R implementation

Sleiman Bassim, PhD

April 2, 2015

Loaded functions:

```
#source("/media/Data/Dropbox/humanR/01funcs.R")
rm(list=ls())
#setwd("/media/Data/Dropbox/humanR/PD/")
#setwd("~/Dropbox/humanR/PD/")
###load("PD.Rdata", .GlobalEnv)
#lsos(pat="")
```

1 Data preprocessing

Load packages.

```
pkgs <- c('xlsx','caret','leaps','glmnet','lattice','latticeExtra','pvclust','gplots')
lapply(pkgs, require, character.only = TRUE)
```

```
[[1]]
[1] TRUE

[[2]]
[1] TRUE

[[3]]
[1] TRUE

[[4]]
[1] TRUE

[[5]]
[1] TRUE

[[6]]
[1] TRUE

[[7]]
[1] TRUE

[[8]]
[1] TRUE
```

2 Load data

Load data.

```
sugars <- read.xlsx("./algae.xlsx", header =TRUE, sheetName = "Cluster")
experiments <- read.xlsx("./algae.xlsx", header = TRUE, sheetName = "Logit")
```

Head of sugars data.

```
head(sugars)
```

```
             sample   PHA  ECA.   SBA  HPA.   PWM ConA  PEA   PNA
1   Rhodomonas lens 0.500 0.851 1.306 0.733 0.586 2.76 1.52 0.763
2   Rhodomonas lens 0.217 0.848 1.713 0.513 0.314 2.45 1.48 0.805
3   Rhodomonas lens 0.628 0.682 0.942 1.143 0.327 2.58 1.57 0.772
4 Rhodomonas salina 1.366 1.112 1.266 0.942 0.383 2.37 1.12 0.592
5 Rhodomonas salina 1.840 1.789 1.533 0.856 0.395 2.35 1.23 0.682
6 Rhodomonas salina 0.789 1.093 1.101 0.741 0.390 2.56 1.33 0.453
    WGA   UEA NA. NA..1 NA..2 NA..3 NA..4 NA..5 NA..6 NA..7 NA..8
1 1.408 1.085  NA    NA    NA    NA    NA    NA    NA    NA    NA
2 0.922 1.053  NA    NA    NA    NA    NA    NA    NA    NA    NA
3 1.047 1.485  NA    NA    NA    NA    NA    NA    NA    NA    NA
4 0.841 0.757  NA    NA    NA    NA    NA    NA    NA    NA    NA
5 1.078 0.831  NA    NA    NA    NA    NA    NA    NA    NA    NA
6 0.666 0.539  NA    NA    NA    NA    NA    NA    NA    NA    NA
  NA..9 NA..10 NA..11 NA..12 NA..13 NA..14 NA..15
1    NA     NA     NA     NA     NA     NA     NA
2    NA     NA     NA     NA     NA     NA     NA
3    NA     NA     NA     NA     NA     NA     NA
4    NA     NA     NA     NA     NA     NA     NA
5    NA     NA     NA     NA     NA     NA     NA
6    NA     NA     NA     NA     NA     NA     NA
```

```
sugars <- sugars[, 1:11]
```

Head of experiment data.

```
head(experiments)
```

```
  total.mussel select.mussel total.oyster select.oyster experiment
1           12         1.000           11         0.909       exp1
2           12         1.000            7         0.714       exp2
3           11         1.000           10         0.900       exp3
4           12         0.833           10         1.000       exp4
5           12         1.000            8         1.000       exp5
6           12         1.000            8         0.875       exp6
      PHA    ECA    SBA     HPA     PWM    ConA    PEA    PNA     WGA
1  0.0553 -0.131  0.141  0.3641  0.0370 -0.0629 1.8389  1.559 -4.6357
2 -0.0210  0.262  0.255  0.0262 -0.2226  0.3573 1.5889  0.463 -2.9769
3 -1.0246 -0.765 -2.187 -0.3418 -0.4397 -0.8196 3.0216  1.593  0.7316
4  0.2474  0.416 -1.309  0.0607 -0.0804  0.2950 0.3390  0.179  0.0755
5 -1.2949 -0.806 -1.478 -0.5816 -0.6671  0.3441 0.0446 -0.214  5.2516
6  0.4231  0.773  1.295  0.7432  0.1936  1.5784 1.0384  0.812 -8.0048
     UEA
1  0.835
2  0.269
3  0.231
4 -0.223
5 -0.599
6  1.148
```

