Model selection in GLMs

Joe Parker

9th November 2018

Root and leaf microbiomes

The data in this practical are taken from Wagner *et al.* (2016; `doi:0.1038/ncomms12151`). This is an in-depth comparison of the root and leave microbiomes (microbial community composition) in experimental populations of the wild mustard plant, *Boechera stricta* (Brassicaceae), investigated using metagenomic DNA sequencing. The plants were planted out carefully-designed plots under controlled conditions, and some individuals from each plot were removed each year, their roots and leaves harvested, and microbial DNA extracted and sequenced. The bioinformatics packages `QIIME` and `Phyloseq` were used to process the data, and statistical analysis was performed using `R` - just as you're about to do…

Experimental design

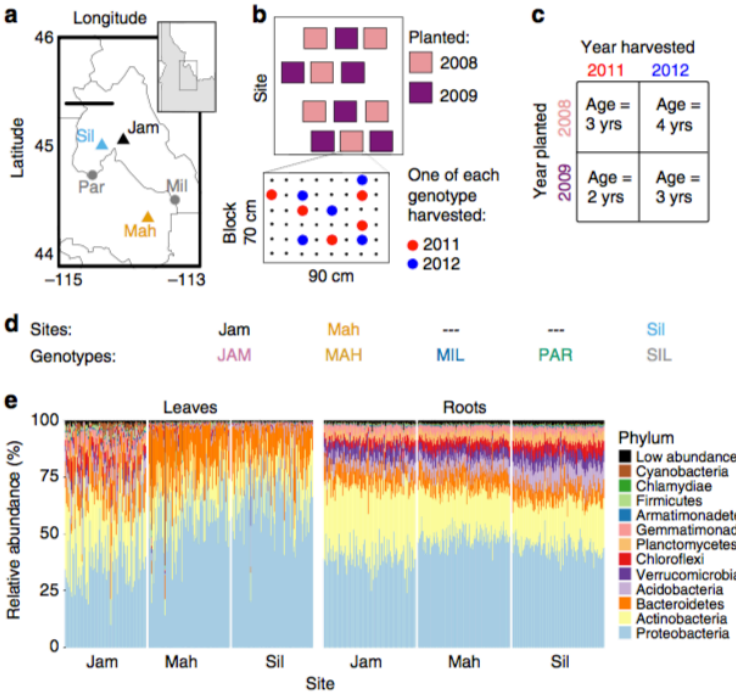


Figure 1 | Summary of experimental design and analysis. (**a**) Map of the study region in central Idaho, USA (map dat
The five genotypes used in this experiment were collected from the five *B. stricta* populations shown. We collected see
plants from each population, for a total of 48 accessions or genetic 'lines'. For our analyses these lines were grouped b
corresponding to the populations from which their ancestors were collected. The populations marked with triangles corr
common gardens where the experiment took place. Scale bar, 50 km. (**b**) Schematic representation of common garden l
replicated, randomized blocks per planting cohort (2008 and 2009). Each block contained one replicate of each 'line', f
'genotype'. In both 2011 and 2012, one individual of each genotype was haphazardly chosen for destructive sampling in
staggered planting/harvesting design disentangled the effects of plant age and year of observation. (**d**) Abbreviations an
five genotypes and three sites featured in this study. (**e**) The relative abundances of major phyla are shown for each le

The experimental design is given below:
And also from the Results:

```
These 616 samples represented 440 individual plants across three common gardens (sites), 36 experimenta
```

Questions:

1. Which factors are blocks, and which are treatments?
2. How many levels are there in this experiment?
3. How many replicates? Are there any pseudoreplicates? Why?
4. Is this design orthogonal? If not, why not?
5. Is there anything else worth noting about the design?

Model selection and fitting by heuristics

The data for this paper is held in Dryad, a large open-access repository of research paper data and analyses. If you want to explore it yourself, head to http://datadryad.org/resource/doi:10.5061/dryad.g60r3; I've simplified the data slightly for this practical.

1. Open the file `wagner_2016_microbiome.tdf` by reading it into `R`. Inspect the data frame and verify that variables have imported correctly (categorical variables as factors, etc - *Hint, you may need to clean your data, or use the* `as.factor()` or `as.numeric()` *conversion functions*).

```
wagner = read.table("R_code_Wagner_etal_2016/wagner_2016_microbiome.tdf",header=T,sep="\t")
wagner$planting = as.factor(wagner$plant_year)
wagner$sampling = as.factor(wagner$sample_year)
wagner$miseq_run = as.factor(wagner$miseq)
attach(wagner)
```

We are seeking to explain microbial diversity (measured by two variables, `CHAO1` and `Shannon`) using the available factors. To start with, let's try and fit a few models using the available parameters. There are a lot of variables here, so rather than eyeball each one separately, we can call `plot()` on the data.frame to produce simple pairwise scatterplots for all variables:

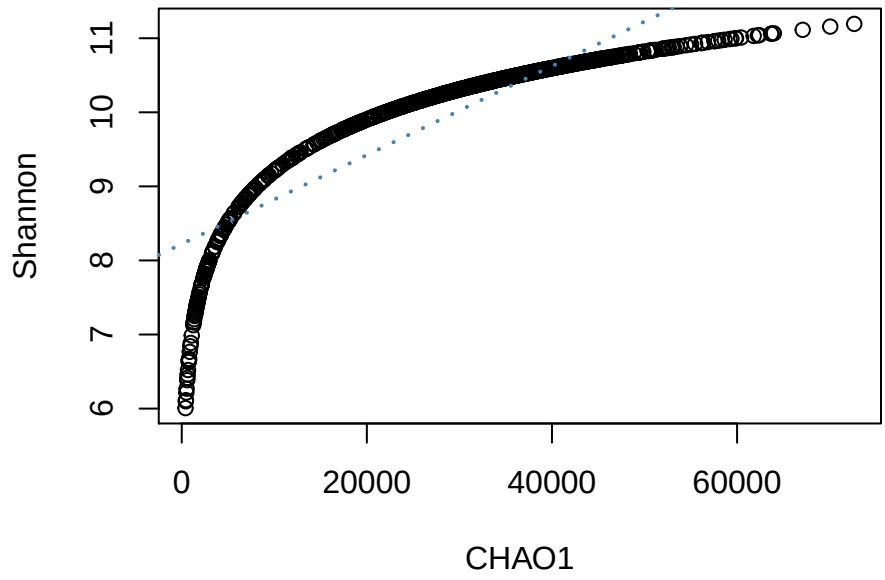
```
plot(wagner)
```

2. There are actually two response variables in this dataset, which represent alternative measures of microbial diversity: `CHAO1` and `Shannon`. Which should we use, we wonder? Surely they should be highly correlated if both measure the same thing - but is there anything weird here? Try plotting them, and fitting a model (*BIG hint: check your assumptions!*)

