Which mutation transitions and drug concentrations influence growth?

Andrew Nguyen

2016-February-27

Contents

regression trees

load libraies and read in data

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


mal_long<-read.csv("../Data/20160227_final_dataset_malaria_pyr_cyc_abs_relative.csv")
```

Regression tree analysis for pyr

pyr delta fitness

```
#construct formula
form<-as.formula(pyr_abs~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: 42.507/320 = 0.13283
##
## n= 320
##
##             CP nsplit rel error  xerror     xstd
## 1  2.9374e-01      0   1.00000 1.01077 0.146218
## 2  1.9681e-01      1   0.70626 0.87232 0.114643
## 3  8.4602e-02      2   0.50945 0.56123 0.056506
## 4  6.9584e-02      3   0.42485 0.55913 0.059138
## 5  5.3793e-02      5   0.28568 0.51592 0.056896
## 6  2.1167e-02      6   0.23188 0.48733 0.057522
## 7  1.3128e-02      7   0.21072 0.42109 0.053948
## 8  7.8837e-03      9   0.18446 0.37832 0.051101
## 9  7.6911e-03     10   0.17658 0.36998 0.050852
## 10 5.5580e-03     11   0.16889 0.37123 0.050942
## 11 3.4965e-03     12   0.16333 0.36032 0.050774
## 12 2.8958e-03     13   0.15983 0.35680 0.050745
## 13 1.1605e-03     15   0.15404 0.35227 0.050757
## 14 8.3833e-04     16   0.15288 0.34977 0.050791
## 15 6.8745e-04     17   0.15204 0.34844 0.050820
## 16 3.6886e-04     18   0.15135 0.34763 0.050833
## 17 2.9953e-04     19   0.15098 0.34726 0.050840
## 18 2.1550e-04     20   0.15068 0.34715 0.050842
## 19 1.9390e-04     21   0.15047 0.34695 0.050844
## 20 7.7118e-05     22   0.15028 0.34684 0.050846
## 21 2.7537e-05     23   0.15020 0.34688 0.050845
## 22 5.0041e-06     24   0.15017 0.34687 0.050845
## 23 0.0000e+00     25   0.15017 0.34687 0.050845
```

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

```
## [1] 7.711776e-05
```

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

```
## Warning: labs do not fit even at cex 0.15, there may be some overplotting
```
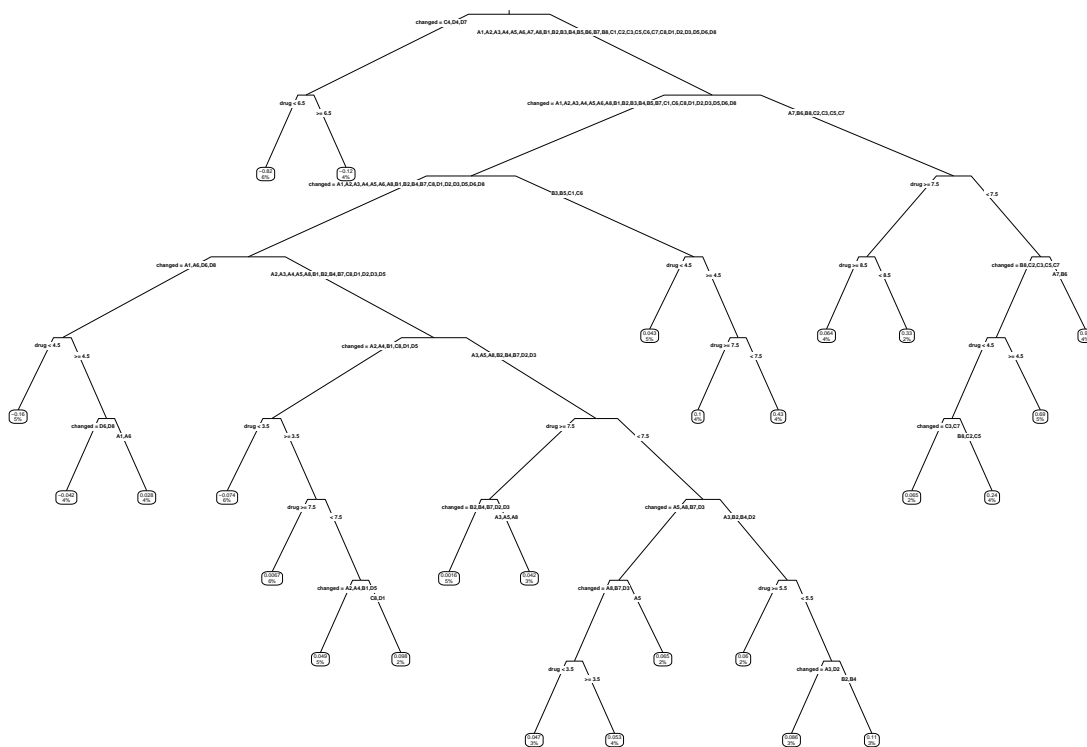
```r
#checking variation explained
mal_long$predicted<-predict(new.tree)
#linear model
mod1<-lm(pyr_abs~predicted,data=mal_long)
summary(mod1)
```