3 Change rownames

Since every 3 samples have the same algae, im keeping the name but adding a numerical suffix to it.

```
rownames(sugars) <- paste(sugars[, 1], rep(1:3, nrow(sugars)/3), sep = "")
```

```
head(sugars)
```

```
                             sample   PHA  ECA.   SBA  HPA.   PWM
Rhodomonas lens1     Rhodomonas lens 0.500 0.851 1.306 0.733 0.586
Rhodomonas lens2     Rhodomonas lens 0.217 0.848 1.713 0.513 0.314
Rhodomonas lens3     Rhodomonas lens 0.628 0.682 0.942 1.143 0.327
Rhodomonas salina1 Rhodomonas salina 1.366 1.112 1.266 0.942 0.383
Rhodomonas salina2 Rhodomonas salina 1.840 1.789 1.533 0.856 0.395
Rhodomonas salina3 Rhodomonas salina 0.789 1.093 1.101 0.741 0.390
                    ConA  PEA   PNA   WGA   UEA
Rhodomonas lens1   2.76 1.52 0.763 1.408 1.085
Rhodomonas lens2   2.45 1.48 0.805 0.922 1.053
Rhodomonas lens3   2.58 1.57 0.772 1.047 1.485
Rhodomonas salina1 2.37 1.12 0.592 0.841 0.757
Rhodomonas salina2 2.35 1.23 0.682 1.078 0.831
Rhodomonas salina3 2.56 1.33 0.453 0.666 0.539
```

```
sugars <- sugars[, -1]
```

4 Hierarchical clustering

Remove missing data.

```
sugars.full <- na.omit(sugars)
dim(sugars)
```

```
[1] 48 10
```

```
dim(sugars.full)
```

```
[1] 47 10
```

Scale the data.

```
sugars.sc <- scale(sugars.full)
```

Transpose the data.

```
sugars.sc.tp <- t(sugars.sc)
sugars.sc.tp[1:3, 1:3]
```

```
     Rhodomonas lens1 Rhodomonas lens2 Rhodomonas lens3
PHA            0.0345           -0.574           0.3097
ECA.           0.9003            0.893           0.4466
SBA            0.3910            0.903          -0.0668
```

Get some compiled fancy colours.

```
source("http://faculty.ucr.edu/~tgirke/Documents/R_BioCond/My_R_Scripts/my.colorFct.R")
```

Cluster ROWS (sugar classes)

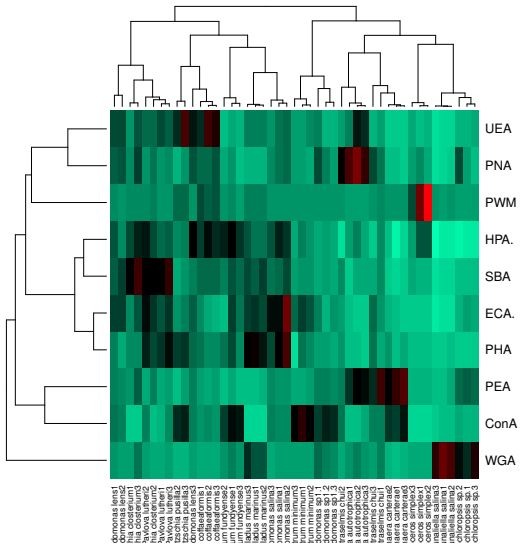
```
sugars.clust <- hclust(as.dist(1-cor((sugars.sc),
                                     method="pearson")),
                       method="complete")
```

Cluster COLUMNS (algae species)

```
algae.clust <- hclust(as.dist(1-cor(sugars.sc.tp,
                                    method="pearson")),
                      method="complete")
```

Draw heatmap.

```
heatmap(sugars.sc.tp,
        Rowv=as.dendrogram(sugars.clust),
        Colv=as.dendrogram(algae.clust),
        col=my.colorFct(), scale="row")
```



5 Cut the tree to extract special patterns of clusterization

Cutting the tree is subjective to ones view of the heatmap.