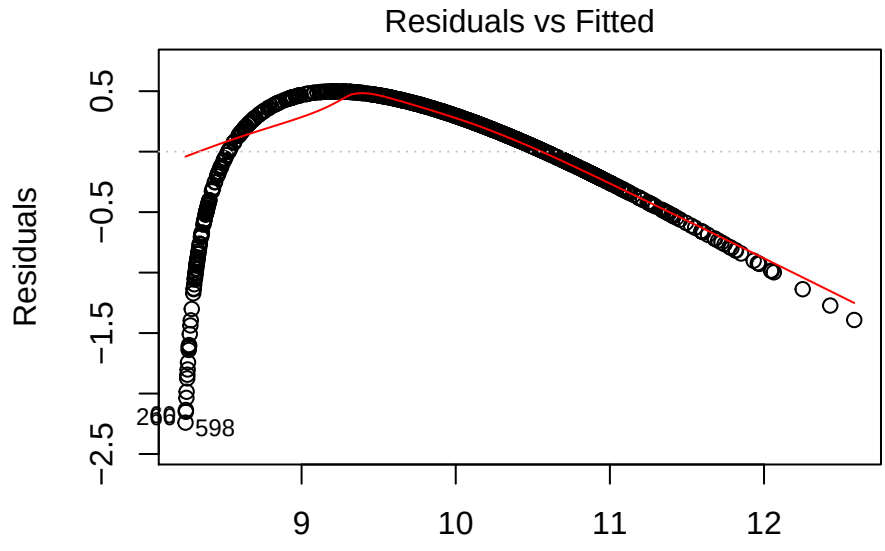
```
plot(CHAO1,Shannon)
response_model=lm(Shannon~CHAO1)
anova(response_model)
```

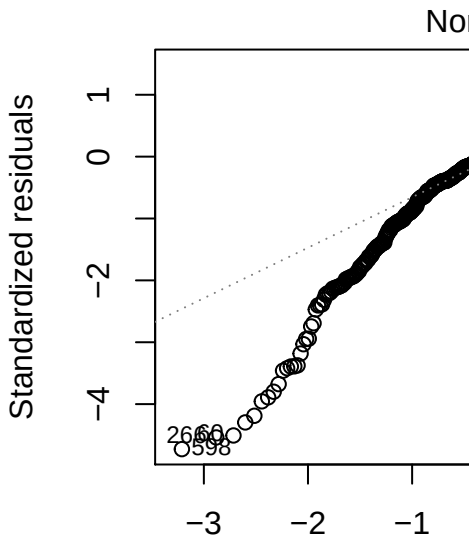
```
## Analysis of Variance Table
##
## Response: Shannon
##            Df Sum Sq Mean Sq F value    Pr(>F)
## CHAO1       1 670.99  670.99  2971.4 < 2.2e-16 ***
## Residuals 754 170.27    0.23
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

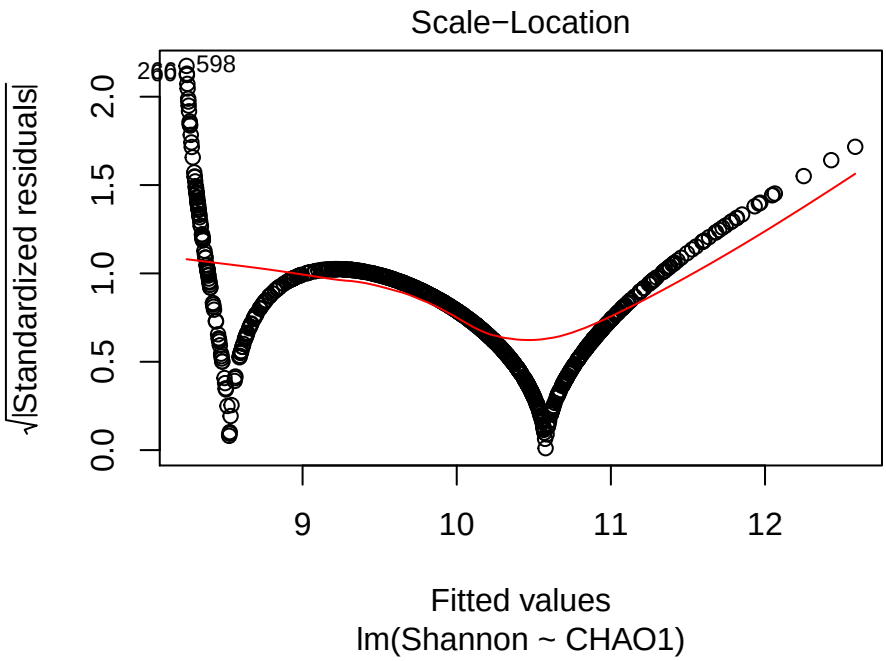
```
abline(response_model,col="steelblue",lwd=2,lty=3)
```

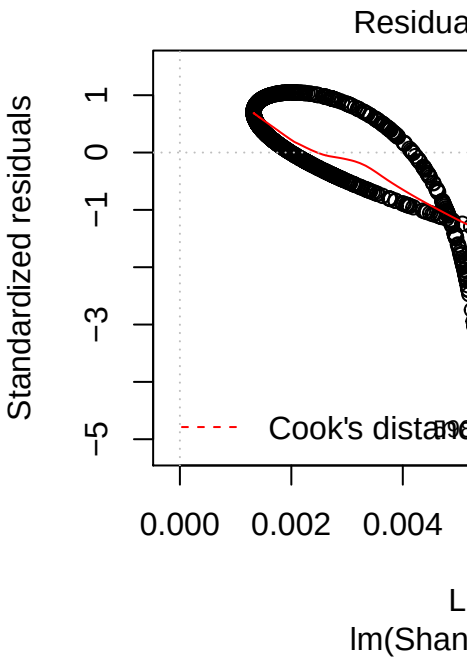



```
plot(response_model)
```



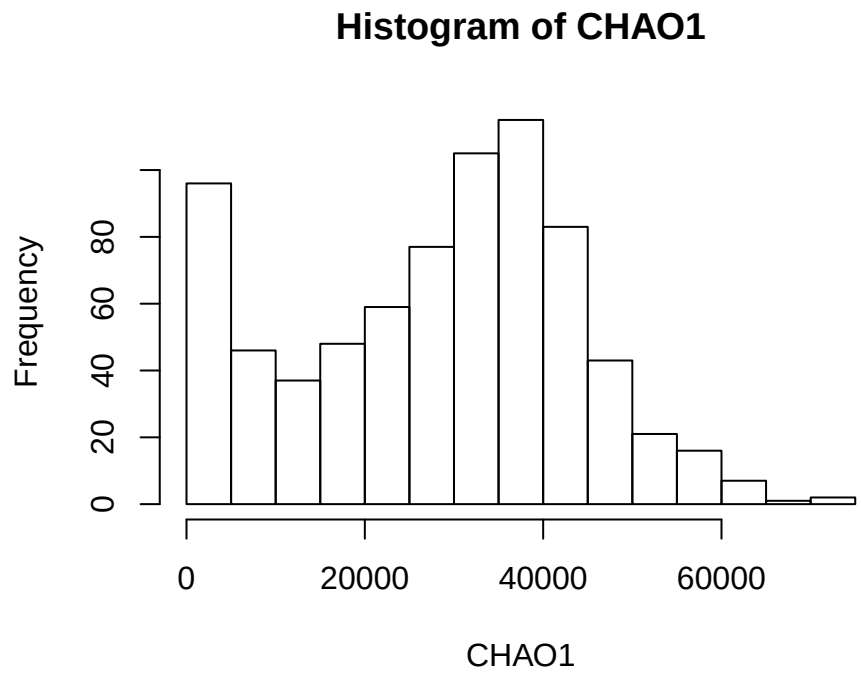






What do you conclude? Try plotting each response variable as a histogram before you decide which to use as the response variable for the rest of the analysis.