```
##
## Call:
## lm(formula = pyr_abs ~ predicted, data = mal_long)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.68744 -0.03795 -0.00444  0.05148  0.52623
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3.103e-16  8.105e-03    0.00        1
## predicted   1.000e+00  2.357e-02   42.42   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1417 on 318 degrees of freedom
## Multiple R-squared:  0.8498, Adjusted R-squared:  0.8494
## F-statistic:  1800 on 1 and 318 DF,  p-value: < 2.2e-16
```

```r
#color by start
plot(mal_long$predicted,mal_long$pyr_abs,col=as.factor(mal_long$start))
```
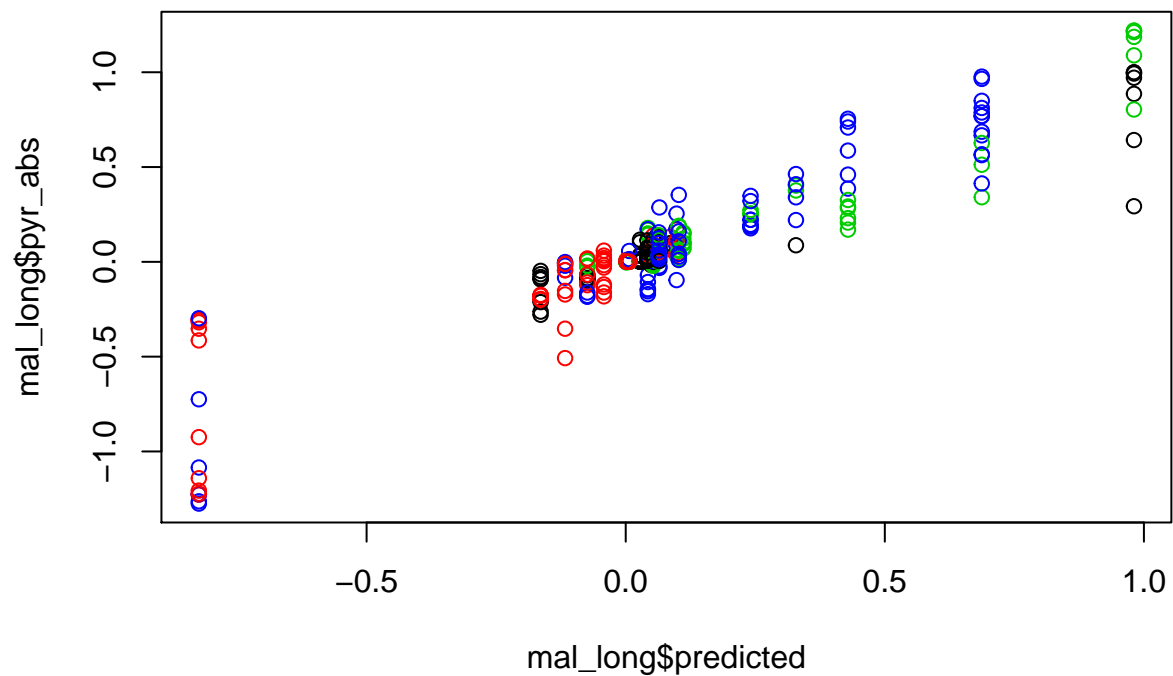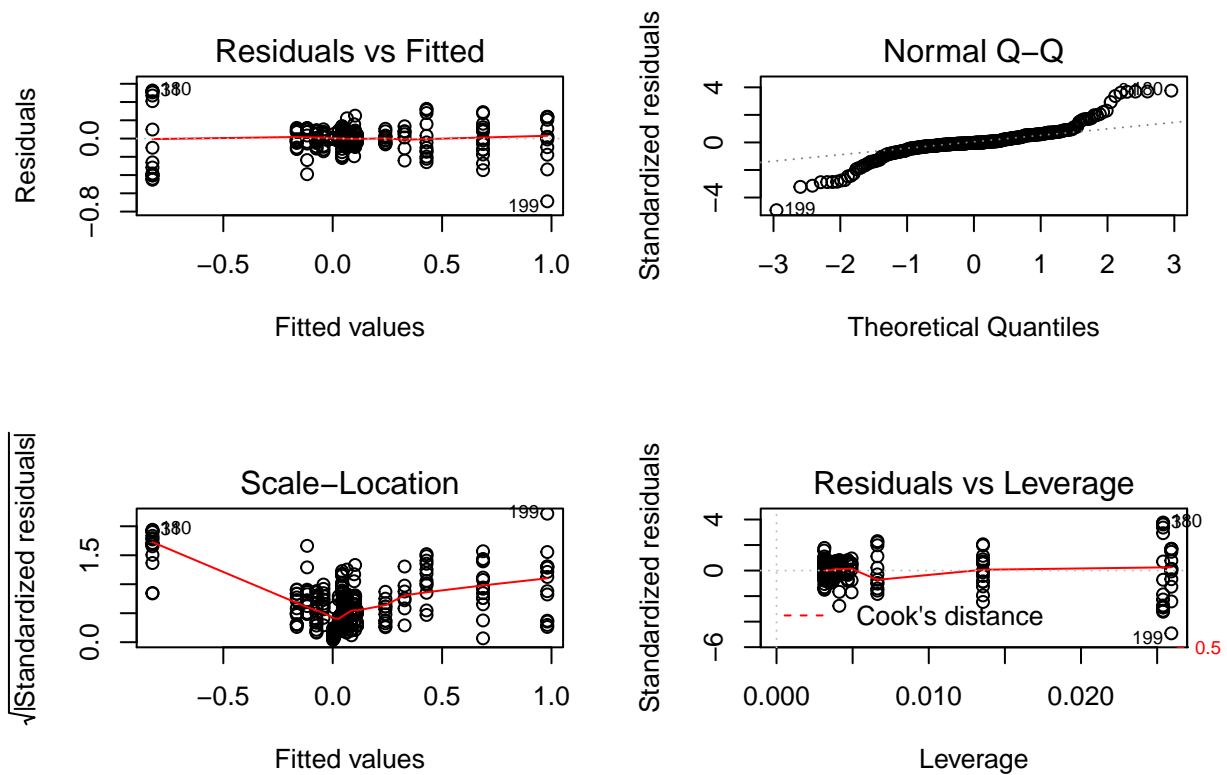
```
#color by change class
#point size correspond to drug
mal_long$sizing<-c(rep(.5,32),rep(2.75,32),rep(1,32),rep(1.25,32),rep(1.5,32),rep(2,32),rep(2.5,32),rep