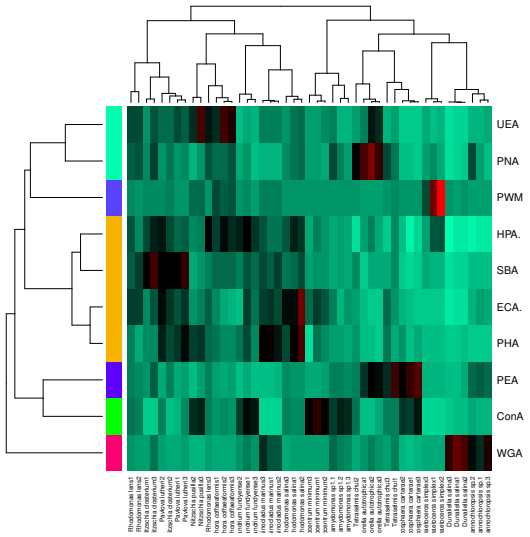
```
sugars.cut <- cutree(sugars.clust, h = max(sugars.clust$height)/2)
```

Prepare some colors.

```
custom.colors <- sample(rainbow(256))
custom.colors <- custom.colors[as.vector(sugars.cut)]
```

Draw another heatmap with identified clusters.

```
heatmap(sugars.sc.tp,
        Rowv=as.dendrogram(sugars.clust),
        Colv=as.dendrogram(algae.clust),
        col=my.colorFct(), scale="row",
        RowSideColors=custom.colors)
```



6 Bootstrapping

Bootstrapping is a resampling technique. One of my favorites. Because it works. Basically the model draws from all the samples a subset of samples. Then it fits a model to that small subset. The model does that many times, eg. 1000-5000 times. Finally, the model will calculate an error for all the fitted subsets. The estimated significance gives the reader an understanding of why the clustering is correct. Resampling is done because our main samples came from a small population. Bootstrap considers that all samples are the whole population.

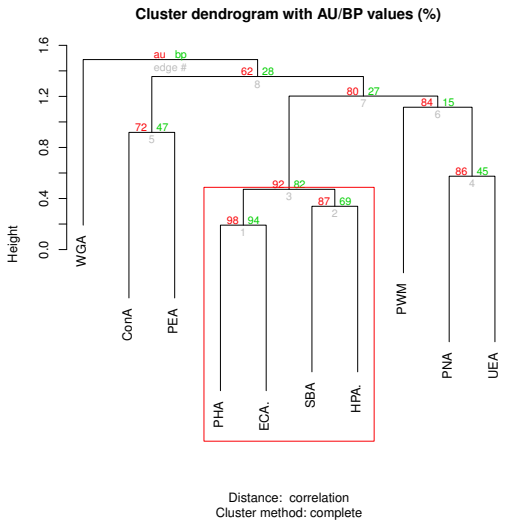
```
sugars.boot <- pvclust(sugars.sc,
                       method.dist = "correlation",
                       method.hclust = "complete",
                       nboot = 5000)

Bootstrap (r = 0.49)... Done.
Bootstrap (r = 0.6)... Done.
Bootstrap (r = 0.68)... Done.
Bootstrap (r = 0.79)... Done.
Bootstrap (r = 0.89)... Done.
Bootstrap (r = 1.0)... Done.
Bootstrap (r = 1.09)... Done.
Bootstrap (r = 1.19)... Done.
Bootstrap (r = 1.3)... Done.
Bootstrap (r = 1.38)... Done.
```

Plot the booststrap results. See significance over 90% in Red. And its like there is 1 significant cluster, To be verified.

```
plot(sugars.boot, hang = 1)
pvrect(sugars.boot, alpha = .90)
```

[1]I also bootstraped the samples they are all significant with percentages over 95.



