Fitness shifts accompany amino acid changes

Andrew Nguyen

2016-February-27

Questions/Objectives

Hypotheses

load libraries first

```r
library(ggplot2)
library(tidyr)
library(rpart)
library(rpart.plot)
```

loading in file and changing to long format

PYR drug

```r
#read in file
#malaria<-read.csv("20160227_aa_shifts_drug.csv")
malaria<-read.csv("20160227_aa_shifts_drug_scaled_fitness.csv")
#change alleles to factors
malaria$starting_allele<-as.factor(malaria$starting_allele)
malaria$end_allele<-as.factor(malaria$end_allele)
#visualizing the whole dataset
str(malaria)
```

```
## 'data.frame':    32 obs. of  14 variables:
##  $ start          : Factor w/ 4 levels "first","fourth",..: 1 1 1 1 1 1 1 1 3 3 ...
##  $ starting_allele: Factor w/ 15 levels "0","1","10","11",..: 1 5 3 2 7 6 4 8 1 9 ...
##  $ end_allele     : Factor w/ 15 levels "1","10","11",..: 8 12 10 9 14 13 11 15 4 12 ...
##  $ changed        : Factor w/ 32 levels "A1","A2","A3",..: 1 2 3 4 5 6 7 8 9 10 ...
##  $ drug1          : num  -0.1996 -0.0693 0.0564 -0.0649 0.0379 ...
##  $ drug2          : num  -0.1912 -0.0657 0.0577 -0.0625 0.0387 ...
##  $ drug3          : num  -0.1644 -0.054 0.0619 -0.0544 0.0413 ...
##  $ drug4          : num  -0.06273 -0.00603 0.0809 -0.01681 0.05297 ...
##  $ drug5          : num  0.211 0.144 0.176 0.142 0.109 ...
##  $ drug6          : num  0.562 0.364 0.718 0.491 0.429 ...
##  $ drug7          : num  0.75 0.488 2.921 0.761 2.19 ...
##  $ drug8          : num  0.807 0.527 6.127 0.858 6.656 ...
##  $ drug9          : num  0.821 0.537 7.771 0.883 10.224 ...
##  $ drug10         : num  0.825 0.539 8.254 0.889 11.498 ...
```

```
#change to long format
mal_long <- gather(malaria, condition, measurement, drug1:drug10)
head(mal_long)
```

```
##   start starting_allele end_allele changed condition measurement
## 1 first               0       1000      A1     drug1 -0.19960115
## 2 first             100       1100      A2     drug1 -0.06929719
## 3 first              10       1010      A3     drug1  0.05640064
## 4 first               1       1001      A4     drug1 -0.06489213
## 5 first             110       1110      A5     drug1  0.03788536
## 6 first             101       1101      A6     drug1 -0.06609239
```

```
mal_long$drug<-c(rep(1,32),rep(2,32),rep(3,32),rep(4,32),rep(5,32),rep(6,32),rep(7,32),rep(8,32),rep(9,3
```