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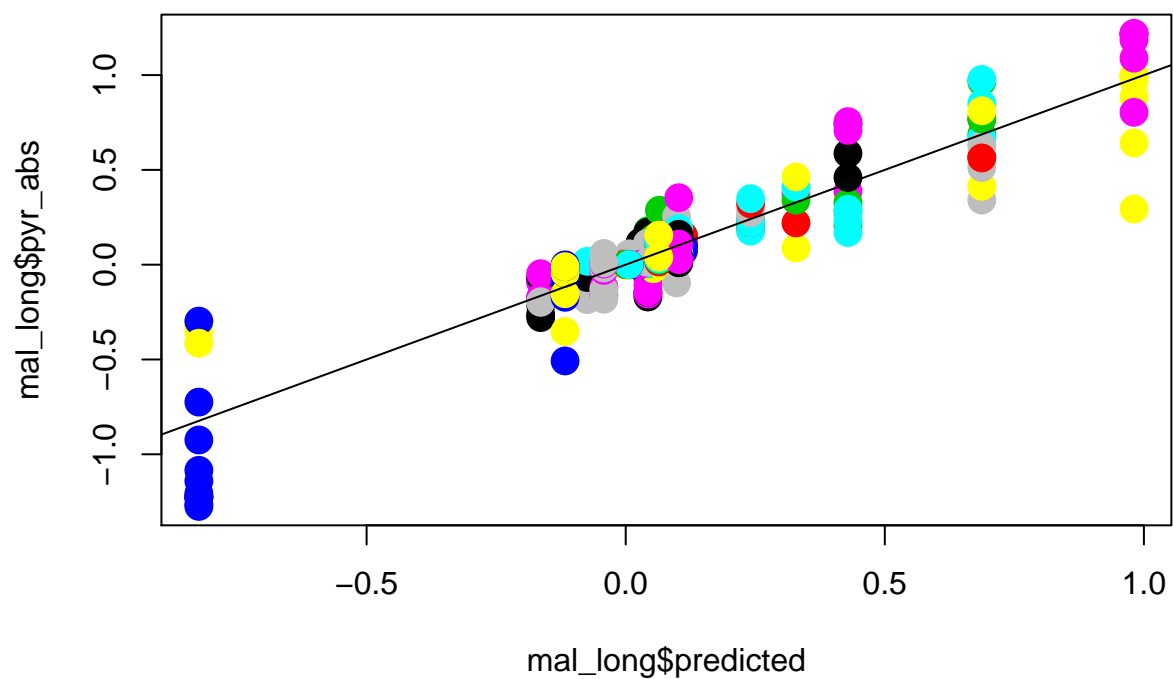
```r
#checking the residuals
par(mfrow=c(2,2))
plot(mod1)
```