Get more colours.

```
source("http://faculty.ucr.edu/~tgirke/Documents/R_BioCond/My_R_Scripts/dendroCol.R")
```

Retrieve the significant clusters.

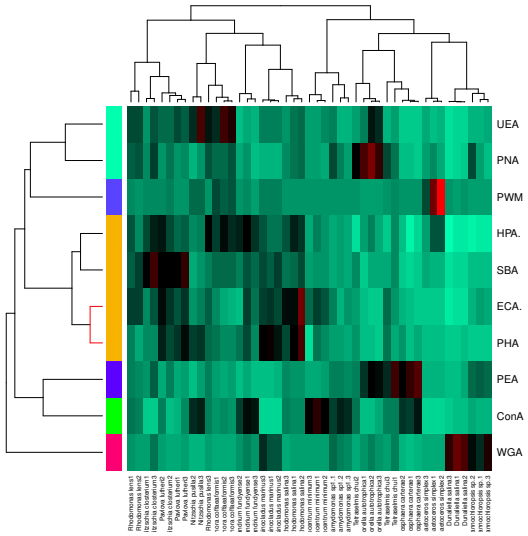
```
clsig <- unlist(pvpick(sugars.boot,
                       alpha=0.95,
                       pv="au",
                       type="geq",
                       max.only=TRUE)$clusters)

dend.colored <- dendrapply(as.dendrogram(sugars.boot$hclust),
                           dendroCol,
                           keys=clsig,
                           xPar="edgePar",
                           bgr="black",
                           fgr="red", pch=20)
```

Draw heatmap with significance.

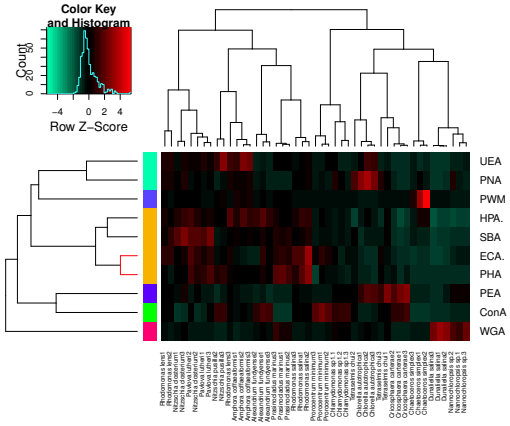
```
heatmap(sugars.sc.tp,
```

```
        Rowv=dend.colored,
        Colv=as.dendrogram(algae.clust),
        col=my.colorFct(),
        scale="row",
        RowSideColors= custom.colors)
```



Draw a better heatmap.

```
heatmap.2(sugars.sc.tp,
          Rowv=dend.colored,
          Colv=as.dendrogram(algae.clust),
          col=my.colorFct(),
          scale="row",
          trace="none",
          RowSideColors=custom.colors,
          margins=c(20,7))
```



7 Logisitc regression

I need a two classes variable.

7.1 Load data

DATA a re in the feeding.xlsx file. I changed the structure of the data. First, algae are either selected (1) or rejected (0). Second, 2 algae used in the same experiment share the same experiment $(E)_n$. *n* is a subscript that design the number of experiments done.

```
feeding <- read.xlsx("./feeding.xlsx", header= TRUE, sheetName = "feeding")
feeding <- feeding[-c(163:165), ]
```