Regression tree analysis for pyr

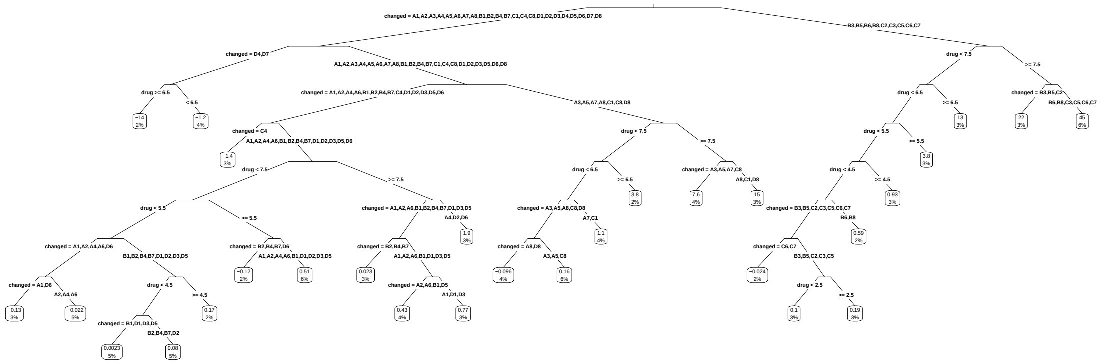
```
#construct formula
form<-as.formula(measurement~changed+drug)
#construct regression tree
tree.1<-rpart(form,data=mal_long,control=rpart.control(minsplit=20,cp=0),method="anova")
printcp(tree.1)
```

```
##
## Regression tree:
## rpart(formula = form, data = mal_long, method = "anova", control = rpart.control(minsplit = 20,
##     cp = 0))
##
## Variables actually used in tree construction:
## [1] changed drug
##
## Root node error: 43344/320 = 135.45
##
## n= 320
##
##            CP nsplit rel error  xerror     xstd
## 1  3.6980e-01      0  1.000000 1.00576 0.186654
## 2  7.0653e-02      2  0.260396 0.33473 0.050771
## 3  2.8034e-02      3  0.189744 0.24549 0.037979
## 4  2.6214e-02      4  0.161709 0.18982 0.027416
## 5  2.3006e-02      5  0.135495 0.17540 0.025184
## 6  1.9273e-02      7  0.089483 0.16930 0.024752
## 7  5.6642e-03      8  0.070210 0.12787 0.019426
## 8  2.1061e-03      9  0.064546 0.12852 0.019436
## 9  1.6657e-03     10  0.062440 0.11945 0.018602
## 10 6.1688e-04     11  0.060774 0.11802 0.018608
## 11 3.1172e-04     12  0.060157 0.11647 0.018584
## 12 2.1415e-04     14  0.059534 0.11634 0.018587
## 13 8.6095e-05     15  0.059320 0.11622 0.018587
## 14 4.5073e-05     16  0.059234 0.11623 0.018590
## 15 4.4113e-05     17  0.059189 0.11622 0.018590
## 16 3.4306e-05     19  0.059100 0.11618 0.018590
## 17 1.3649e-05     20  0.059066 0.11610 0.018592
## 18 1.0991e-05     21  0.059052 0.11603 0.018592
## 19 6.1184e-06     22  0.059041 0.11604 0.018592
## 20 3.8852e-06     23  0.059035 0.11603 0.018592
## 21 2.6356e-06     24  0.059031 0.11603 0.018592
## 22 1.5120e-06     25  0.059029 0.11602 0.018592
## 23 1.1214e-06     26  0.059027 0.11602 0.018592
## 24 8.5731e-07     27  0.059026 0.11602 0.018592
## 25 0.0000e+00     28  0.059025 0.11602 0.018592
```

```
#prune tree
tree.1$cptable[which.min(tree.1$cptable[,"xerror"]),"CP"]
```

```
## [1] 0
```

```
new.tree<-prune(tree.1, cp=0)
#plot tree
rpart.plot(new.tree,type=3,extra=100)
```

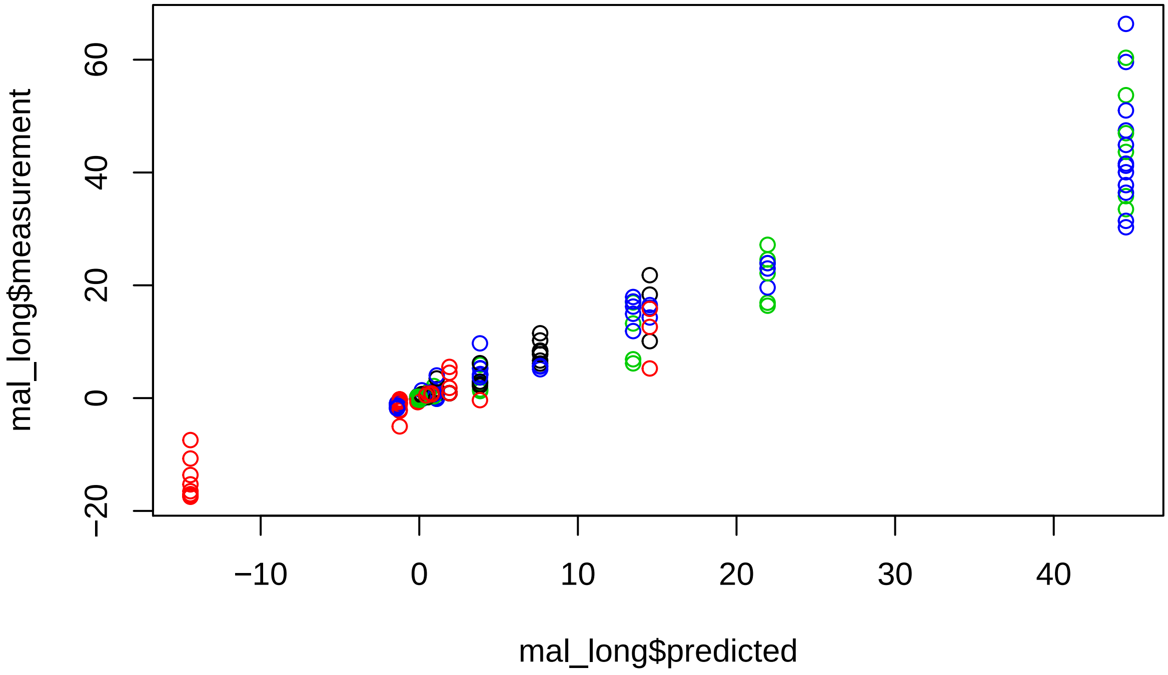


```
#checking variation explained
mal_long$predicted<-predict(new.tree)
#linear model
mod1<-lm(measurement~predicted,data=mal_long)
summary(mod1)
```

```
##
## Call:
## lm(formula = measurement ~ predicted, data = mal_long)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.2639  -0.2048  -0.0101   0.1661  21.7932
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.589e-15  1.690e-01     0.0        1
## predicted   1.000e+00  1.404e-02    71.2   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 2.836 on 318 degrees of freedom
## Multiple R-squared:  0.941,  Adjusted R-squared:  0.9408
## F-statistic:  5070 on 1 and 318 DF,  p-value: < 2.2e-16
```