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

pyr relative fitness

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

```
##
## 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 1.01410 0.188212
## 2  7.0653e-02      2  0.260396 0.29694 0.045982
## 3  2.8034e-02      3  0.189744 0.27946 0.041213
## 4  2.6214e-02      4  0.161709 0.21938 0.031603
## 5  2.3006e-02      5  0.135495 0.19544 0.028702
## 6  1.9273e-02      7  0.089483 0.16484 0.024204
## 7  5.6642e-03      8  0.070210 0.14044 0.021961
## 8  2.1061e-03      9  0.064546 0.13436 0.021377
## 9  1.6657e-03     10  0.062440 0.13061 0.021211
## 10 6.1688e-04     11  0.060774 0.12503 0.020591
## 11 3.1172e-04     12  0.060157 0.12365 0.020596
## 12 2.1415e-04     14  0.059534 0.12355 0.020598
## 13 8.6095e-05     15  0.059320 0.12300 0.020604
## 14 4.5073e-05     16  0.059234 0.12278 0.020607
## 15 4.4113e-05     17  0.059189 0.12271 0.020608
## 16 3.4306e-05     19  0.059100 0.12270 0.020608
## 17 1.3649e-05     20  0.059066 0.12256 0.020610
## 18 1.0991e-05     21  0.059052 0.12254 0.020610
## 19 6.1184e-06     22  0.059041 0.12254 0.020610
## 20 3.8852e-06     23  0.059035 0.12253 0.020610
## 21 2.6356e-06     24  0.059031 0.12254 0.020610
## 22 1.5120e-06     25  0.059029 0.12253 0.020610
## 23 1.1214e-06     26  0.059027 0.12253 0.020610
## 24 8.5731e-07     27  0.059026 0.12253 0.020610
## 25 0.0000e+00     28  0.059025 0.12253 0.020610
```
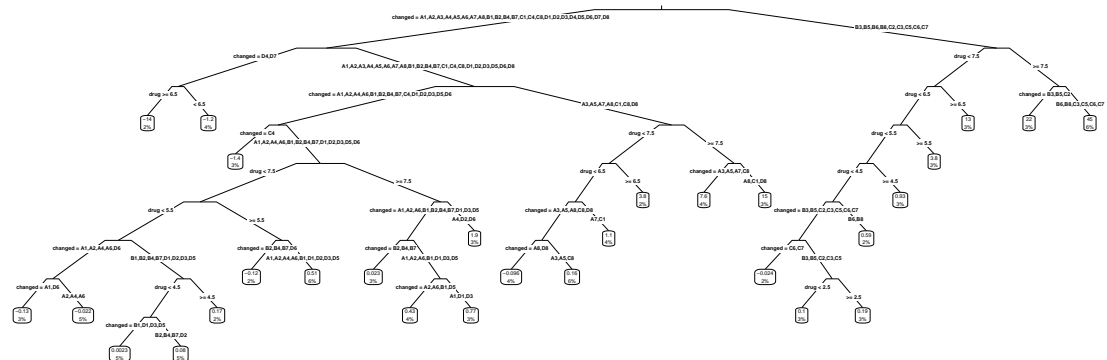
```r
#prune tree
pyr_rel_meas$cptable[which.min(pyr_rel_meas$cptable[,"xerror"]),"CP"]
```