Change rownames and samples.

```
head(feeding)
```

```
                Selected      PHA  ECA.   SBA   HPA.      PWM ConA
1 Chlorella autotrophica  0.08772 0.208 0.223 0.4007  0.01095 1.17
2 Chlorella autotrophica  0.09415 0.117 0.548 0.5549  0.05270 1.16
3 Chlorella autotrophica -0.00308 0.129 0.493 0.3748  0.02538 1.60
4    Nannochloropsis sp.  0.00540 0.328 0.296 0.0868 -0.00597 1.34
5    Nannochloropsis sp. -0.01043 0.242 0.262 0.0657  0.01718 1.50
6    Nannochloropsis sp.  0.01789 0.278 0.282 0.0856 -0.03317 1.28
   PEA   PNA  WGA   UEA
1 3.22 1.907 1.27 0.705
2 4.45 2.274 1.91 1.471
3 4.06 1.788 1.38 1.314
4 2.18 0.312 4.90 0.326
5 2.03 0.858 5.86 0.265
6 2.01 0.122 7.70 0.395
```

```
rownames(feeding) <- paste(feeding[, 1],
                           "sp",
                           gl(nrow(feeding)/3, 3, nrow(feeding)),
                           ".",
                           gl(nrow(feeding)/6, 6,nrow(feeding), labels = c(letters, "xx")),
                           ".",
                           seq(1:162),
                           sep = "")

feeding <- feeding[, -1]
head(feeding)
```

```
                                    PHA  ECA.   SBA   HPA.      PWM
Chlorella autotrophicasp1.a.1  0.08772 0.208 0.223 0.4007  0.01095
Chlorella autotrophicasp1.a.2  0.09415 0.117 0.548 0.5549  0.05270
Chlorella autotrophicasp1.a.3 -0.00308 0.129 0.493 0.3748  0.02538
Nannochloropsis sp.sp2.a.4     0.00540 0.328 0.296 0.0868 -0.00597
Nannochloropsis sp.sp2.a.5    -0.01043 0.242 0.262 0.0657  0.01718
Nannochloropsis sp.sp2.a.6     0.01789 0.278 0.282 0.0856 -0.03317
                               ConA  PEA   PNA  WGA   UEA
Chlorella autotrophicasp1.a.1 1.17 3.22 1.907 1.27 0.705
Chlorella autotrophicasp1.a.2 1.16 4.45 2.274 1.91 1.471
Chlorella autotrophicasp1.a.3 1.60 4.06 1.788 1.38 1.314
Nannochloropsis sp.sp2.a.4    1.34 2.18 0.312 4.90 0.326
Nannochloropsis sp.sp2.a.5    1.50 2.03 0.858 5.86 0.265
Nannochloropsis sp.sp2.a.6    1.28 2.01 0.122 7.70 0.395
```

47 Add 2 more columns. By using *gl* both columns are factors, which is good, for the *glmnet* package.

```
selected <- gl(2, 3, nrow(feeding), labels = c("1", "0"))
```

```
#exp <- paste("E", gl(nrow(feeding)/6, 6, nrow(feeding)), sep = "")
exp <- paste(gl(nrow(feeding)/6, 6, nrow(feeding)))
feeding <- data.frame(feeding, selected = selected, experiments = exp)
feeding[1:6, ]
```

```
                                    PHA  ECA.    SBA   HPA.      PWM
Chlorella autotrophicasp1.a.1  0.08772 0.208 0.223 0.4007  0.01095
Chlorella autotrophicasp1.a.2  0.09415 0.117 0.548 0.5549  0.05270
Chlorella autotrophicasp1.a.3 -0.00308 0.129 0.493 0.3748  0.02538
Nannochloropsis sp.sp2.a.4     0.00540 0.328 0.296 0.0868 -0.00597
Nannochloropsis sp.sp2.a.5    -0.01043 0.242 0.262 0.0657  0.01718
Nannochloropsis sp.sp2.a.6     0.01789 0.278 0.282 0.0856 -0.03317
                               ConA  PEA   PNA  WGA   UEA selected
Chlorella autotrophicasp1.a.1 1.17 3.22 1.907 1.27 0.705        1
Chlorella autotrophicasp1.a.2 1.16 4.45 2.274 1.91 1.471        1
Chlorella autotrophicasp1.a.3 1.60 4.06 1.788 1.38 1.314        1
Nannochloropsis sp.sp2.a.4    1.34 2.18 0.312 4.90 0.326        0
Nannochloropsis sp.sp2.a.5    1.50 2.03 0.858 5.86 0.265        0
Nannochloropsis sp.sp2.a.6    1.28 2.01 0.122 7.70 0.395        0
                               experiments
Chlorella autotrophicasp1.a.1           1
Chlorella autotrophicasp1.a.2           1
Chlorella autotrophicasp1.a.3           1
Nannochloropsis sp.sp2.a.4              1
Nannochloropsis sp.sp2.a.5              1
Nannochloropsis sp.sp2.a.6              1
```

