Model-based multivariate analyses ‘mvabund’ package

Natasha Karenyi

SEEC - Statistics in Ecology, Environment and Conservation

University of Cape Town • iYunivesithi yaseKapa • Universiteit van Kaapstad

SPES BONA

Happy Halloween

Introduction

- When do we use multivariate analyses?
  - Multiple response variables

Data types

- Species
  - Presence/absence, ordinal, count, biomass, percentage cover
- Environmental
  - Geological, oceanographic, climate
- Morphological/Traits
  - Size or shape measurements, life history traits, sex, etc.
- Molecular data

Introduction

- When do we use multivariate analyses?
  - Multiple response variables
- Broad types of analyses:
  - association-based: http://www.seec.uct.ac.za/introduction-multivariate-analyses
  - model-based

‘mvabund’ package in R

- Statistical methods for analysing multivariate abundance data
- Used for **statistical inference**, not exploratory analyses
  - the theory, methods, and practice of forming judgements about the parameters of a population and the reliability of statistical relationships, typically on the basis of random sampling. http://www.seec.uct.ac.za/experimental-and-survey-design

Data types

- Species
  - Presence/absence, ordinal, count, biomass, ~~percentage cover~~
- Environmental
  - Geological, oceanographic, climate
- Morphological/Traits
  - Size or shape measurements, life history traits, sex, etc.
- Molecular data?

Characteristics of multivariate data

- **Multivariate** – many correlated response variables and often more variables than observations
- **Abundance** – abundance or presence/absence data usually has a strong mean-variance relationship

The `manyglm` function

- The `manyglm` function is designed for mv abundance data. It deals with key data properties:
- **Multivariate:** It uses row (site) resampling for inference and to preserve the correlation between variables (species).
- **Abundance:** manyglm fits a separate GLM to each species. (`family` & `plot`).

What do we want to know?

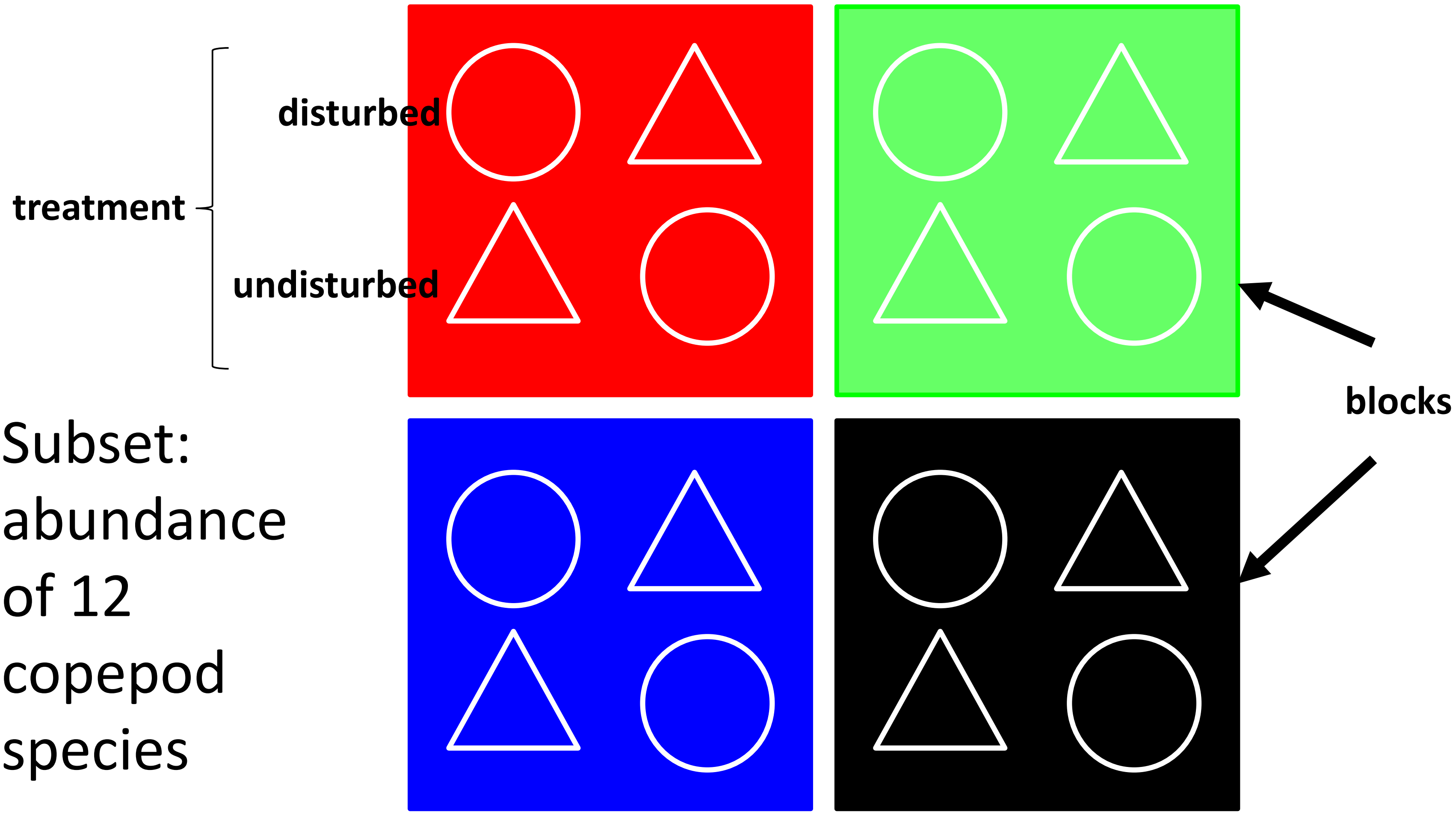
- Does treatment have an effect on assemblage?
  - discrete explanatory variables
- What are the indicator species?