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

```r
pyr_rel_meas2<-prune(pyr_rel_meas, cp=0)
#plot tree
rpart.plot(pyr_rel_meas2,type=3,extra=100)
```



```r
#checking variation explained
mal_long$pyr_rel_meas_predicted<-predict(new.tree)
#linear model
mod2<-lm(pyr_abs~pyr_rel_meas_predicted,data=mal_long)
summary(mod2)
```

```
## 
## Call:
## lm(formula = pyr_abs ~ pyr_rel_meas_predicted, data = mal_long)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.68744 -0.03795 -0.00444  0.05148  0.52623
## 
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)            3.103e-16  8.105e-03    0.00        1
## pyr_rel_meas_predicted 1.000e+00  2.357e-02   42.42   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.1417 on 318 degrees of freedom
## Multiple R-squared:  0.8498, Adjusted R-squared:  0.8494
## F-statistic:  1800 on 1 and 318 DF,  p-value: < 2.2e-16
```

```r
#color by start
plot(mal_long$pyr_rel_meas_predicted,mal_long$pyr_abs,col=as.factor(mal_long$start))
abline(mod2)
```
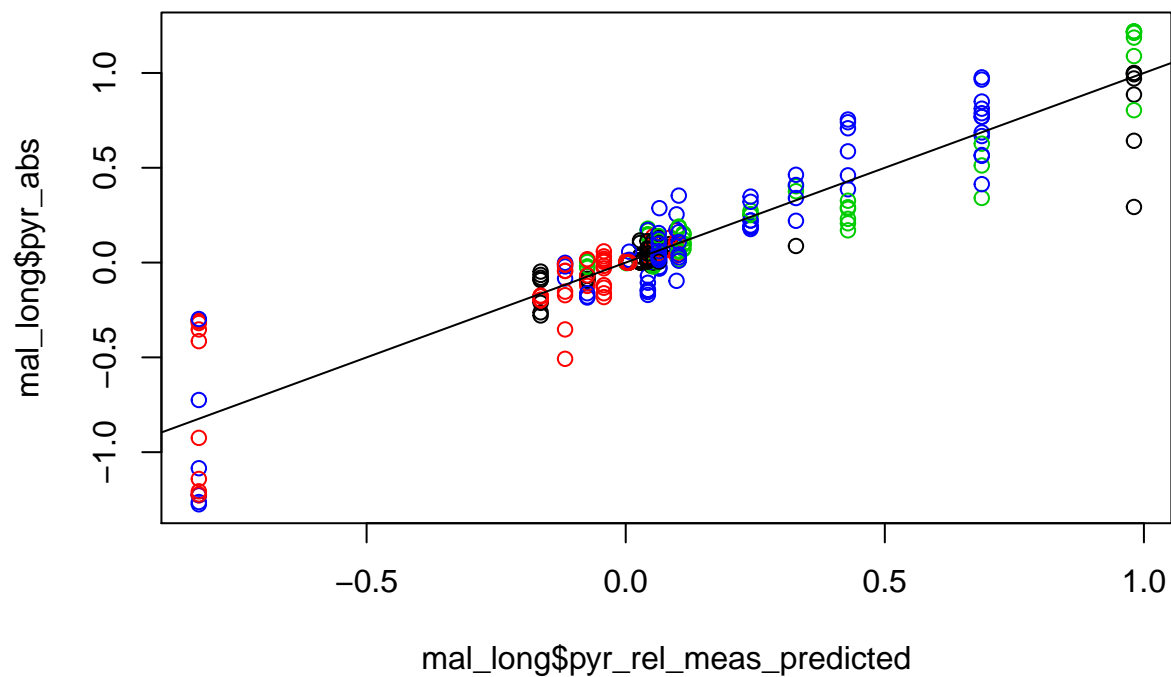


```r
#checking the residuals
par(mfrow=c(2,2))
plot(mod2)
```
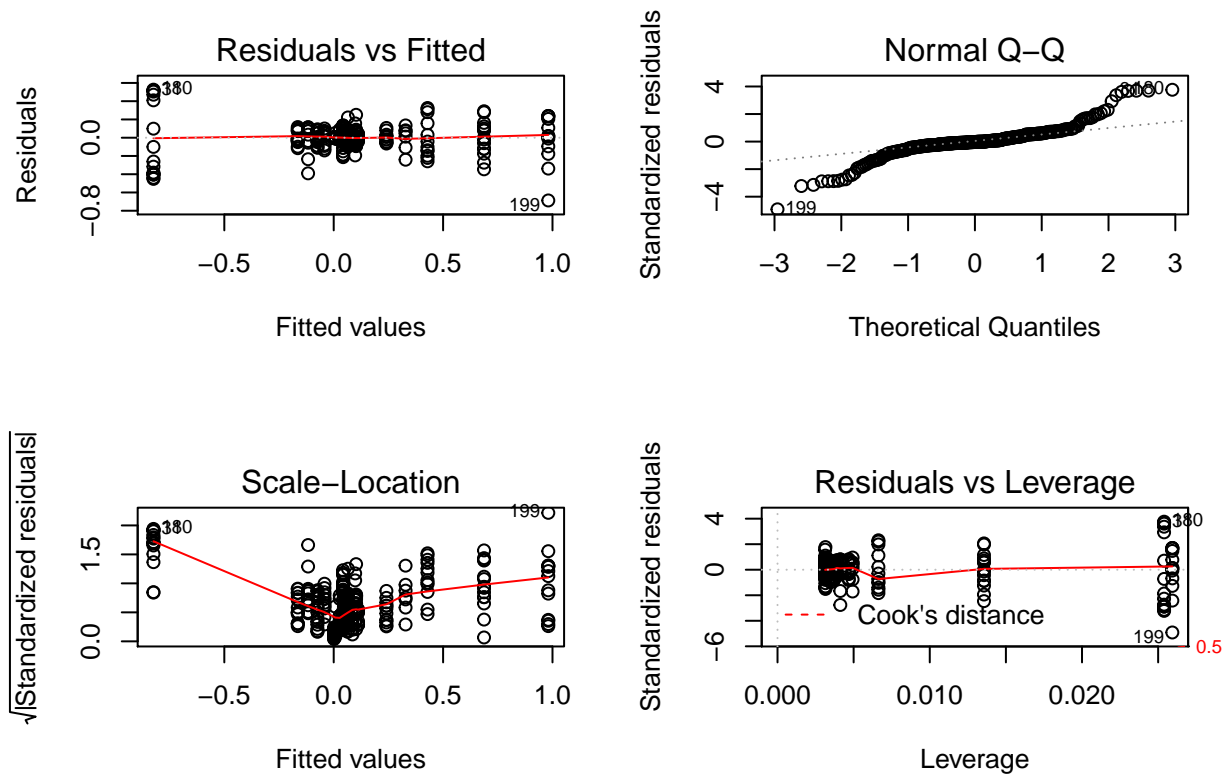
```r
par(mfrow=c(1,1))
```

Regression tree analysis for cyc

cyc delta fitness

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

```
##
## 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: 50.434/320 = 0.15761
##
## n= 320
##
##           CP nsplit rel error  xerror     xstd
## 1 0.31963248      0   1.00000 1.00891 0.127066
## 2 0.15097929      1   0.68037 0.71417 0.087692
```

```
## 3  0.10172411      2   0.52939 0.61830 0.059530
## 4  0.06755882      3   0.42766 0.58707 0.050474
## 5  0.04475129      4   0.36011 0.54863 0.050777
## 6  0.03477959      5   0.31535 0.51440 0.049556
## 7  0.02558626      6   0.28057 0.49060 0.049653
## 8  0.02482108      7   0.25499 0.47981 0.049109
## 9  0.00938617      8   0.23017 0.46503 0.048444
## 10 0.00613666     10   0.21139 0.43871 0.048551
## 11 0.00498082     11   0.20526 0.43615 0.048351
## 12 0.00481884     12   0.20028 0.43517 0.048352
## 13 0.00382639     13   0.19546 0.43037 0.048361
## 14 0.00377140     15   0.18781 0.43006 0.048335
## 15 0.00210706     16   0.18403 0.42636 0.048065
## 16 0.00164684     18   0.17982 0.42371 0.048062
## 17 0.00103491     20   0.17653 0.42335 0.048106
## 18 0.00063426     21   0.17549 0.42304 0.048116
## 19 0.00036380     23   0.17422 0.42012 0.048178
## 20 0.00021676     24   0.17386 0.41966 0.048176
## 21 0.00000000     25   0.17364 0.41978 0.048175
```