```
hist(CHAO1)
```

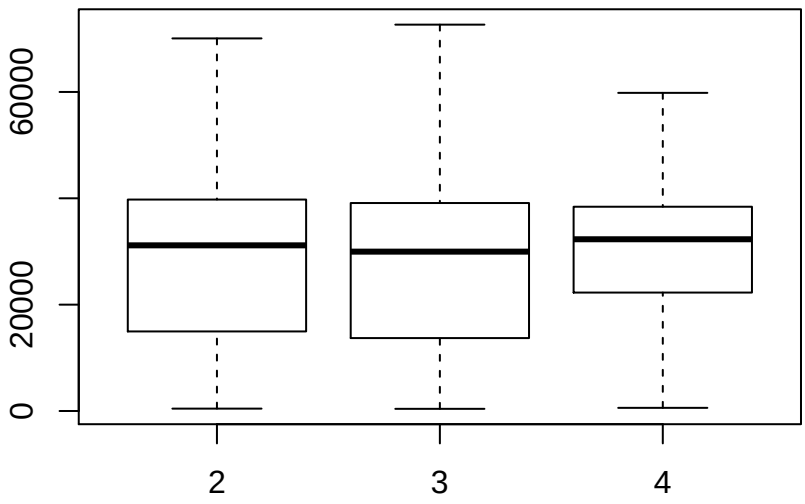



```
hist(Shannon)
```

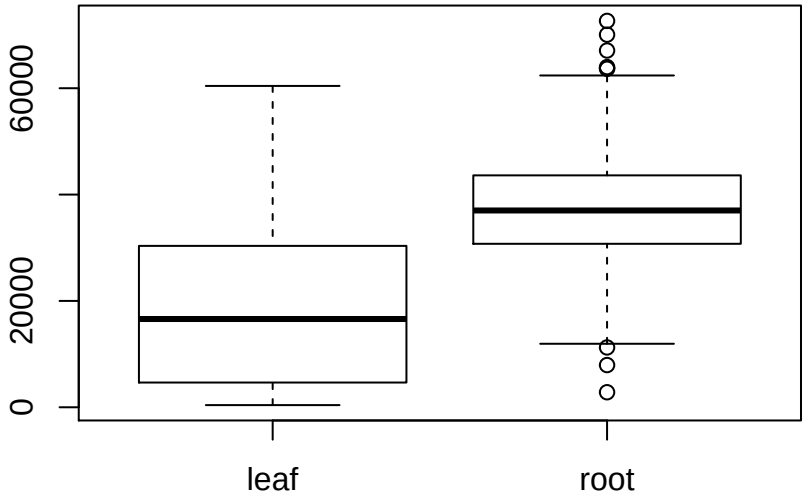
CHAO1 is normally distributed, but Shannon is not. We'll use Shannon for the analysis as it makes things simple for us.

3. Some of these look particularly interesting in relation to our response variable (either `CHAO1` or `Shannon`). Produce three boxplots of variables that seem like they may have explanatory power.

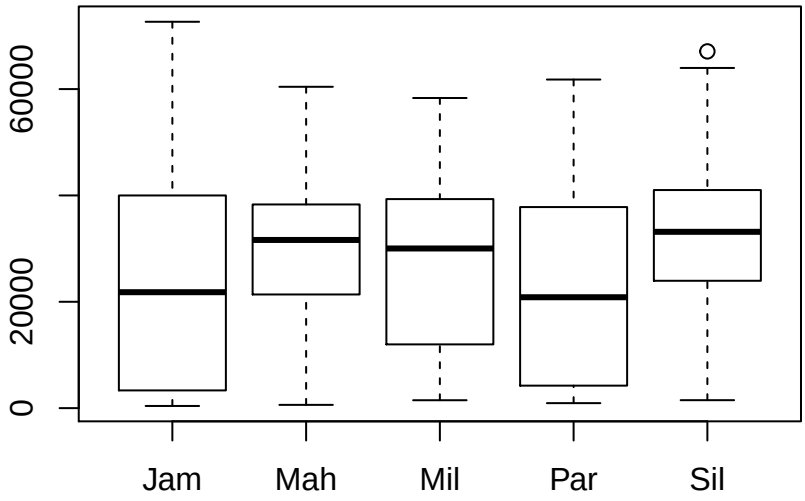
```
# good / interesting / simple ones:
boxplot(CHAO1 ~ age)
```



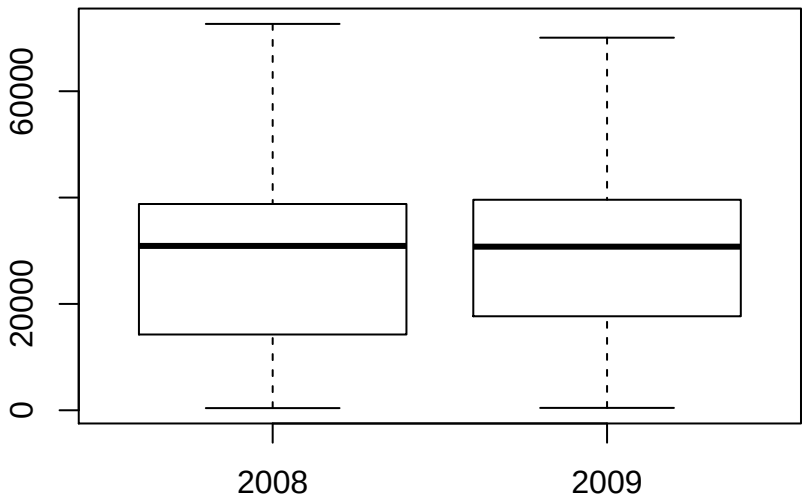

