ALSIA_StatisticalAnalysis

Elise

2024-06-09

Statistical modeling for phenotypic traits

The explanatory variable (X) refers to the genotype, which is a categorical variable. The response variables (Y) are the phenotypic data of the variables dataframe.

In this case, the variables are:

And the genotypes are:

```
print(variables)
```

```
## [1] "FW_shoot_g"      "Plant_height_cm"
```

```
unique(endpoint$Genotype)
```

```
## [1] EPPN1_L  EPPN2_L  EPPN3_L  EPPN4_L  EPPN1_H  EPPN2_H  EPPN3_H  EPPN4_H
## [9] EPPN20_T
## 9 Levels: EPPN1_H EPPN1_L EPPN2_H EPPN2_L EPPN20_T EPPN3_H EPPN3_L ... EPPN4_L
```

1. First linear models

Firstly, we model the Y = X + r + c + e Where - Y is the phenotypic trait; - X the genotype (fixed effect); - r the row effect (fixed or random); - c the column effect (fixed or random);

```
fit_models_fixed <- function(data, trait_name) {
  fixed_formula <- as.formula(paste(trait_name, "~ Genotype + Row + Column"))
  fixed_model <- lm(fixed_formula, data)
  print(paste("Summary for fixed effects model of", trait_name))
  print(summary(fixed_model))
  print(anova(fixed_model))
}

fit_models_random <- function(data, trait_name)  {
  random_formula <- as.formula(paste(trait_name, "~ Genotype + (1|Row) + (1|Column)"))
  random_model <- lmer(random_formula, data)
  print(paste("Summary for random effects model of", trait_name))
  print(summary(random_model))
  print(anova(random_model))
  print(ranova(random_model))
}

for (trait in variables) {
  fit_models_fixed(endpoint_clean, trait)
}
```

```
## [1] "Summary for fixed effects model of FW_shoot_g"
##
## Call:
## lm(formula = fixed_formula, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -63.077 -22.632   1.314  21.174  58.747
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)        73.720     25.909   2.845  0.00637 **
## GenotypeEPPN1_L    -1.975     19.785  -0.100  0.92086
## GenotypeEPPN2_H    31.321     19.834   1.579  0.12047
## GenotypeEPPN2_L    14.783     19.076   0.775  0.44195
## GenotypeEPPN20_T   11.576     19.722   0.587  0.55982
## GenotypeEPPN3_H    37.202     20.622   1.804  0.07714 .
## GenotypeEPPN3_L     6.103     19.788   0.308  0.75902
## GenotypeEPPN4_H    19.470     23.014   0.846  0.40150
## GenotypeEPPN4_L    16.976     20.505   0.828  0.41159
## Row2               -4.615     30.840  -0.150  0.88164
## Row3              -24.274     26.540  -0.915  0.36469
## Row4               -7.264     35.534  -0.204  0.83883
## Row5               -7.707     26.555  -0.290  0.77281
## Row6                3.880     29.488   0.132  0.89583
## Row7              -20.113     28.694  -0.701  0.48651
## Row8              -37.240     35.056  -1.062  0.29310
## Row9                4.535     27.915   0.162  0.87159
## Row10             -26.989     28.125  -0.960  0.34177
## Row11             -17.914     30.755  -0.582  0.56281
## Row12               6.955     28.669   0.243  0.80930
## Row13               6.923     25.498   0.271  0.78711
## Row14              13.495     28.838   0.468  0.64182
## Row15              14.709     30.149   0.488  0.62774
## Row16             -13.665     