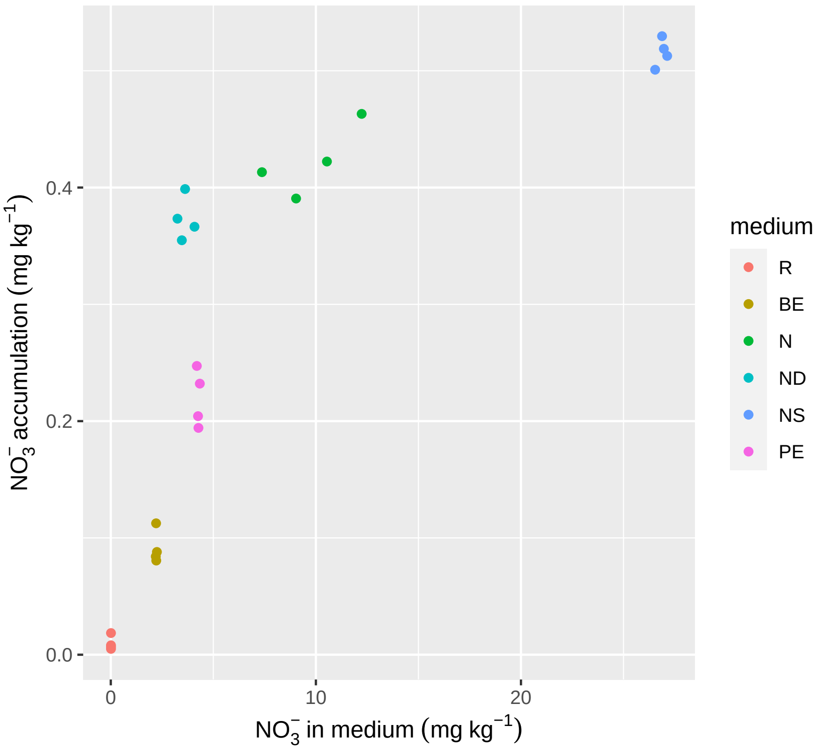
```
# lm(formula = NO3.accum ~ medium, data = lem)
#
# Residuals:
#       Min        1Q    Median        3Q       Max
# -0.031673 -0.009509 -0.002861  0.009905  0.040788
#
# Coefficients:
#             Estimate Std. Error t value Pr(>|t|)
# (Intercept) 0.009586   0.009686   0.990    0.335
# mediumBE    0.081738   0.013698   5.967 1.21e-05 ***
# mediumN     0.412723   0.013698  30.130  < 2e-16 ***
# mediumND    0.363780   0.013698  26.557 6.87e-16 ***
# mediumNS    0.505934   0.013698  36.935  < 2e-16 ***
# mediumPE    0.209885   0.013698  15.322 9.03e-12 ***
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#
# Residual standard error: 0.01937 on 18 degrees of freedom
# Multiple R-squared:  0.9915,Adjusted R-squared:  0.9891
# F-statistic: 418.8 on 5 and 18 DF,  p-value: < 2.2e-16
```

As expected, very clear differences. Does it matter exactly which ones differed? Seems everything was higher than R.

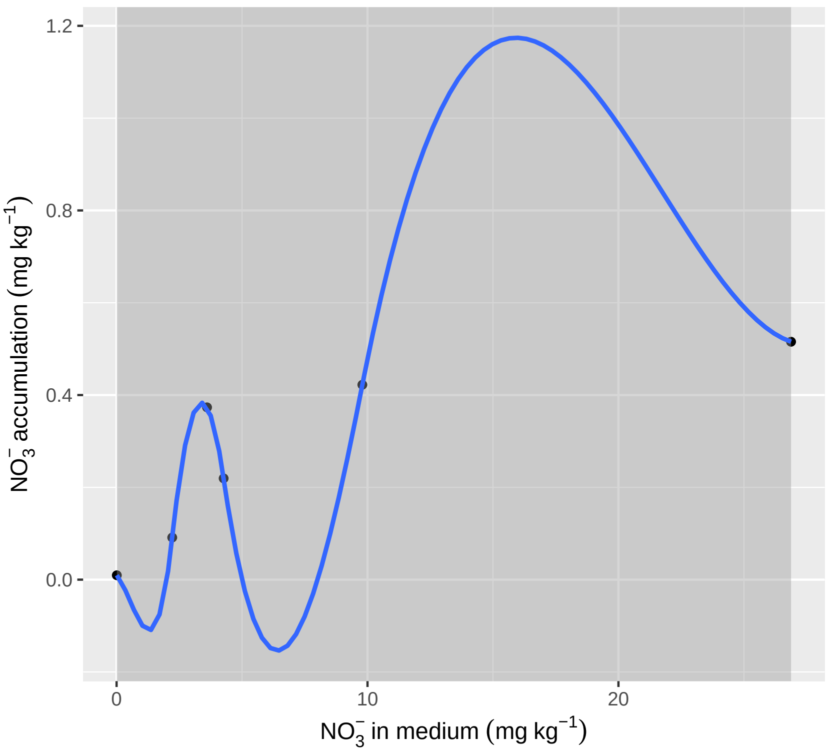
```
m2 <- aov(NO3.accum ~ medium, data = lem)
anova(m2)


# Analysis of Variance Table
#
# Response: NO3.accum
#           Df  Sum Sq  Mean Sq F value    Pr(>F)
# medium     5 0.78574 0.157147  418.76 < 2.2e-16 ***
# Residuals 18 0.00675 0.000375
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
TukeyHSD(m2)


#   Tukey multiple comparisons of means
#     95% family-wise confidence level
#
# Fit: aov(formula = NO3.accum ~ medium, data = lem)
#
# $medium
#               diff         lwr          upr     p adj
# BE-R   0.08173821  0.03820541  0.125271009 0.0001514
# N-R    0.41272254  0.36918974  0.456255345 0.0000000
# ND-R   0.36378044  0.32024763  0.407313236 0.0000000
# NS-R   0.50593448  0.46240168  0.549467284 0.0000000
# PE-R   0.20988491  0.16635211  0.253417711 0.0000000
# N-BE   0.33098434  0.28745154  0.374517136 0.0000000
# ND-BE  0.28204223  0.23850943  0.325575027 0.0000000
# NS-BE  0.42419628  0.38066347  0.467729076 0.0000000
# PE-BE  0.12814670  0.08461390  0.171679503 0.0000003
# ND-N  -0.04894211 -0.09247491 -0.005409309 0.0224993
# NS-N   0.09321194  0.04967914  0.136744740 0.0000291
# PE-N  -0.20283763 -0.24637043 -0.159304833 0.0000000
# NS-ND  0.14215405  0.09862125  0.185686849 0.0000001
# PE-ND -0.15389552 -0.19742832 -0.110362724 0.0000000
# PE-NS -0.29604957 -0.33958237 -0.252516773 0.0000000
```

```
ggplot(lem, aes(NO3.med, NO3.accum, colour = medium)) +
  geom_point() +
  labs(x = expression(NO[3]^'-'~'in medium'~(mg~kg^'-1')), y = expression(NO[3]^'-'~'accumulation'~(mg
```