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

```
## [1] 0.0002167576
```

```r
cyc_relative_tree2<-prune(cyc_relative_tree, cp=00498)
#plot tree
rpart.plot(cyc_relative_tree,type=3,extra=100)
```
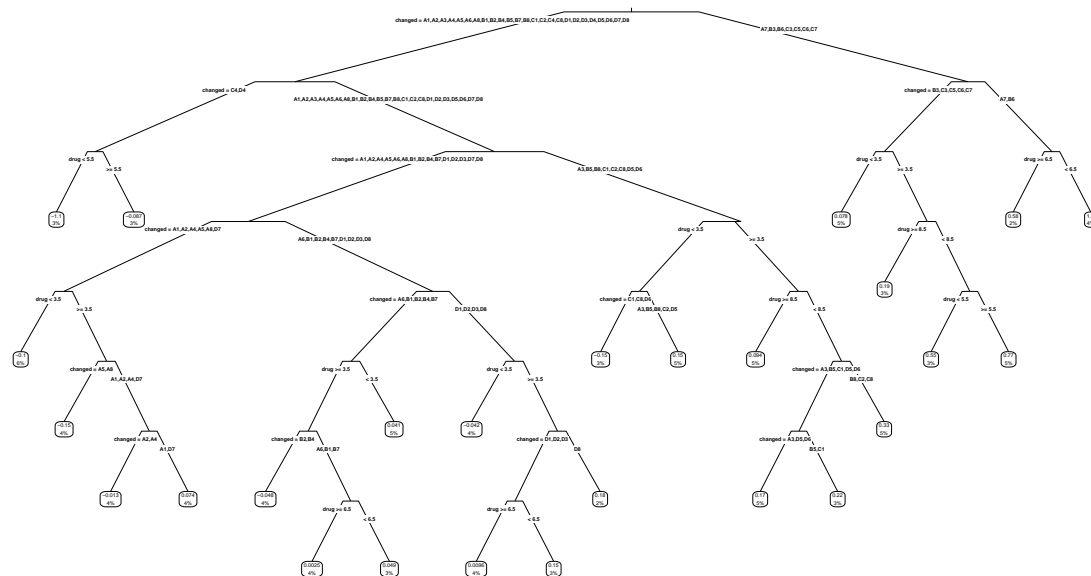


```r
mal_long$cyc_relative_prediction<-predict(cyc_relative_tree)
mod6<-lm(cyc_relative~cyc_relative_prediction,data=mal_long)
summary(mod6)
```

```
##
```

```
## Call:
## lm(formula = cyc_relative ~ cyc_relative_prediction, data = mal_long)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.61332 -0.08801 -0.00158  0.08314  0.63363
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             2.483e-17  9.756e-03     0.0        1
## cyc_relative_prediction 1.000e+00  2.571e-02    38.9   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1659 on 318 degrees of freedom
## Multiple R-squared:  0.8264, Adjusted R-squared:  0.8258
## F-statistic:  1513 on 1 and 318 DF,  p-value: < 2.2e-16
```

```
plot(mal_long$cyc_relative,mal_long$cyc_relative_prediction)
abline(mod6)
```
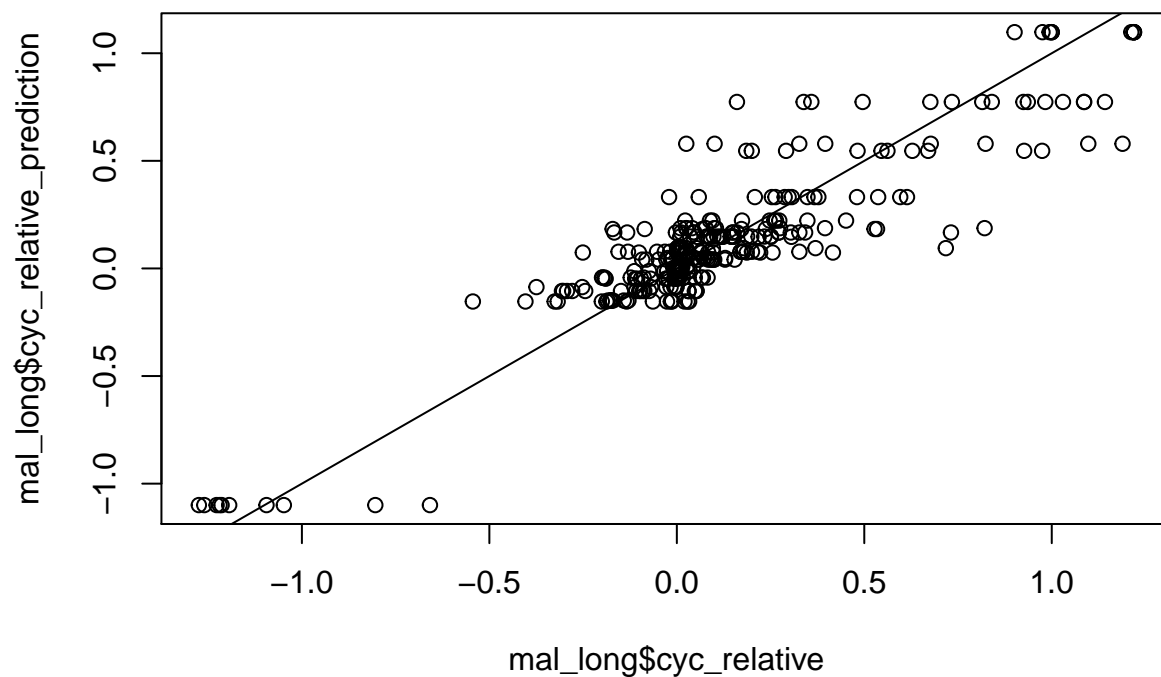