```
boxplot(CHAO1 ~ habitat)
```



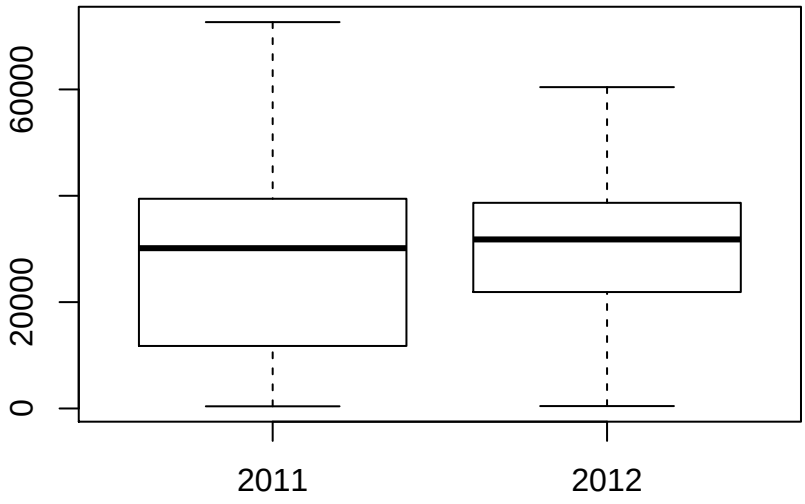

```
boxplot(CHAO1 ~ site)
```



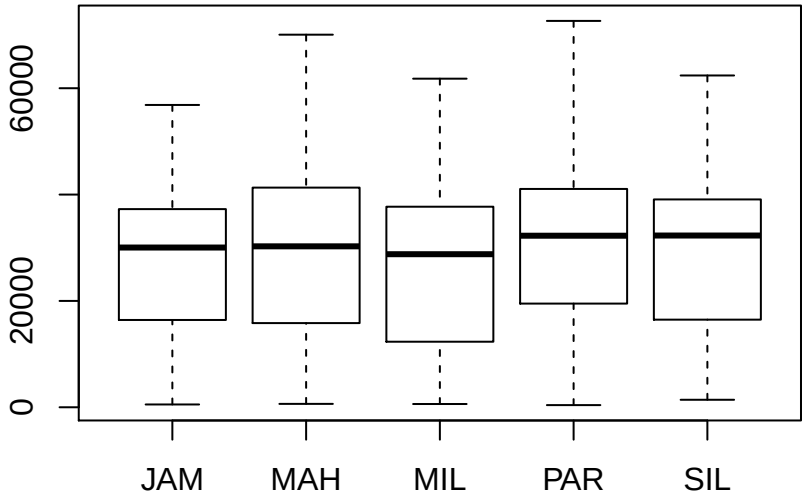

```
boxplot(CHAO1 ~ planting)
```



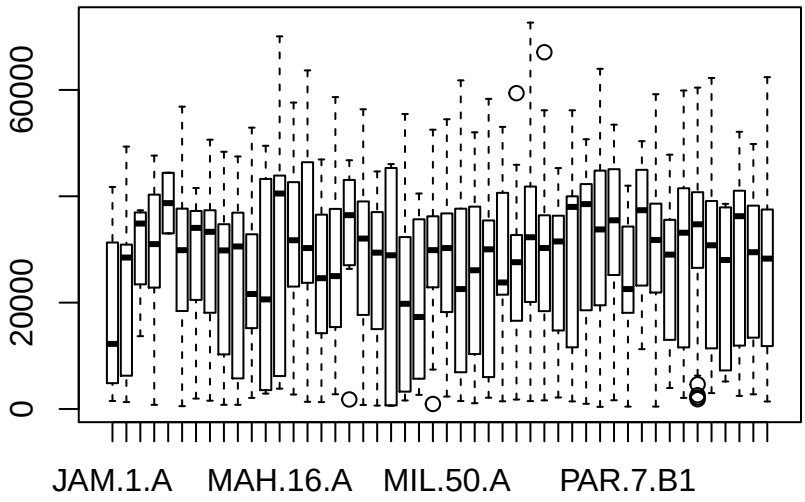

```
boxplot(CHAO1 ~ sampling)
```



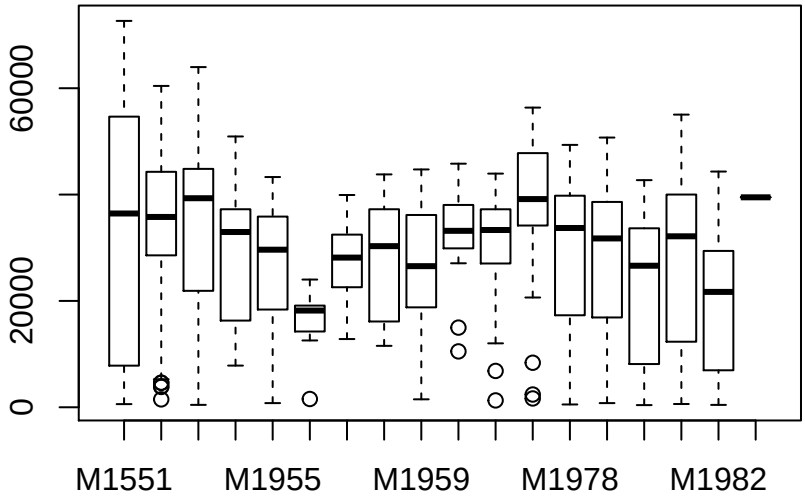

```
boxplot(CHAO1 ~ geno)
```



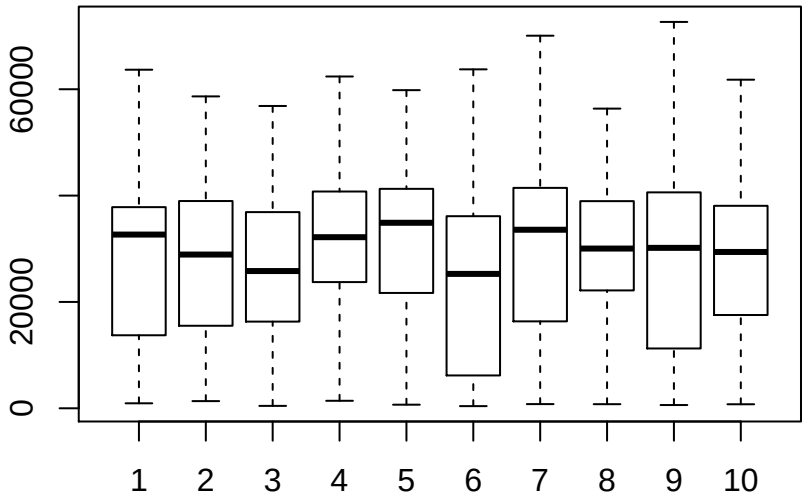

```
# orrible ones, a lot of levels so maybe notsomuch
boxplot(CHAO1 ~ line1)
```




```
boxplot(CHAO1 ~ block)
```




```
boxplot(CHAO1 ~ miseq_run)
```



4. Now fit a model using these terms. Start with a maximal model, and try to refine it to produce a minimal adequate model, by any means you see fit.