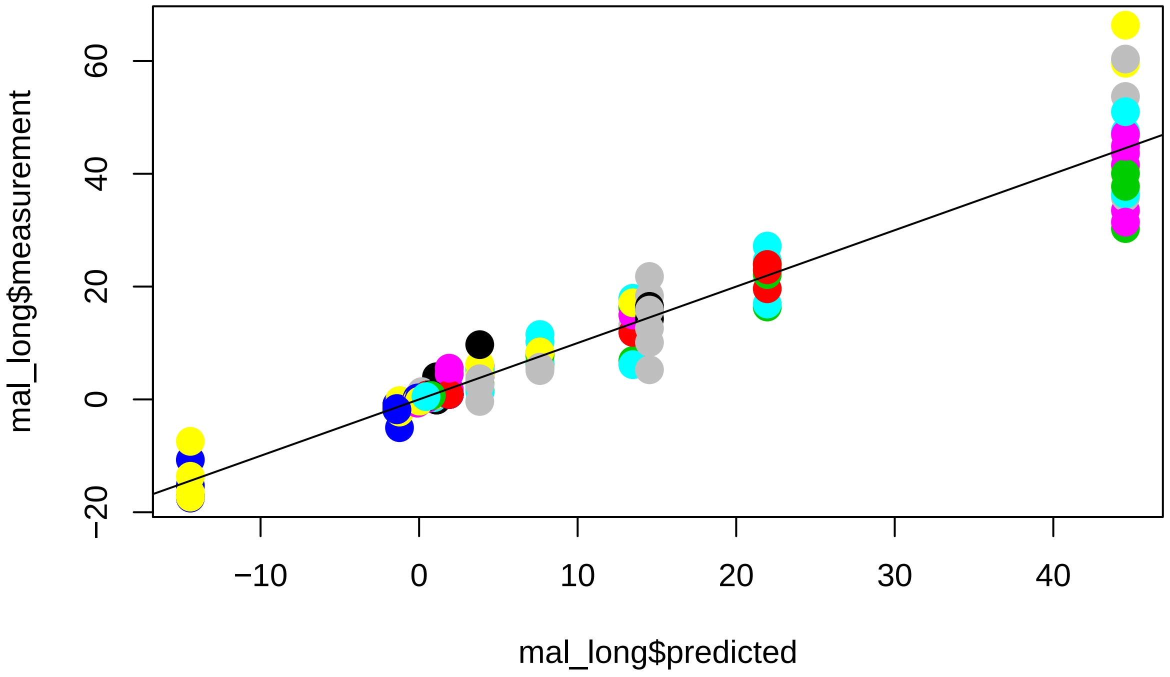
```
#color by start
plot(mal_long$predicted,mal_long$measurement,col=as.factor(mal_long$start))
```