30.655  -0.446  0.65764
## Row17             -23.566     29.765  -0.792  0.43218
## Row18              -6.173     28.172  -0.219  0.82743
## Row19             -10.481     24.808  -0.422  0.67445
## Row20             -55.875     27.010  -2.069  0.04366 *
## Row21              35.602     34.083   1.045  0.30114
## Row22             -28.280     27.940  -1.012  0.31623
## Row23             -14.303     28.701  -0.498  0.62037
## Row24             -16.645     31.276  -0.532  0.59690
## Row25             -41.022     29.946  -1.370  0.17673
## Row26             -14.696     29.409  -0.500  0.61944
## Column15            8.928     14.055   0.635  0.52815
## Column16            5.219     13.781   0.379  0.70647
## Column17            5.602     13.130   0.427  0.67141
## Column18           -6.826     13.497  -0.506  0.61523
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 37.91 on 51 degrees of freedom
##   (1 observation effacée parce que manquante)
## Multiple R-squared:  0.3465, Adjusted R-squared:  -0.1277
```

```
## F-statistic: 0.7307 on 37 and 51 DF,  p-value: 0.8401
##
## Analysis of Variance Table
##
## Response: FW_shoot_g
##           Df Sum Sq Mean Sq F value Pr(>F)
## Genotype   8  12268 1533.50  1.0671 0.4006
## Row       25  24227  969.08  0.6744 0.8568
## Column     4   2356  589.07  0.4099 0.8007
## Residuals 51  73290 1437.05
## [1] "Summary for fixed effects model of Plant_height_cm"
##
## Call:
## lm(formula = fixed_formula, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -37.440 -10.337  -1.093  11.616  33.871
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      123.4694    14.6346   8.437 6.77e-11 ***
## GenotypeEPPN1_L    1.5682    11.1554   0.141   0.8888
## GenotypeEPPN2_H   12.4396    11.1169   1.119   0.2690
## GenotypeEPPN2_L    4.5911    10.6976   0.429   0.6698
## GenotypeEPPN20_T  -0.2525    10.9218  -0.023   0.9817
## GenotypeEPPN3_H   18.4233    11.8906   1.549   0.1281
## GenotypeEPPN3_L   -3.7525    11.0177  -0.341   0.7350
## GenotypeEPPN4_H   -3.2318    13.4447  -0.240   0.8111
## GenotypeEPPN4_L    2.9740    11.3167   0.263   0.7939
## Row2              -7.1041    17.2496  -0.412   0.6824
## Row3             -10.2809    14.6069  -0.704   0.4851
## Row4              -7.7104    25.0795  -0.307   0.7599
## Row5              -7.4500    14.7832  -0.504   0.6167
## Row6              -7.6719    16.3898  -0.468   0.6419
## Row7              -8.5612    15.7745  -0.543   0.5899
## Row8             -24.8878    19.5494  -1.273   0.2094
## Row9               3.5684    15.4867   0.230   0.8188
## Row10            -23.6697    15.7574  -1.502   0.1399
## Row11            -16.7567    17.1780  -0.975   0.3344
## Row12             -1.3952    15.7561  -0.089   0.9298
## Row13              7.1836    15.5991   0.461   0.6473
## Row14             -6.4771    15.8941  -0.408   0.6855
## Row15              3.3166    16.6924   0.199   0.8434
## Row16             -4.5917    17.1077  -0.268   0.7896
## Row17            -14.3931    16.4185  -0.877   0.3852
## Row18              6.7502    17.4629   0.387   0.7009
## Row19            -16.1112    13.6801  -1.178   0.2450
## Row20            -30.8323    14.9026  -2.069   0.0442 *
## Row21             12.5159    18.8870   0.663   0.5108
## Row22            -10.5671    15.4548  -0.684   0.4976
## Row23            -10.1168    15.7491  -0.642   0.5238
## Row24            -21.0547    17.4764  -1.205   0.2345
## Row25            -41.9089    25.6357  -1.635   0.1089
## Row26            -14.2372    16.2096  -0.878   0.3843
## Column15          -7.5272     8.4164  -0.894   0.3758
```

```
## Column16          -1.6026     7.8591  -0.204   0.8393
## Column17          -1.1364     7.6101  -0.149   0.8819
## Column18          -8.2376     7.9085  -1.042   0.3030
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20.78 on 46 degrees of freedom
##   (6 observations effacées parce que manquantes)
## Multiple R-squared:  0.3835, Adjusted R-squared:  -0.1123
## F-statistic: 0.7735 on 37 and 46 DF,  p-value: 0.7885
##
## Analysis of Variance Table
##
## Response: Plant_height_cm
##           Df  Sum Sq Mean Sq F value Pr(>F)
## Genotype   8  4055.0  506.87  1.1739 0.3352
## Row       25  7515.3  300.61  0.6962 0.8335
## Column     4   787.7  196.91  0.4560 0.7675
## Residuals 46 19862.3  431.79
```

```
for (trait in variables) {
  fit_models_random(endpoint_clean, trait)
}
```

```
## boundary (singular) fit: see help('isSingular')
```

```
## [1] "Summary for random effects model of FW_shoot_g"
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: random_formula
##    Data: data
##
## REML criterion at convergence: 818
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -1.8252 -0.8810  0.1228  0.7135  1.9163
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  Row      (Intercept)    0      0.00
##  Column   (Intercept)    0      0.00
##  Residual             1248     35.33
## Number of obs: 89, groups:  Row, 26; Column, 5
##
## Fixed effects:
##                  Estimate Std. Error    df t value Pr(>|t|)
## (Intercept)         62.35      11.17 80.00   5.580 3.17e-07 ***
## GenotypeEPPN1_L      7.64      15.80 80.00   0.484    0.630
## GenotypeEPPN2_H     19.64      15.80 80.00   1.243    0.218
## GenotypeEPPN2_L     21.84      15.80 80.00   1.382    0.171
## GenotypeEPPN20_T    18.36      15.80 80.00   1.162    0.249
## GenotypeEPPN3_H     46.87      16.23 80.00   2.887    0.005 **
## GenotypeEPPN3_L     13.88      15.80 80.00   0.878    0.382
## GenotypeEPPN4_H     14.46      15.80 80.00   0.915    0.363
## GenotypeEPPN4_L     16.81      15.80 80.00   1.064    0.291
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr) GEPPN1 GEPPN2_H GEPPN2_L GEPPN20 GEPPN3_H GEPPN3_L GEPPN4_H
## GntyEPPN1_L -0.707
## GntyEPPN2_H -0.707  0.500
## GntyEPPN2_L -0.707  0.500  0.500
## GntEPPN20_T -0.707  0.500  0.500    0.500
## GntyEPPN3_H -0.688  0.487  0.487    0.487    0.487
## GntyEPPN3_L -0.707  0.500  0.500    0.500    0.500   0.487
## GntyEPPN4_H -0.707  0.500  0.500    0.500    0.500   0.487    0.500
## GntyEPPN4_L -0.707  0.500  0.500    0.500    0.500   0.487    0.500    0.500
## optimizer (nloptwrap) convergence code: 0 (OK)
## boundary (singular) fit: see help('isSingular')
##
## Type III Analysis of Variance Table with Satterthwaite's method
##          Sum Sq Mean Sq NumDF DenDF F value Pr(>F)
## Genotype  12268  1533.5     8    80  1.2284 0.2935
```