```
my_model = lm(CHAO1 ~ habitat * sampling * line1)
anova(my_model)
```

```
## Analysis of Variance Table
##
## Response: CHAO1
##                          Df     Sum Sq    Mean Sq  F value    Pr(>F)
```

```
## habitat                  1 6.4797e+10 6.4797e+10 436.9022 < 2.2e-16 ***
## sampling                 1 1.3761e+09 1.3761e+09   9.2784  0.002415 **
## line1                   47 6.5391e+09 1.3913e+08   0.9381  0.592572
## habitat:sampling         1 1.1049e+10 1.1049e+10  74.5000 < 2.2e-16 ***
## habitat:line1           47 5.9893e+09 1.2743e+08   0.8592  0.736259
## sampling:line1          16 1.3172e+09 8.2327e+07   0.5551  0.916795
## habitat:sampling:line1   8 9.9782e+08 1.2473e+08   0.8410  0.566680
## Residuals              634 9.4029e+10 1.4831e+08
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
drop1(my_model,test = "F")
```

```
## Single term deletions
##
## Model:
## CHAO1 ~ habitat * sampling * line1
##                         Df Sum of Sq        RSS   AIC F value Pr(>F)
## <none>                               9.4029e+10 14335
## habitat:sampling:line1   8 997816895 9.5027e+10 14327   0.841 0.5667
```

```
# line1 looks useless, drop it
my_model2 = update(my_model, . ~ . - line1)
anova(my_model2)
```

```
## Analysis of Variance Table
##
## Response: CHAO1
##                          Df     Sum Sq    Mean Sq  F value    Pr(>F)
## habitat                   1 6.4797e+10 6.4797e+10 436.9022 < 2.2e-16 ***
## sampling                  1 1.3761e+09 1.3761e+09   9.2784  0.002415 **
## habitat:sampling          1 1.1285e+10 1.1285e+10  76.0914 < 2.2e-16 ***
## habitat:line1            94 1.2292e+10 1.3077e+08   0.8817  0.774379
## sampling:line1           16 1.3172e+09 8.2327e+07   0.5551  0.916795
## habitat:sampling:line1    8 9.9782e+08 1.2473e+08   0.8410  0.566680
## Residuals               634 9.4029e+10 1.4831e+08
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# interactions aren't significant, or sensible. lose them
my_model3 = lm(CHAO1 ~ habitat * sampling)
anova(my_model3)
```

```
## Analysis of Variance Table
##
## Response: CHAO1
##                   Df     Sum Sq    Mean Sq  F value    Pr(>F)
## habitat            1 6.4797e+10 6.4797e+10 448.5382 < 2.2e-16 ***
## sampling           1 1.3761e+09 1.3761e+09   9.5255  0.002101 **
## habitat:sampling   1 1.1285e+10 1.1285e+10  78.1179 < 2.2e-16 ***
## Residuals        752 1.0864e+11 1.4446e+08
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
drop1(my_model3)
```

```
## Single term deletions
##
## Model:
## CHAO1 ~ habitat * sampling
##                   Df  Sum of Sq        RSS   AIC
## <none>                          1.0864e+11 14208
## habitat:sampling   1 1.1285e+10 1.1992e+11 14281
```

```r
# looks appropriate - try adding another variable
add1(my_model3,. ~ . + block,test="F")
```

```
## Single term additions
##
## Model:
## CHAO1 ~ habitat * sampling
##         Df  Sum of Sq        RSS   AIC F value    Pr(>F)
## <none>                1.0864e+11 14208
## block   17 1.7425e+10 9.1211e+10 14110  8.2598 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
my_model4 = update(my_model3, . ~ . + block)
anova(my_model4)
```

```
## Analysis of Variance Table
##
## Response: CHAO1
##                   Df     Sum Sq    Mean Sq F value    Pr(>F)
## habitat            1 6.4797e+10 6.4797e+10 522.151 < 2.2e-16 ***
## sampling           1 1.3761e+09 1.3761e+09  11.089 0.0009118 ***
## block             17 2.4085e+10 1.4168e+09  11.417 < 2.2e-16 ***
## habitat:sampling   1 4.6250e+09 4.6250e+09  37.269 1.665e-09 ***
## Residuals        735 9.1211e+10 1.2410e+08
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# etc...
```

5. Did you forget to check anything? ;) *Inspect residuals/diagnostic plots*

Model selection by stepwise AIC

So far, so good - but there are a *lot* of possible combinations here. We'll use an automated model selection function, `step()` to have a go.

5. Use the `step()` function to fit a model describing microbial diversity in terms of tissue (leaf/root), site, age, genotype, year, block, line, and MiSeq run, using forwards search. Start with a simple linear regression of age vs. diversity.

```r
# backwards
backwards_final=step(lm(CHAO1 ~ habitat * sampling * block * site),direction="backward")
```

```
## Start:  AIC=13964.62
## CHAO1 ~ habitat * sampling * block * site
##
##
## Step:  AIC=13964.62
## CHAO1 ~ habitat + sampling + block + site + habitat:sampling +
##     habitat:block + sampling:block + habitat:site + sampling:site +
##     block:site + habitat:sampling:block + habitat:sampling:site +
##     habitat:block:site + sampling:block:site
##
##                           Df Sum of Sq        RSS   AIC
## - habitat:block:site      15 927984422 5.4004e+10 13948
## - sampling:block:site      7 410314129 5.3487e+10 13956
## - habitat:sampling:block   4  70248006 5.3146e+10 13958
## - habitat:sampling:site    2  17991546 5.3094e+10 13961
## <none>                                 5.3076e+10 13965
##
## Step:  AIC=13947.72
## CHAO1 ~ habitat + sampling + block + site + habitat:sampling +
##     habitat:block + sampling:block + habitat:site + sampling:site +
```