cyc relative fitness

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

```
##
```

```
## 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: 7173248/320 = 22416
##
## n= 320
##
##            CP nsplit rel error  xerror    xstd
## 1  2.6701e-01      0   1.00000 1.00501 0.39953
## 2  3.4668e-02      2   0.46599 0.86187 0.34233
## 3  1.5447e-02      3   0.43132 0.72226 0.25766
## 4  6.3356e-03      5   0.40042 0.69290 0.25605
## 5  5.7819e-04      6   0.39409 0.67085 0.24741
## 6  3.5239e-04      7   0.39351 0.66941 0.24742
## 7  2.4010e-04      8   0.39316 0.66939 0.24742
## 8  1.5515e-04     10   0.39268 0.66908 0.24742
## 9  3.2616e-05     11   0.39252 0.66857 0.24743
## 10 2.2929e-05     12   0.39249 0.66846 0.24743
## 11 1.4699e-05     13   0.39247 0.66845 0.24743
## 12 4.1893e-06     15   0.39244 0.66843 0.24743
## 13 3.2353e-06     16   0.39243 0.66842 0.24743
## 14 2.5064e-06     17   0.39243 0.66842 0.24743
## 15 1.1500e-06     18   0.39243 0.66842 0.24743
## 16 1.0166e-06     19   0.39243 0.66842 0.24743
## 17 5.3571e-07     20   0.39242 0.66842 0.24743
## 18 4.2244e-07     21   0.39242 0.66842 0.24743
## 19 3.1688e-07     22   0.39242 0.66842 0.24743
## 20 2.6210e-07     23   0.39242 0.66842 0.24743
## 21 1.8543e-07     24   0.39242 0.66842 0.24743
## 22 2.7221e-08     25   0.39242 0.66841 0.24743
## 23 2.1421e-08     26   0.39242 0.66841 0.24743
## 24 1.0355e-08     27   0.39242 0.66841 0.24743
## 25 0.0000e+00     28   0.39242 0.66841 0.24743
```
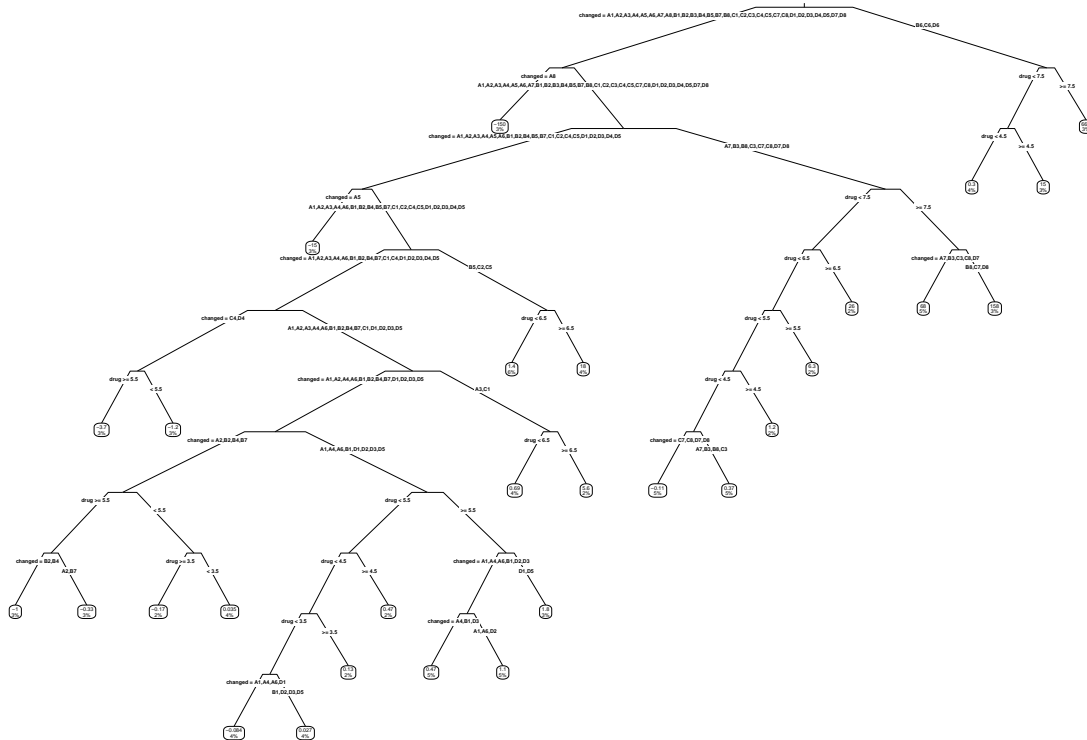
```r
#prune tree
cyc_abs_tree$cptable[which.min(cyc_abs_tree$cptable[,"xerror"]),"CP"]
```

```
## [1] 2.722086e-08
```

```r