```
## boundary (singular) fit: see help('isSingular')
## boundary (singular) fit: see help('isSingular')
```

```
## ANOVA-like table for random-effects: Single term deletions
##
## Model:
## FW_shoot_g ~ Genotype + (1 | Row) + (1 | Column)
##              npar  logLik    AIC LRT Df Pr(>Chisq)
## <none>         12 -409.01 842.02
## (1 | Row)      11 -409.01 840.02   0  1          1
## (1 | Column)   11 -409.01 840.02   0  1          1
```

```
## boundary (singular) fit: see help('isSingular')
```

```
## [1] "Summary for random effects model of Plant_height_cm"
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: random_formula
##    Data: data
##
## REML criterion at convergence: 677.5
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -1.6926 -0.6683 -0.1045  0.6599  1.8319
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  Row      (Intercept)   0.0     0.00
##  Column   (Intercept)   0.0     0.00
##  Residual             375.5    19.38
## Number of obs: 84, groups:  Row, 26; Column, 5
##
## Fixed effects:
##                  Estimate Std. Error      df t value Pr(>|t|)
## (Intercept)       105.500      6.128  75.000  17.216  < 2e-16 ***
## GenotypeEPPN1_L     9.944      8.904  75.000   1.117  0.26762
## GenotypeEPPN2_H     9.500      8.666  75.000   1.096  0.27651
## GenotypeEPPN2_L    13.389      8.904  75.000   1.504  0.13686
## GenotypeEPPN20_T    7.300      8.666  75.000   0.842  0.40228
## GenotypeEPPN3_H    26.625      9.192  75.000   2.896  0.00494 **
## GenotypeEPPN3_L     7.300      8.666  75.000   0.842  0.40228
## GenotypeEPPN4_H     1.750      9.192  75.000   0.190  0.84953
## GenotypeEPPN4_L     5.800      8.666  75.000   0.669  0.50539
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr) GEPPN1 GEPPN2_H GEPPN2_L GEPPN20 GEPPN3_H GEPPN3_L GEPPN4_H
## GntyEPPN1_L -0.688
## GntyEPPN2_H -0.707  0.487
## GntyEPPN2_L -0.688  0.474  0.487
## GntEPPN20_T -0.707  0.487  0.500    0.487
## GntyEPPN3_H -0.667  0.459  0.471    0.459    0.471
## GntyEPPN3_L -0.707  0.487  0.500    0.487    0.500   0.471
## GntyEPPN4_H -0.667  0.459  0.471    0.459    0.471   0.444    0.471
## GntyEPPN4_L -0.707  0.487  0.500    0.487    0.500   0.471    0.500    0.471
## optimizer (nloptwrap) convergence code: 0 (OK)
## boundary (singular) fit: see help('isSingular')
##
## Type III Analysis of Variance Table with Satterthwaite's method
##          Sum Sq Mean Sq NumDF DenDF F value Pr(>F)
## Genotype   4055  506.87     8    75  1.3497 0.2328
```

```
## boundary (singular) fit: see help('isSingular')
## boundary (singular) fit: see help('isSingular')
```

```
## ANOVA-like table for random-effects: Single term deletions
##
## Model:
## Plant_height_cm ~ Genotype + (1 | Row) + (1 | Column)
##              npar  logLik    AIC LRT Df Pr(>Chisq)
## <none>         12 -338.77 701.53
## (1 | Row)      11 -338.77 699.53   0  1          1
## (1 | Column)   11 -338.77 699.53   0  1          1
```