8 Logit

Run a logistic regression on the numerical data.

```
fit <- glm(selected ~ .,
           data = feeding,
           #data = feeding[, -ncol(feeding)],
           family = "binomial")
```

Show a summary of the fit with coefficients and p-values.

```
summary(fit)
```

```
Call:
glm(formula = selected ~ ., family = "binomial", data = feeding)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0082  -0.6166   0.0029   0.6346   2.0455

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   2.15901    2.99574    0.72   0.4711
PHA           2.08470    1.67669    1.24   0.2137
ECA.         -6.80105    2.21435   -3.07   0.0021 **
SBA           2.29879    1.06143    2.17   0.0303 *
HPA.         -4.63538    2.64149   -1.75   0.0793 .
PWM           5.25779    2.17706    2.42   0.0157 *
ConA          0.18631    0.32745    0.57   0.5694
PEA          -0.77410    0.37097   -2.09   0.0369 *
PNA           0.00985    1.04684    0.01   0.9925
WGA           0.05475    0.26083    0.21   0.8337
UEA           1.90730    1.73538    1.10   0.2717
experiments10 -0.49517    1.46808   -0.34   0.7359
experiments11 -2.52790    2.34399   -1.08   0.2808
experiments12 -6.12649   11.36591   -0.54   0.5899
experiments13 -0.12664    1.45566   -0.09   0.9307
experiments14 -2.53450    1.46707   -1.73   0.0841 .
experiments15 -1.43869    1.65074   -0.87   0.3835
experiments16  1.23229    1.48425    0.83   0.4064
experiments17 -0.74682    1.34743   -0.55   0.5794
experiments18  0.33383    1.31724    0.25   0.7999
experiments19 -6.77770    8.40604   -0.81   0.4201
experiments2  -1.75993    1.51350   -1.16   0.2449
experiments20  0.64356    1.62467    0.40   0.6920
experiments21 -2.06295    1.58497   -1.30   0.1931
experiments22 -1.10536    1.55571   -0.71   0.4774
experiments23 -5.53478   15.11977   -0.37   0.7143
experiments24 -7.64155    5.77969   -1.32   0.1861
experiments25 -0.76227    1.54114   -0.49   0.6209
experiments26 -2.40676    1.71779   -1.40   0.1612
experiments27 -0.82325    1.94624   -0.42   0.6723
experiments3  -1.79062    1.64217   -1.09   0.2755
experiments4  -0.82007    2.10541   -0.39   0.6969
experiments5  -4.55254    1.66537   -2.73   0.0063 **
experiments6  -2.23694    1.47670   -1.51   0.1298
experiments7  -7.25685    6.76331   -1.07   0.2833
experiments8  -3.64978    1.83436   -1.99   0.0466 *
experiments9  -6.95653    7.73904   -0.90   0.3687
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 224.58  on 161  degrees of freedom
Residual deviance: 122.33  on 125  degrees of freedom
AIC: 196.3

Number of Fisher Scoring iterations: 7
```

51 Calculate the confidence intervals using the log-likelihood from the logit model.

```
confint(fit)
```

```
Warning:  glm.fit:  fitted probabilities numerically 0 or 1 occurred
```

```
                 2.5 %  97.5 %
(Intercept)    -3.603  8.2840
PHA            -1.105  5.5566
ECA.          -11.631 -2.8432
SBA             0.371  4.5588
HPA.          -10.154  0.3303
PWM             1.628  9.9986
ConA           -0.452  0.8450
PEA            -1.553 -0.0837
PNA            -2.084  2.0825
WGA            -0.470  0.5633
UEA            -1.387  5.5220
experiments10  -3.416  2.4277
experiments11  -7.234  1.5376
experiments12 -18.423  1.6361
experiments13  -3.028  2.7754
experiments14  -5.485  0.3503
experiments15  -4.715  1.8192
experiments16  -1.678  4.2328
experiments17  -3.478  1.9245
experiments18  -2.282  2.9931
experiments19 -18.468  0.4432
experiments2   -4.859  1.1492
experiments20  -2.584  3.8861
experiments21  -5.189  1.1282
experiments22  -4.196  1.9795
experiments23 -18.626  2.9382
experiments24 -18.475 -1.0471
experiments25  -3.826  2.3006
experiments26  -6.022  0.8683
experiments27  -4.497  3.0216
experiments3   -5.450  1.2494
experiments4   -5.185  3.1116
experiments5   -8.084 -1.4088
experiments6   -5.207  0.6941
experiments7  -18.475 -0.9211
experiments8   -7.564 -0.2419
experiments9  -18.473 -0.2001
```