```
##     block:site + habitat:sampling:block + habitat:sampling:site +
##     sampling:block:site
##
##                          Df  Sum of Sq        RSS   AIC
## - sampling:block:site    13 1261365216 5.5266e+10 13939
## - habitat:sampling:block  5  272591897 5.4277e+10 13942
## - habitat:sampling:site   2   63924277 5.4068e+10 13945
## <none>                                 5.4004e+10 13948
##
## Step:  AIC=13939.18
## CHAO1 ~ habitat + sampling + block + site + habitat:sampling +
##     habitat:block + sampling:block + habitat:site + sampling:site +
##     block:site + habitat:sampling:block + habitat:sampling:site
##
##                          Df  Sum of Sq        RSS   AIC
## - block:site             53 3529506212 5.8795e+10 13880
## - habitat:sampling:block  5  312752041 5.5578e+10 13933
## <none>                                 5.5266e+10 13939
## - habitat:sampling:site   3  545657803 5.5811e+10 13941
##
## Step:  AIC=13879.98
## CHAO1 ~ habitat + sampling + block + site + habitat:sampling +
##     habitat:block + sampling:block + habitat:site + sampling:site +
##     habitat:sampling:block + habitat:sampling:site
##
##                          Df Sum of Sq        RSS   AIC
## - habitat:sampling:block  5 324223402 5.9119e+10 13874
## <none>                               5.8795e+10 13880
## - habitat:sampling:site   3 906608258 5.9702e+10 13886
##
## Step:  AIC=13874.14
## CHAO1 ~ habitat + sampling + block + site + habitat:sampling +
##     habitat:block + sampling:block + habitat:site + sampling:site +
##     habitat:sampling:site
##
##                          Df  Sum of Sq        RSS   AIC
## - sampling:block         15 1429683241 6.0549e+10 13862
## <none>                               5.9119e+10 13874
## - habitat:sampling:site   3  899478509 6.0019e+10 13880
## - habitat:block          15 3481185477 6.2601e+10 13887
##
## Step:  AIC=13862.2
## CHAO1 ~ habitat + sampling + block + site + habitat:sampling +
##     habitat:block + habitat:site + sampling:site + habitat:sampling:site
##
##                          Df  Sum of Sq        RSS   AIC
## <none>                               6.0549e+10 13862
## - habitat:block          16 3471411436 6.4020e+10 13872
## - habitat:sampling:site   3 1332654239 6.1882e+10 13873
```

```
# forwards
forward_final=step(lm(CHAO1 ~ age * habitat),scope=(~age*habitat*block*sampling),direction="forward")
```

```
## Start:  AIC=14269.89
## CHAO1 ~ age * habitat
##
##            Df  Sum of Sq        RSS   AIC
## + block    17 2.0875e+10 9.7010e+10 14156
## + sampling  1 1.3059e+09 1.1658e+11 14264
## <none>                   1.1789e+11 14270
##
```

```
## Step:  AIC=14156.55
## CHAO1 ~ age + habitat + block + age:habitat
##
##                  Df  Sum of Sq        RSS   AIC
## + habitat:block 16 7588696956 8.9421e+10 14127
## + sampling       1 1611284604 9.5399e+10 14146
## <none>                         9.7010e+10 14156
## + age:block     16 3464094765 9.3546e+10 14161
##
## Step:  AIC=14126.97
## CHAO1 ~ age + habitat + block + age:habitat + habitat:block
##
##             Df  Sum of Sq        RSS   AIC
## + sampling   1 4088775580 8.5332e+10 14094
## <none>                    8.9421e+10 14127
## + age:block 16 2040251929 8.7381e+10 14142
##
## Step:  AIC=14093.59
## CHAO1 ~ age + habitat + block + sampling + age:habitat + habitat:block
##
##                     Df  Sum of Sq        RSS   AIC
## + block:sampling   15 6212344973 7.9120e+10 14066
## + habitat:sampling  1 2963176807 8.2369e+10 14069
## <none>                           8.5332e+10 14094
## + age:sampling      1   35266820 8.5297e+10 14095
## + age:block        16 1986436761 8.3346e+10 14108
##
## Step:  AIC=14066.44
## CHAO1 ~ age + habitat + block + sampling + age:habitat + habitat:block +
##     block:sampling
##
##                     Df  Sum of Sq        RSS   AIC
## + habitat:sampling  1 1242090386 7.7878e+10 14056
## <none>                           7.9120e+10 14066
## + age:sampling      1    9676451 7.9110e+10 14068
## + age:block        16 1178958898 7.7941e+10 14087
##
## Step:  AIC=14056.48
## CHAO1 ~ age + habitat + block + sampling + age:habitat + habitat:block +
##     block:sampling + habitat:sampling
##
##                           Df  Sum of Sq        RSS   AIC
## <none>                                   7.7878e+10 14056
## + age:sampling             1   30616725 7.7847e+10 14058
## + habitat:block:sampling  5  521277697 7.7357e+10 14061
## + age:block               16 1209768501 7.6668e+10 14077
```

6. Write out the full model equation, including fitted terms, for the final model.

```
anova(my_model,my_model2,my_model3,my_model4,forward_final)
```

```
## Analysis of Variance Table
##
## Model 1: CHAO1 ~ habitat * sampling * line1
## Model 2: CHAO1 ~ habitat + sampling + habitat:sampling + habitat:line1 +
##     sampling:line1 + habitat:sampling:line1
## Model 3: CHAO1 ~ habitat * sampling
## Model 4: CHAO1 ~ habitat + sampling + block + habitat:sampling
## Model 5: CHAO1 ~ age + habitat + block + sampling + age:habitat + habitat:block +
##     block:sampling + habitat:sampling
##   Res.Df        RSS   Df   Sum of Sq      F    Pr(>F)
## 1    634 9.4029e+10
```

```
## 2    634 9.4029e+10    0  0.0000e+00
## 3    752 1.0864e+11 -118 -1.4607e+10 0.8347    0.8878
## 4    735 9.1211e+10   17  1.7425e+10 6.9112 1.828e-15 ***
## 5    702 7.7878e+10   33  1.3333e+10 2.7243 1.264e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(forward_final)
```

```
##
## Call:
## lm(formula = CHAO1 ~ age + habitat + block + sampling + age:habitat +
##     habitat:block + block:sampling + habitat:sampling)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -31669  -6039   -256   6149  37060
##
## Coefficients: (3 not defined because of singularities)
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  14346       4136   3.468 0.000556 ***
## age                          -2592       1152  -2.249 0.024799 *
## habitatroot                  33068       5286   6.256 6.85e-10 ***
## blockM1554                    8572       4077   2.103 0.035859 *
## blockM1555                    8224       3807   2.160 0.031093 *
## blockM1690                   18087       7876   2.296 0.021950 *
## blockM1955                    8332       3683   2.262 0.023979 *
## blockM1956                    7123       4739   1.503 0.133267
## blockM1957                   14939       5033   2.968 0.003099 **
## blockM1958                   12582       5033   2.500 0.012659 *
## blockM1959                    9427       4125   2.285 0.022585 *
## blockM1960                   23323       4364   5.345 1.23e-07 ***
## blockM1961                   19629       4076   4.816 1.79e-06 ***
## blockM1977                   17798       4174   4.264 2.28e-05 ***
## blockM1978                    7113       4205   1.691 0.091214 .
## blockM1979                    2986       3625   0.824 0.410327
## blockM1980                    4762       3076   1.548 0.122031
## blockM1981                    5725       3115   1.838 0.066459 .
## blockM1982                   -1985       3223  -0.616 0.538288
## blockM1992                  -28004      11664  -2.401 0.016609 *
## sampling2012                 25966       3751   6.923 9.97e-12 ***
## age:habitatroot               4103       1602   2.562 0.010628 *
## habitatroot:blockM1554      -18055       4904  -3.682 0.000249 ***
## habitatroot:blockM1555      -15036       4623  -3.253 0.001199 **
## habitatroot:blockM1690       -2936      12905  -0.228 0.820073
## habitatroot:blockM1955      -26909       4799  -5.608 2.95e-08 ***
## habitatroot:blockM1956      -36190       9359  -3.867 0.000121 ***
## habitatroot:blockM1957      -54260       8651  -6.272 6.23e-10 ***
## habitatroot:blockM1958       -8829      12170  -0.725 0.468414
## habitatroot:blockM1959      -35989      11803  -3.049 0.002380 **
## habitatroot:blockM1960      -20884       8420  -2.480 0.013358 *
## habitatroot:blockM1961      -14422      11846  -1.217 0.223854
## habitatroot:blockM1977      -25716       5086  -5.056 5.47e-07 ***
## habitatroot:blockM1978      -21059       4973  -4.235 2.60e-05 ***
## habitatroot:blockM1979      -18196       4407  -4.129 4.09e-05 ***
## habitatroot:blockM1980      -24989       4157  -6.011 2.96e-09 ***
## habitatroot:blockM1981      -19124       4132  -4.629 4.38e-06 ***
## habitatroot:blockM1982      -21592       4113  -5.250 2.02e-07 ***
## habitatroot:blockM1992          NA         NA      NA       NA
## blockM1554:sampling2012      -7977       5328  -1.497 0.134790
## blockM1555:sampling2012     -15401       5283  -2.915 0.003668 **
```