2. Linear models with Plant_type

Model with X as Plant_type instead of Genotype, and row and column effects as random effects. Plant_type is defined as H for Hybrid, L for pure Line and T for Tester.

```
endpoint_clean$Plant_type <- as.factor(endpoint_clean$Plant_type)
endpoint_clean$Plant_type <- relevel(endpoint_clean$Plant_type, ref = "T") # T as base level

fit_model_plant_type <- function(data, trait) {
  # Random effects model with Plant_type as a fixed effect
  model_formula <- as.formula(paste(trait, "~ Plant_type + (1|Row) + (1|Column)"))
  model <- lmer(model_formula, data)
  print(paste("Summary for mixed effects model of", trait))
  print(summary(model))
  print(anova(model))
  print(ranova(model))
}

for (trait in variables) {
  fit_model_plant_type(endpoint_clean, trait)
}
```

```
## boundary (singular) fit: see help('isSingular')
```

```
## [1] "Summary for mixed effects model of FW_shoot_g"
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: model_formula
##    Data: data
##
## REML criterion at convergence: 870.3
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -1.7593 -0.9242  0.1694  0.6961  2.1970
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  Row      (Intercept)    0      0.00
##  Column   (Intercept)    0      0.00
##  Residual             1299     36.04
## Number of obs: 89, groups:  Row, 26; Column, 5
##
## Fixed effects:
##             Estimate Std. Error     df t value Pr(>|t|)
## (Intercept)   80.710     11.398 86.000   7.081 3.67e-10 ***
## Plant_typeH    1.200     12.776 86.000   0.094    0.925
## Plant_typeL   -3.318     12.743 86.000  -0.260    0.795
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr) Plnt_H
## Plant_typeH -0.892
## Plant_typeL -0.894  0.798
## optimizer (nloptwrap) convergence code: 0 (OK)
## boundary (singular) fit: see help('isSingular')
##
## Type III Analysis of Variance Table with Satterthwaite's method
##            Sum Sq Mean Sq NumDF DenDF F value Pr(>F)
## Plant_type 413.53  206.76     2    86  0.1592 0.8531
```

```
## boundary (singular) fit: see help('isSingular')
## boundary (singular) fit: see help('isSingular')
```

```
## ANOVA-like table for random-effects: Single term deletions
##
## Model:
## FW_shoot_g ~ Plant_type + (1 | Row) + (1 | Column)
##              npar  logLik    AIC LRT Df Pr(>Chisq)
## <none>          6 -435.14 882.29
## (1 | Row)       5 -435.14 880.29   0  1          1
## (1 | Column)    5 -435.14 880.29   0  1          1
```

```
## boundary (singular) fit: see help('isSingular')
```

```
## [1] "Summary for mixed effects model of Plant_height_cm"
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: model_formula
##    Data: data
##
## REML criterion at convergence: 724.2
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.0286 -0.7750  0.0264  0.7788  2.0844
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  Row      (Intercept)   0.0     0.00
##  Column   (Intercept)   0.0     0.00
##  Residual             397.5    19.94
## Number of obs: 84, groups:  Row, 26; Column, 5
##
## Fixed effects:
##             Estimate Std. Error      df t value Pr(>|t|)
## (Intercept)  112.800      6.305  81.000  17.892   <2e-16 ***
## Plant_typeH    1.644      7.127  81.000   0.231    0.818
## Plant_typeL    1.674      7.086  81.000   0.236    0.814
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr) Plnt_H
## Plant_typeH -0.885
## Plant_typeL -0.890  0.787
## optimizer (nloptwrap) convergence code: 0 (OK)
## boundary (singular) fit: see help('isSingular')
##
## Type III Analysis of Variance Table with Satterthwaite's method
##            Sum Sq Mean Sq NumDF DenDF F value Pr(>F)
## Plant_type 24.276  12.138     2    81  0.0305 0.9699
```

```
## boundary (singular) fit: see help('isSingular')
## boundary (singular) fit: see help('isSingular')
```