Data required for `manyglm`

- List of 2 data.frames
  - Species abundance data per site
  - Treatment or Environmental data per site

- Convert abundance to `mvabund` object
- Treatment or environmental variables as vector or data.frame

Tasmania copepod data



treatment

disturbed

undisturbed

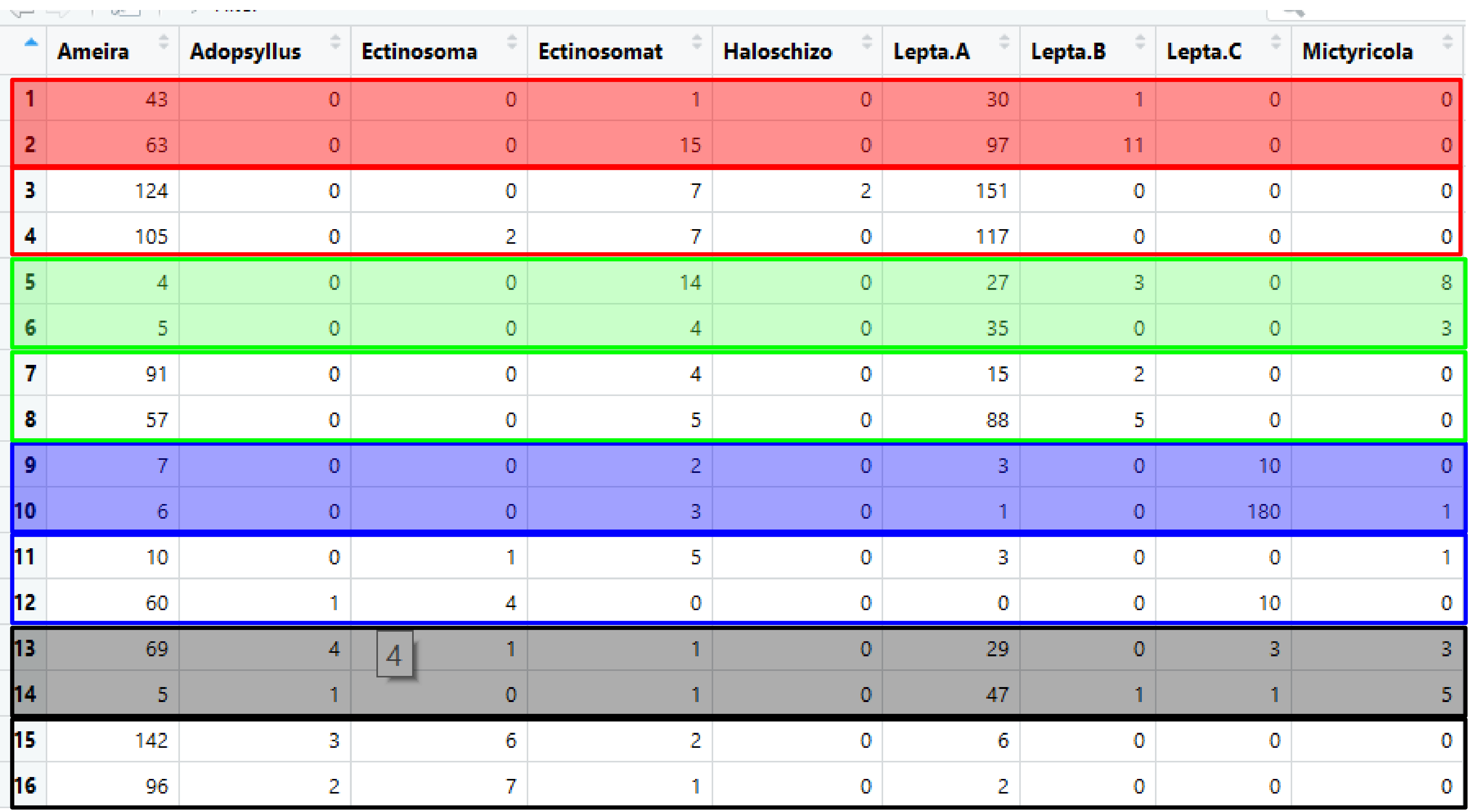
blocks

Subset: abundance of 12 copepod species

To visualise Tasmania copepod data

```
> data(Tasmania)
> tasm.cop <- mvabund(Tasmania$copepods)
> treatment <- Tasmania$treatment
> block <- Tasmania$block
> plot(tasm.cop ~ treatment, col=as.numeric(block))
```