```
## blockM1690:sampling2012    -42787      13434  -3.185 0.001512 **
## blockM1955:sampling2012    -17306       5383  -3.215 0.001363 **
## blockM1956:sampling2012    -20681       9212  -2.245 0.025080 *
## blockM1957:sampling2012        NA         NA      NA       NA
## blockM1958:sampling2012    -33567      11954  -2.808 0.005122 **
## blockM1959:sampling2012     -7076      11660  -0.607 0.544147
## blockM1960:sampling2012    -30761       7947  -3.871 0.000119 ***
## blockM1961:sampling2012    -37551      11611  -3.234 0.001278 **
## blockM1977:sampling2012    -12906       6030  -2.140 0.032690 *
## blockM1978:sampling2012    -14809       5401  -2.742 0.006260 **
## blockM1979:sampling2012     -9207       4780  -1.926 0.054482 .
## blockM1980:sampling2012    -15453       5350  -2.889 0.003989 **
## blockM1981:sampling2012    -16653       5064  -3.288 0.001058 **
## blockM1982:sampling2012    -15453       4952  -3.120 0.001879 **
## blockM1992:sampling2012        NA         NA      NA       NA
## habitatroot:sampling2012   -10427       3116  -3.346 0.000863 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10530 on 702 degrees of freedom
## Multiple R-squared:  0.5815, Adjusted R-squared:  0.5499
## F-statistic: 18.41 on 53 and 702 DF,  p-value: < 2.2e-16
```

7. How does this compare to the model you produced by heuristic search in (3) above?

```
anova(my_model4,forward_final,test="F")
```

```
## Analysis of Variance Table
##
## Model 1: CHAO1 ~ habitat + sampling + block + habitat:sampling
## Model 2: CHAO1 ~ age + habitat + block + sampling + age:habitat + habitat:block +
##     block:sampling + habitat:sampling
##   Res.Df        RSS Df  Sum of Sq     F    Pr(>F)
## 1    735 9.1211e+10
## 2    702 7.7878e+10 33 1.3333e+10 3.642 9.643e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