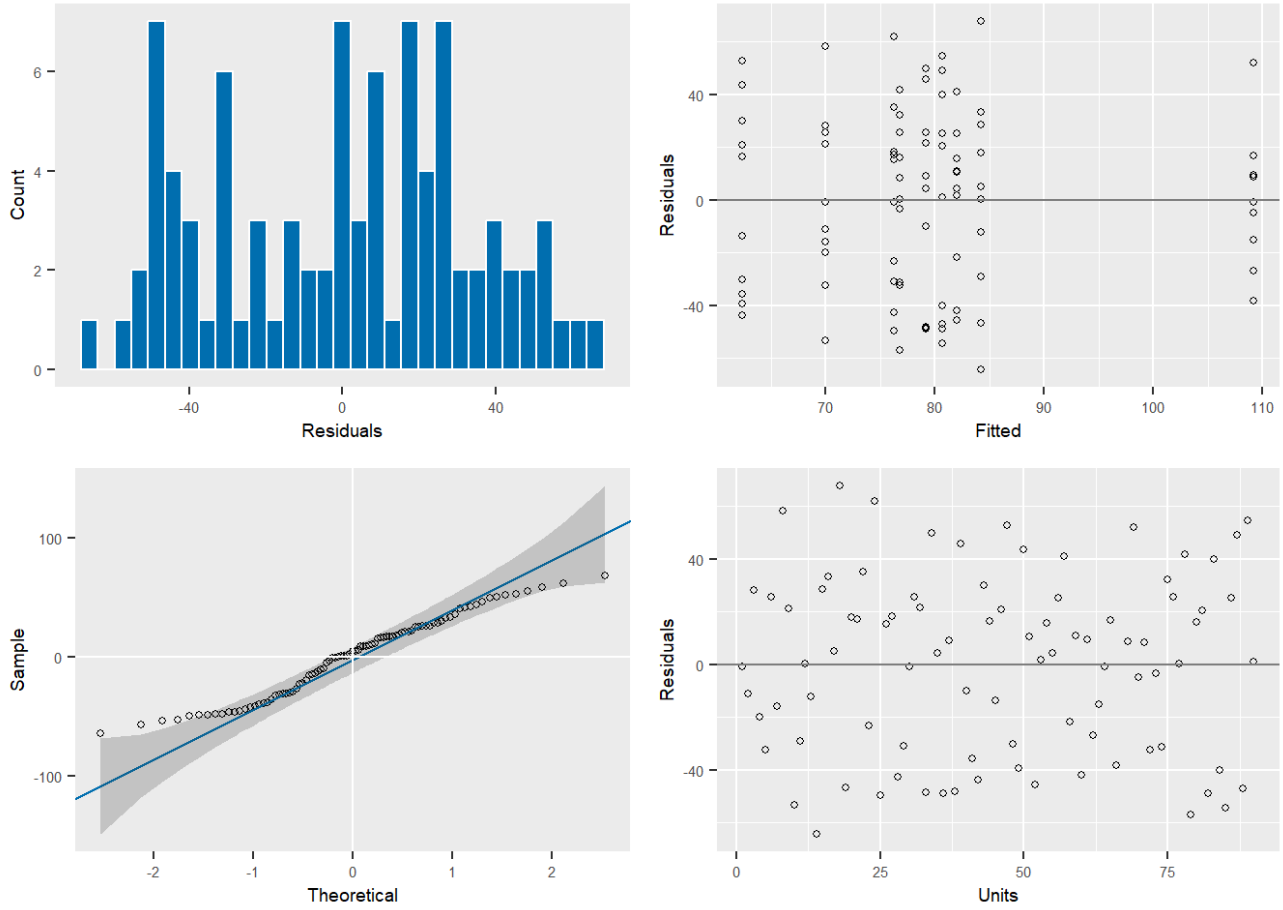
```
## ANOVA-like table for random-effects: Single term deletions
##
## Model:
## Plant_height_cm ~ Plant_type + (1 | Row) + (1 | Column)
##              npar  logLik    AIC LRT Df Pr(>Chisq)
## <none>          6 -362.09 736.19
## (1 | Row)       5 -362.09 734.19   0  1          1
## (1 | Column)    5 -362.09 734.19   0  1          1
```

3. Linear models with asreml library

```
modasreml <- asreml(fixed = FW_shoot_g ~ Genotype,
                    random = ~ Row + Column,
                    residual = ~  NULL,
                    data = endpoint_clean)
```

```
## ASReml Version 4.2 09/06/2024 16:51:23
##           LogLik        Sigma2     DF     wall
##  1     -338.0533      1163.695     80   16:51:23  (  2 restrained)
##  2     -335.6864      1240.189     80   16:51:23  (  2 restrained)
##  3     -335.5063      1247.866     80   16:51:23  (  2 restrained)
##  4     -335.4948      1248.376     80   16:51:23  (  2 restrained)
##  5     -335.4941      1248.408     80   16:51:23  (  2 restrained)
```

```
plot(modasreml)
```




```
summary(modasreml)$varcomp
```

```
##             component std.error z.ratio bound %ch
## Column  1.263301e-04        NA      NA     B  NA
## Row     1.263301e-04        NA      NA     B  NA
## units!R 1.248408e+03  226.5693 5.51005     P   0
```

4. Linear models with Soil variable

Model with Soil as explicative variable.

5. Linear models with Soil variable with Plant_type