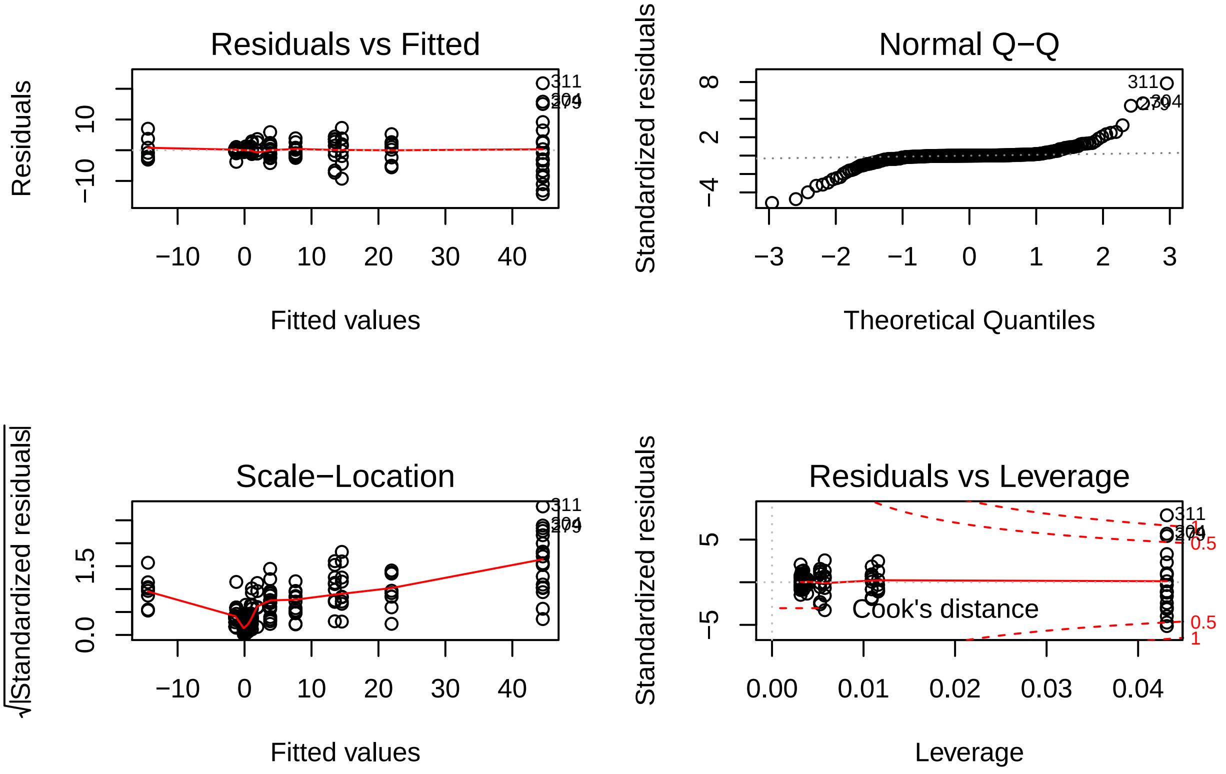
```
#color by change class
#point size correspond to drug
mal_long$sizing<-c(rep(.5,32),rep(2.75,32),rep(1,32),rep(1.25,32),rep(1.5,32),rep(2,32),rep(2.5,32),rep

plot(mal_long$predicted,mal_long$measurement,col=as.factor(mal_long$changed),pch=16,cex=2)
#,cex=mal_long$sizing

#,xlim=c(-2,2),ylim=c(-1.5,1.5)
abline(mod1)
```



```r
#checking the residuals
par(mfrow=c(2,2))
plot(mod1)
```