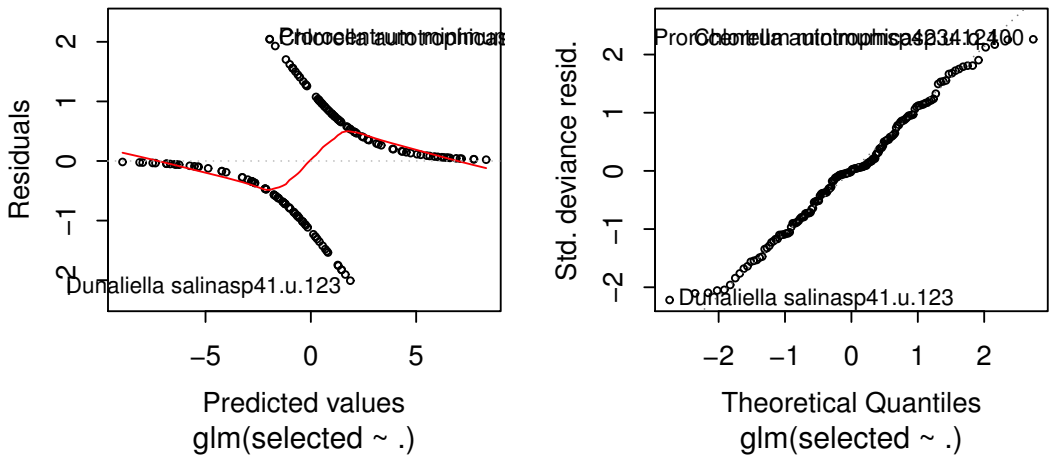
52 Another way is to get the CIs from the standard errors. Same as above.

```
confint.default(fit)
```

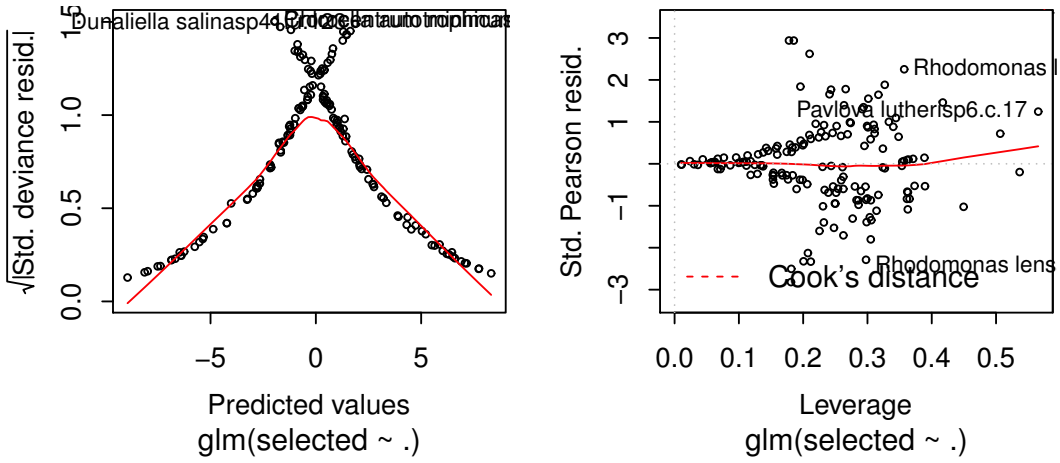
```
                 2.5 %  97.5 %
(Intercept)     -3.713  8.0306
PHA             -1.202  5.3710
ECA.           -11.141 -2.4610
SBA              0.218  4.3792
HPA.            -9.813  0.5418
PWM              0.991  9.5247
ConA            -0.455  0.8281
PEA             -1.501 -0.0470
PNA             -2.042  2.0616
WGA             -0.456  0.5660
UEA             -1.494  5.3086
experiments10   -3.373  2.3822
experiments11   -7.122  2.0662
experiments12 -28.403 16.1503
experiments13   -2.980  2.7264
experiments14   -5.410  0.3409
experiments15   -4.674  1.7967
experiments16   -1.677  4.1414
experiments17   -3.388  1.8941
experiments18   -2.248  2.9156
experiments19 -23.253  9.6978
experiments2    -4.726  1.2065
experiments20   -2.541  3.8279
experiments21   -5.169  1.0435
experiments22   -4.154  1.9438
experiments23 -35.169 24.0994
experiments24 -18.970  3.6864
experiments25   -3.783  2.2583
experiments26   -5.774  0.9601
experiments27   -4.638  2.9913
experiments3    -5.009  1.4280
experiments4    -4.947  3.3065
experiments5    -7.817 -1.2885
experiments6    -5.131  0.6573
experiments7   -20.513  5.9990
experiments8    -7.245 -0.0545
experiments9   -22.125  8.2117
```