8. Now try the same thing, but this time start with a different model (any of your choosing). What do you notice? *The MAM may not be the same*
9. Perform a backwards search (`direction=backward`), starting with a more complex model containing interaction terms (e.g. in `*` combination). What do you notice? *ibid. see above*
10. Finally, perform a bidirectional search, starting with your model from (3) above (`direction=both`). How does this compare to your model?

```
final_final=step(lm(CHAO1~age*habitat),scope=c(lower=~habitat,upper=~ age* habitat * sampling * block *
```

```
## Start:  AIC=14269.89
## CHAO1 ~ age * habitat
##
##               Df  Sum of Sq        RSS   AIC
## + block       17 2.0875e+10 9.7010e+10 14156
## + site         4 1.0886e+10 1.0700e+11 14205
## + sampling     1 1.3059e+09 1.1658e+11 14264
## <none>                      1.1789e+11 14270
## - age:habitat  1 3.1850e+09 1.2107e+11 14288
##
## Step:  AIC=14156.55
## CHAO1 ~ age + habitat + block + age:habitat
##
##                 Df  Sum of Sq        RSS   AIC
## + site           4 1.4484e+10 8.2526e+10 14042
## + habitat:block 16 7.5887e+09 8.9421e+10 14127
```

```
## + sampling       1 1.6113e+09 9.5399e+10 14146
## <none>                        9.7010e+10 14156
## - age:habitat    1 2.6619e+08 9.7276e+10 14157
## + age:block     16 3.4641e+09 9.3546e+10 14161
## - block         17 2.0875e+10 1.1789e+11 14270
##
## Step:  AIC=14042.31
## CHAO1 ~ age + habitat + block + site + age:habitat
##
##                 Df  Sum of Sq        RSS   AIC
## + habitat:site   4 1.3328e+10 6.9198e+10 13917
## + habitat:block 16 6.7690e+09 7.5757e+10 14010
## + age:site       4 1.8930e+09 8.0633e+10 14033
## - age:habitat    1 8.3840e+06 8.2535e+10 14040
## + block:site    53 1.0958e+10 7.1568e+10 14041
## <none>                        8.2526e+10 14042
## + sampling       1 1.0294e+08 8.2423e+10 14043
## + age:block     16 2.4920e+09 8.0034e+10 14051
## - site           4 1.4484e+10 9.7010e+10 14156
## - block         17 2.4473e+10 1.0700e+11 14205
##
## Step:  AIC=13917.14
## CHAO1 ~ age + habitat + block + site + age:habitat + habitat:site
##
##                 Df  Sum of Sq        RSS   AIC
## + age:site       4 1.8830e+09 6.7315e+10 13904
## + habitat:block 16 3.7453e+09 6.5453e+10 13907
## - age:habitat    1 1.4667e+08 6.9345e+10 13917
## <none>                        6.9198e+10 13917
## + sampling       1 9.7752e+07 6.9101e+10 13918
## + age:block     16 1.4049e+09 6.7793e+10 13934
## + block:site    53 6.5482e+09 6.2650e+10 13948
## - habitat:site   4 1.3328e+10 8.2526e+10 14042
## - block         17 2.1461e+10 9.0660e+10 14087
##
## Step:  AIC=13904.29
## CHAO1 ~ age + habitat + block + site + age:habitat + habitat:site +
##     age:site
##
##                     Df  Sum of Sq        RSS   AIC
## + age:habitat:site   3 1.5144e+09 6.5801e+10 13893
## + habitat:block     16 3.5992e+09 6.3716e+10 13895
## - age:habitat        1 6.3186e+07 6.7379e+10 13903
## <none>                            6.7315e+10 13904
## + sampling           1 3.6522e+07 6.7279e+10 13906
## - age:site           4 1.8830e+09 6.9198e+10 13917
## + age:block         16 1.4422e+09 6.5873e+10 13920
## + block:site        53 5.6130e+09 6.1702e+10 13944
## - habitat:site       4 1.3318e+10 8.0633e+10 14033
## - block             17 1.8438e+10 8.5753e+10 14053
##
## Step:  AIC=13893.09
## CHAO1 ~ age + habitat + block + site + age:habitat + habitat:site +
##     age:site + age:habitat:site
##
##                     Df  Sum of Sq        RSS   AIC
## + habitat:block     16 3.5199e+09 6.2281e+10 13884
## <none>                            6.5801e+10 13893
## + sampling           1 1.8285e+07 6.5783e+10 13895
## - age:habitat:site   3 1.5144e+09 6.7315e+10 13904
## + age:block         16 1.3625e+09 6.4439e+10 13909
```

```
## + block:site       53 5.1902e+09 6.0611e+10 13937
## - block            17 1.8873e+10 8.4674e+10 14050
##
## Step:  AIC=13883.52
## CHAO1 ~ age + habitat + block + site + age:habitat + habitat:site +
##     age:site + habitat:block + age:habitat:site
##
##                     Df  Sum of Sq        RSS   AIC
## + sampling           1  300504765 6.1981e+10 13882
## <none>                             6.2281e+10 13884
## - habitat:block     16 3519867347 6.5801e+10 13893
## - age:habitat:site   3 1435129637 6.3716e+10 13895
## + age:block         16  746944444 6.1534e+10 13906
## + block:site        53 4114276577 5.8167e+10 13938
##
## Step:  AIC=13881.87
## CHAO1 ~ age + habitat + block + site + sampling + age:habitat +
##     habitat:site + age:site + habitat:block + age:habitat:site
##
##                     Df  Sum of Sq        RSS   AIC
## + sampling:site      4 1694850943 6.0286e+10 13869
## <none>                             6.1981e+10 13882
## + habitat:sampling   1  146904962 6.1834e+10 13882
## + age:sampling       1   69026864 6.1912e+10 13883
## - sampling           1  300504765 6.2281e+10 13884
## - habitat:block     16 3802087263 6.5783e+10 13895
## - age:habitat:site   3 1580795499 6.3561e+10 13895
## + sampling:block    15 1268874473 6.0712e+10 13896
## + age:block         16  736808789 6.1244e+10 13905
## + block:site        53 3983646356 5.7997e+10 13938
##
## Step:  AIC=13868.91
## CHAO1 ~ age + habitat + block + site + sampling + age:habitat +
##     habitat:site + age:site + habitat:block + site:sampling +
##     age:habitat:site
##
##                     Df  Sum of Sq        RSS   AIC
## + habitat:sampling   1  354357055 5.9931e+10 13866
## <none>                             6.0286e+10 13869
## + age:sampling       1   48613035 6.0237e+10 13870
## - age:habitat:site   3 1027451834 6.1313e+10 13876
## + sampling:block    15 1508676921 5.8777e+10 13880
## - habitat:block     16 3592320344 6.3878e+10 13881
## - site:sampling      4 1694850943 6.1981e+10 13882
## + age:block         16  623057055 5.9663e+10 13893
## + block:site        53 3733579753 5.6552e+10 13927
##
## Step:  AIC=13866.45
## CHAO1 ~ age + habitat + block + site + sampling + age:habitat +
##     habitat:site + age:site + habitat:block + site:sampling +
##     habitat:sampling + age:habitat:site
##
##                          Df  Sum of Sq        RSS   AIC
## + habitat:sampling:site   3 1574580116 5.8357e+10 13852
## <none>                                5.9931e+10 13866
## + age:sampling            1   36211505 5.9895e+10 13868
## - age:habitat:site        3  627240746 6.0559e+10 13868
## - habitat:sampling        1  354357055 6.0286e+10 13869
## + sampling:block         15 1651848042 5.8280e+10 13875
## - habitat:block          16 3430396072 6.3362e+10 13876
## - site:sampling           4 1902303036 6.1834e+10 13882
```

```
## + age:block             16  737366975 5.9194e+10 13889
## + block:site            53 3938548826 5.5993e+10 13921
##
## Step:  AIC=13852.32
## CHAO1 ~ age + habitat + block + site + sampling + age:habitat +
##     habitat:site + age:site + habitat:block + site:sampling +
##     habitat:sampling + age:habitat:site + habitat:site:sampling
##
##                          Df  Sum of Sq        RSS   AIC
## <none>                                  5.8357e+10 13852
## + age:sampling            1   18610970 5.8338e+10 13854
## - age:habitat:site        3 1149765524 5.9507e+10 13861
## - habitat:block          16 3350460211 6.1707e+10 13862
## + sampling:block         15 1266641959 5.7090e+10 13866
## - habitat:site:sampling   3 1574580116 5.9931e+10 13866
## + age:block              16  815076344 5.7542e+10 13874
## + block:site             53 3833784674 5.4523e+10 13907
```

```
anova(forward_final,backwards_final,final_final)
```

```
## Analysis of Variance Table
##
## Model 1: CHAO1 ~ age + habitat + block + sampling + age:habitat + habitat:block +
##     block:sampling + habitat:sampling
## Model 2: CHAO1 ~ habitat + sampling + block + site + habitat:sampling +
##     habitat:block + habitat:site + sampling:site + habitat:sampling:site
## Model 3: CHAO1 ~ age + habitat + block + site + sampling + age:habitat +
##     habitat:site + age:site + habitat:block + site:sampling +
##     habitat:sampling + age:habitat:site + habitat:site:sampling
##   Res.Df        RSS Df  Sum of Sq      F   Pr(>F)
## 1    702 7.7878e+10
## 2    704 6.0549e+10 -2 1.7329e+10
## 3    695 5.8357e+10  9 2.1922e+09 2.9009 0.002238 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

ibid.++; and the final final one seems best.

Bonus question if you finish early

Hopefully by now, you won't find it hard to work out which variables are continuous, and which are categorical. Do this now - take a piece of paper and write them out. However… there is something special about `block` and MiSeq run - and something else special about `age` and planting/sampling year. Can you think that they are? *Block, line and one are blocking factors, we may want to treat these differently somehow. Data are pooled potentially. Age, and sample/planting year are time-series/autocorrelated*

Super-bonus

It's Friday! Enjoy the weekend, you've earnt it.

Reference

Wagner MR, Lundberg DS, del Rio TG, Tringe SG, Dangl JL, Mitchell-Olds T (2016) Host genotype and age shape the leaf and root microbiomes of a wild perennial plant. *Nature Communications* **7**:12151. https://doi.org/10.1038/ncomms12151