cyc_abs_tree2<-prune(cyc_abs_tree, cp=0)
#plot tree
rpart.plot(cyc_abs_tree2,type=3,extra=100)
```

```
## Warning: labs do not fit even at cex 0.15, there may be some overplotting
```

```r
mal_long$cyc_abs_prediction<-predict(cyc_abs_tree2)
mod5<-lm(cyc_abs~cyc_abs_prediction,data=mal_long)
summary(mod5)
```

```
##
## Call:
## lm(formula = cyc_abs ~ cyc_abs_prediction, data = mal_long)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -873.05   -0.39    0.00    0.32  607.89
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        3.178e-15  5.364e+00    0.00        1
## cyc_abs_prediction 1.000e+00  4.507e-02   22.19   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 94.09 on 318 degrees of freedom
## Multiple R-squared:  0.6076, Adjusted R-squared:  0.6063
## F-statistic: 492.4 on 1 and 318 DF,  p-value: < 2.2e-16
```

```r
plot(mal_long$cyc_abs,mal_long$cyc_abs_prediction)
abline(mod5)
```
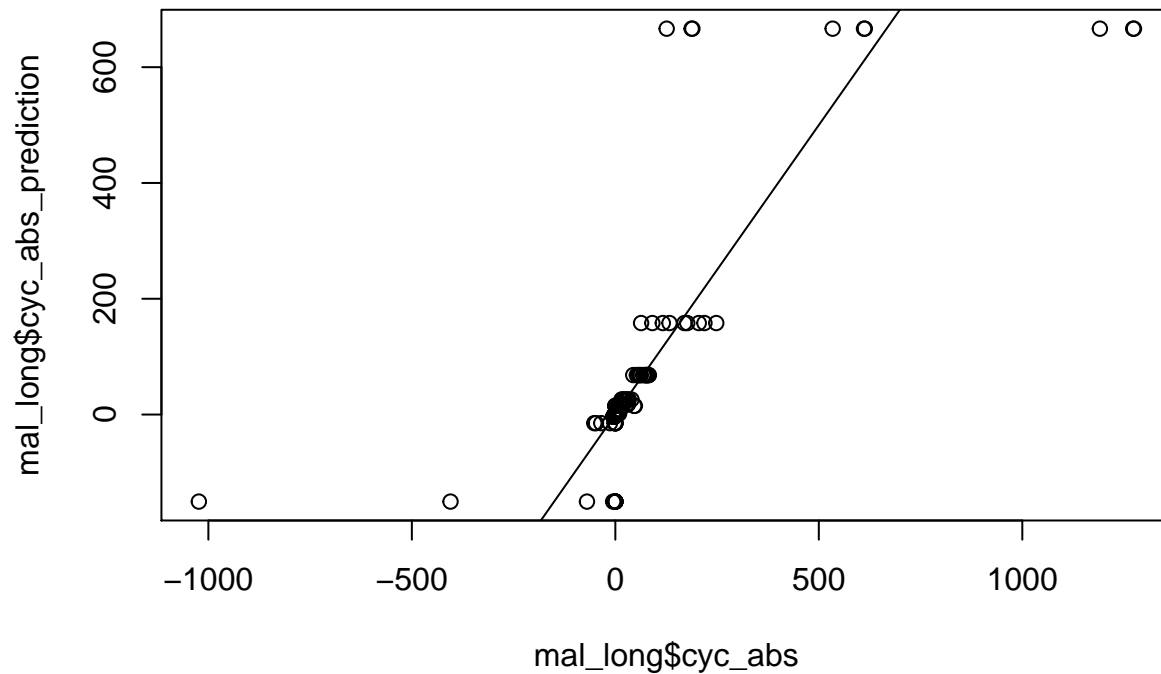
SessionInfo

```
sessionInfo()
```

```
## R version 3.2.2 (2015-08-14)
## Platform: x86_64-apple-darwin13.4.0 (64-bit)
## Running under: OS X 10.11.3 (El Capitan)
##
## locale:
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
## [1] rpart.plot_1.5.2 rpart_4.1-10     tidyr_0.2.0      ggplot2_1.0.1
##
## loaded via a namespace (and not attached):
##  [1] Rcpp_0.11.6      digest_0.6.8     MASS_7.3-43      grid_3.2.2
##  [5] plyr_1.8.2       gtable_0.1.2     formatR_1.2.1    magrittr_1.5
##  [9] scales_0.2.4     evaluate_0.8     stringi_1.0-1    reshape2_1.4.1
## [13] rmarkdown_0.9.3  proto_0.3-10     tools_3.2.2      stringr_1.0.0
## [17] munsell_0.4.2    yaml_2.1.13      colorspace_1.2-6 htmltools_0.3
## [21] knitr_1.10.5
```