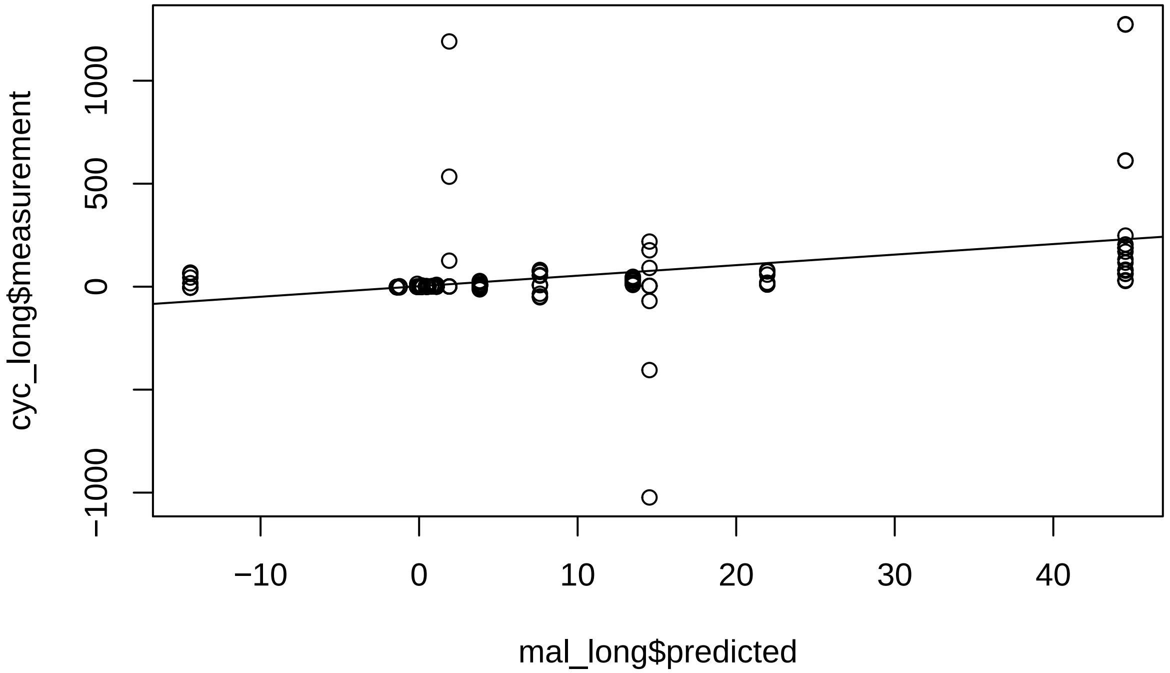
```r
par(mfrow=c(1,1))
```

compare regression tree with CYC growth data

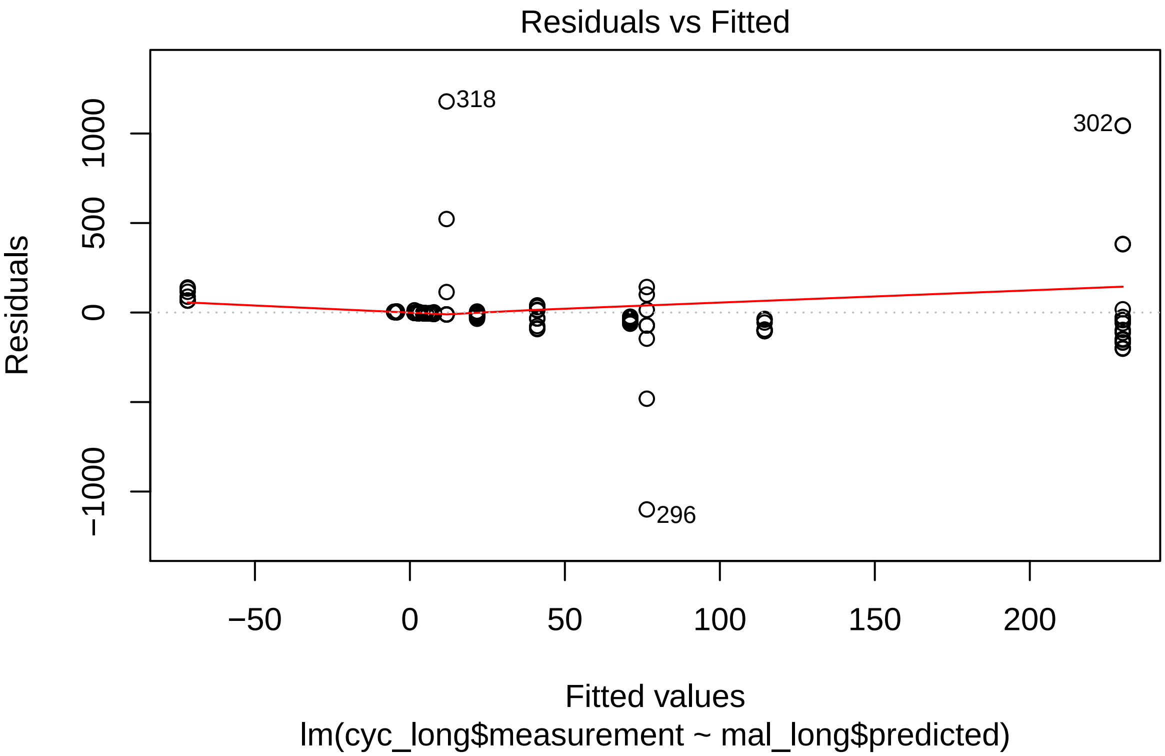
```
cyc<-read.csv("20160227_cyc_drug_dose_aa_shifts.csv")
cyc_long <- gather(cyc, condition, measurement, drug1:drug10)
mod2<-lm(cyc_long$measurement~mal_long$predicted)
summary(mod2)
```