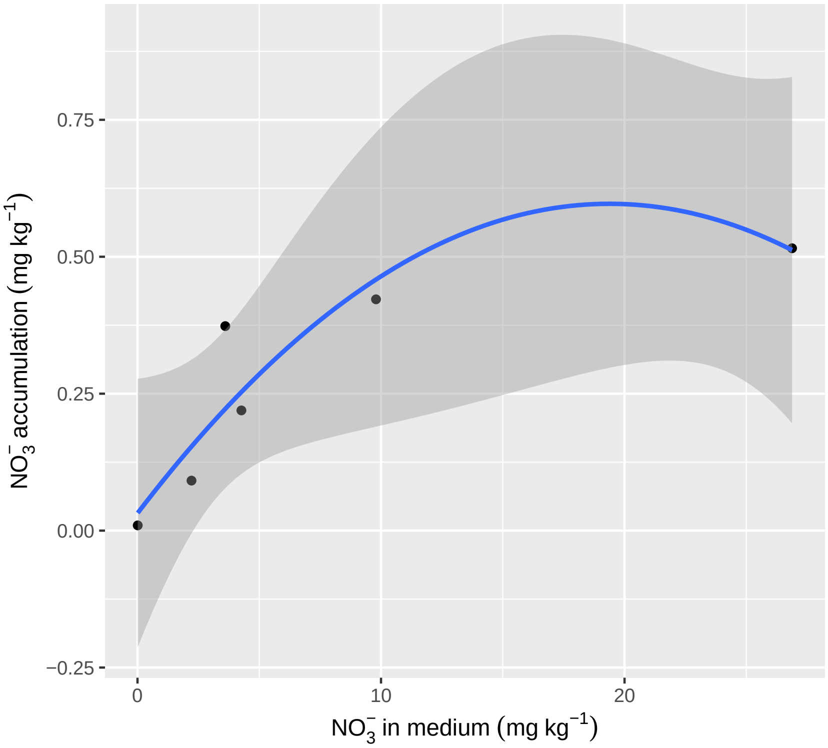


```
ggplot(lemsum, aes(NO3.med.mean, NO3.accum.mean)) +
  geom_point() +
  geom_smooth() +
  labs(x = expression(NO[3]^'-'~'in medium'~(mg~kg^'-1')), y = expression(NO[3]^'-'~'accumulation'~(mg

# ‘geom_smooth()‘ using method = ’loess’ and formula ’y ~ x’
```



Whoa! Overfitting by default!

```
ggplot(lemsum, aes(NO3.med.mean, NO3.accum.mean)) +
  geom_point() +
  geom_smooth(method = lm, formula = y ~ poly(x, 2)) +
  labs(x = expression(NO[3]^'-'~'in medium'~(mg~kg^'-1')), y = expression(NO[3]^'-'~'accumulation'~(mg
```



6 Bibliography

R. Devlamynck, M. Fernandes de Souza, M. Bog, J. Leenknegt, M. Eeckhout, and E. Meers. Effect of the growth medium composition on nitrate accumulation in the novel protein crop Lemna minor. *Ecotoxicology and Environmental Safety*, 206:111380, Dec. 2020. ISSN 0147-6513. doi: 10.1016/j.ecoenv.2020.111380. URL https://www.sciencedirect.com/science/article/pii/S0147651320312173.

J. J. Faraway. *Linear Models with R.* Number v. 63 in Texts in Statistical Science. Chapman & Hall/CRC, Boca Raton, 2005. ISBN 1-58488-425-8.

K. Koch, T. Lippert, and J. E. Drewes. The role of inoculum's origin on the methane yield of different substrates in biochemical methane potential (BMP) tests. *Bioresource Technology*, 243 (Supplement C):457–463, Nov. 2017. ISSN 0960-8524. doi: 10.1016/j.biortech.2017.06.142.