53 Plot the logit.

```
par(mar=c(4,4,.1,.1),cex.lab=.95,cex.axis=.9,mgp=c(2,.7,0),tcl=-.3)
plot(fit, cex = .5)
```



54



9 System Information

The version number of R and packages loaded for generating the vignette were:

```
###save(list=ls(pattern=".*|.*"),file="PD.Rdata")
sessionInfo()

R version 3.1.2 (2014-10-31)
Platform: x86_64-unknown-linux-gnu (64-bit)

locale:
 [1] LC_CTYPE=en_US.UTF-8          LC_NUMERIC=C
 [3] LC_TIME=en_US.UTF-8           LC_COLLATE=en_US.UTF-8
 [5] LC_MONETARY=en_US.UTF-8       LC_MESSAGES=en_US.UTF-8
 [7] LC_PAPER=en_US.UTF-8          LC_NAME=en_US.UTF-8
 [9] LC_ADDRESS=en_US.UTF-8        LC_TELEPHONE=en_US.UTF-8
[11] LC_MEASUREMENT=en_US.UTF-8    LC_IDENTIFICATION=en_US.UTF-8

attached base packages:
[1] stats     graphics  grDevices utils     datasets  methods
[7] base

other attached packages:
 [1] ISLR_1.0            boot_1.3-13         knitr_1.8
 [4] gplots_2.14.2       pvclust_1.3-0       latticeExtra_0.6-26
 [7] RColorBrewer_1.0-5  glmnet_1.9-8        Matrix_1.1-4
[10] leaps_2.9           caret_6.0-37        ggplot2_1.0.0
[13] lattice_0.20-29     xlsx_0.5.7          xlsxjars_0.6.1
[16] rJava_0.9-6

loaded via a namespace (and not attached):
 [1] bitops_1.0-6         BradleyTerry2_1.0-5 brglm_0.5-9
 [4] car_2.0-22           caTools_1.17.1      codetools_0.2-9
 [7] colorspace_1.2-4     compiler_3.1.2      digest_0.6.4
[10] evaluate_0.5.5       foreach_1.4.2       formatR_1.0
[13] gdata_2.13.3         grid_3.1.2          gtable_0.1.2
[16] gtools_3.4.1         highr_0.4           iterators_1.0.7
[19] KernSmooth_2.23-13   lme4_1.1-7          MASS_7.3-35
[22] minqa_1.2.4          munsell_0.4.2       nlme_3.1-118
[25] nloptr_1.0.4         nnet_7.3-8          plyr_1.8.1
[28] proto_0.3-10         Rcpp_0.11.3         reshape2_1.4
[31] scales_0.2.4         splines_3.1.2       stringr_0.6.2
[34] tools_3.1.2
```