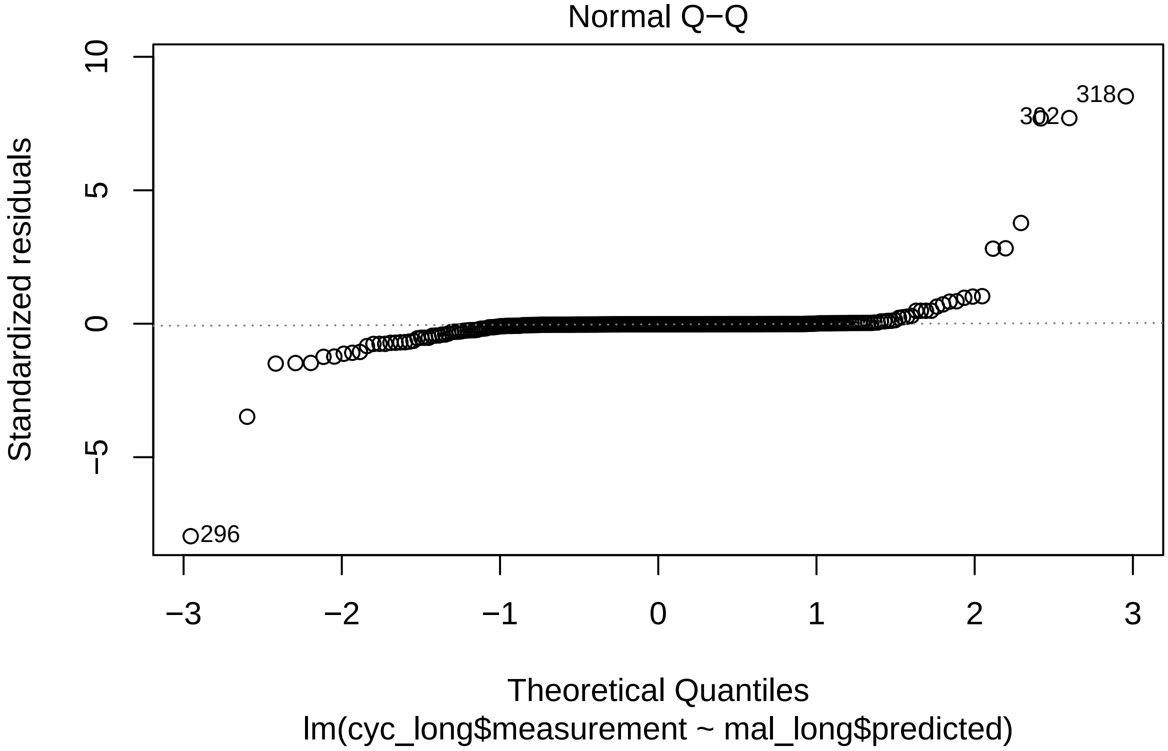
```
##
## Call:
## lm(formula = cyc_long$measurement ~ mal_long$predicted)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1099.82    -4.99    -2.65    -1.83  1178.83
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)          2.1031     8.2544   0.255    0.799
## mal_long$predicted   5.1151     0.6861   7.455  8.6e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 138.6 on 318 degrees of freedom
## Multiple R-squared:  0.1488, Adjusted R-squared:  0.1461
## F-statistic: 55.58 on 1 and 318 DF,  p-value: 8.6e-13
```

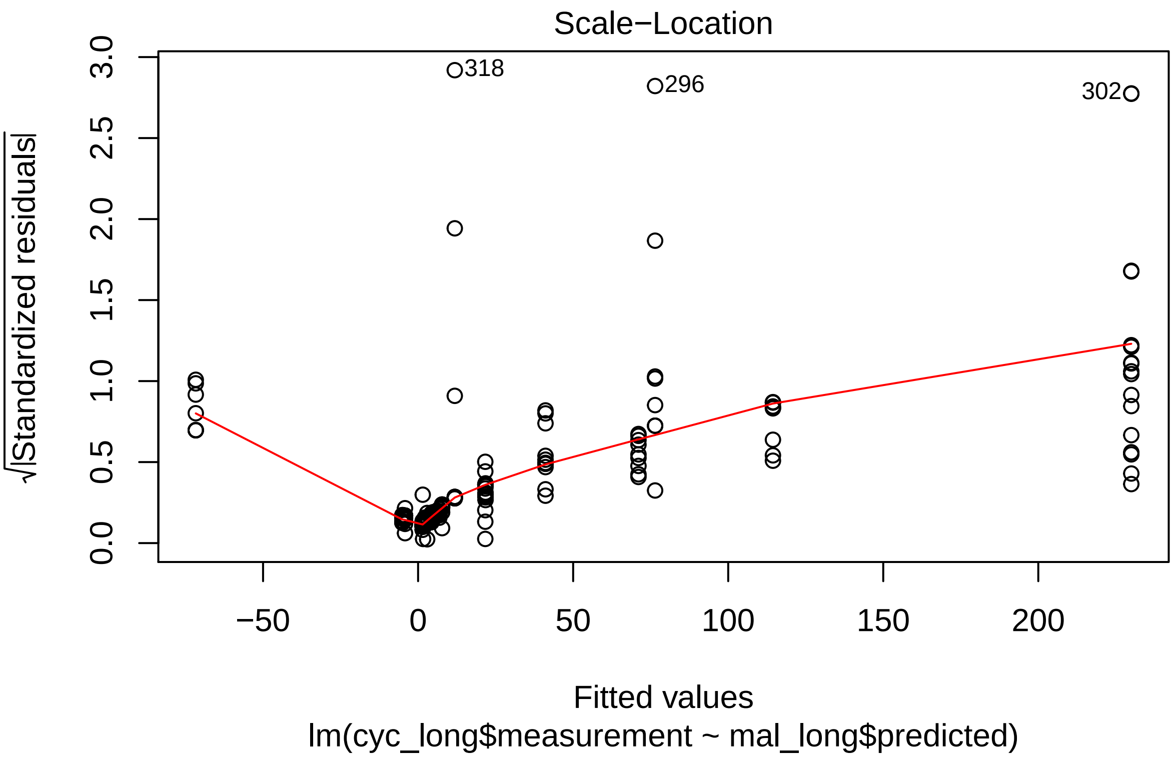
```
plot(mal_long$predicted,cyc_long$measurement)
abline(mod2)
```

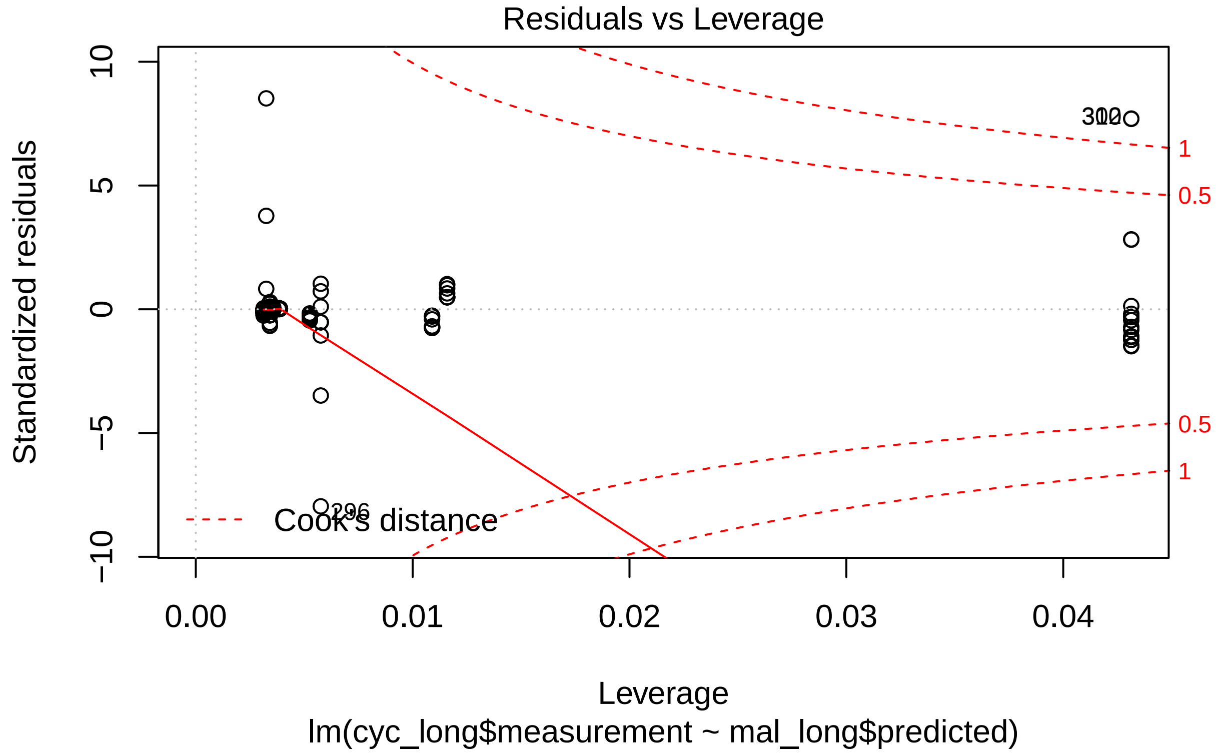


```
plot(mod2)
```


Residuals vs Fitted
Residuals
1000
500
0
−1000
−50
0
50
100
150
200
Fitted values
lm(cyc_long$measurement ~ mal_long$predicted)
318
302
296


Normal Q−Q
Standardized residuals
10
5
0
−5
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(cyc_long$measurement ~ mal_long$predicted)
318
302
296



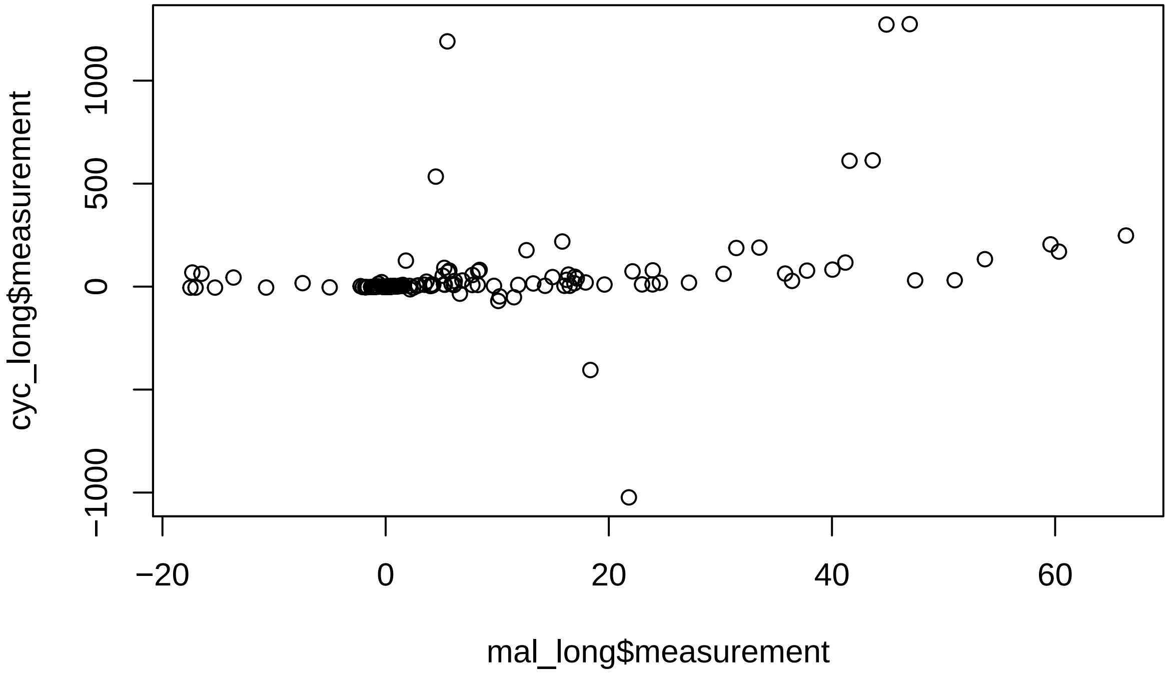


```
#comparing direct measurements of pyr with cyc
mod3<-lm(cyc_long$measurement~mal_long$measurement)
summary(mod3)
```

```
##
## Call:
```

```
## lm(formula = cyc_long$measurement ~ mal_long$measurement)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1132.46    -5.96    -3.61    -2.81  1160.61
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)             3.171      8.255   0.384    0.701
## mal_long$measurement    4.858      0.668   7.273 2.76e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 139.1 on 318 degrees of freedom
## Multiple R-squared:  0.1426, Adjusted R-squared:  0.1399
## F-statistic:  52.9 on 1 and 318 DF,  p-value: 2.757e-12
```

```
plot(mal_long$measurement,cyc_long$measurement)
```


cyc_long$measurement
mal_long$measurement