All code taken directly from mvabund package reference manual



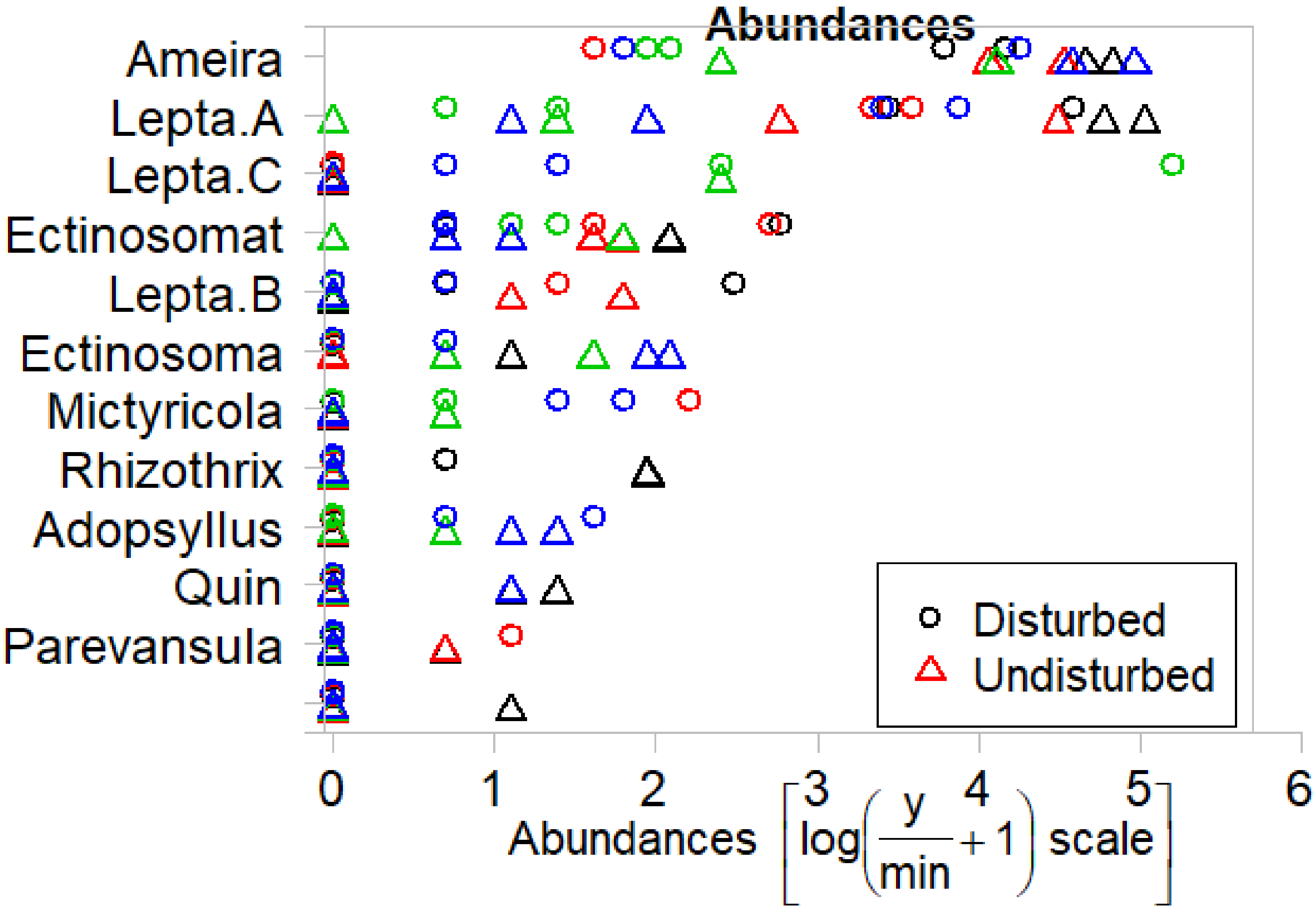
	Ameira	Adopsyllus	Ectinosoma	Ectinosomat	Haloschizo	Lepta.A	Lepta.B	Lepta.C	Mictyricola
1	43	0	0	1	0	30	1	0	0
2	63	0	0	15	0	97	11	0	0
3	124	0	0	7	2	151	0	0	0
4	105	0	2	7	0	117	0	0	0
5	4	0	0	14	0	27	3	0	8
6	5	0	0	4	0	35	0	0	3
7	91	0	0	4	0	15	2	0	0
8	57	0	0	5	0	88	5	0	0
9	7	0	0	2	0	3	0	10	0
10	6	0	0	3	0	1	0	180	1
11	10	0	1	5	0	3	0	0	1
12	60	1	4	0	0	0	0	10	0
13	69	4	1	1	0	29	0	3	3
14	5	1	0	1	0	47	1	1	5
15	142	3	6	2	0	6	0	0	0
16	96	2	7	1	0	2	0	0	0

To visualise Tasmania copepod data

```
> data(Tasmania)
> tasm.cop <- mvabund(Tasmania$copepods)
> treatment <- Tasmania$treatment
> block <- Tasmania$block
> plot(tasm.cop ~ treatment, col=as.numeric(block))
```

All code taken directly from mvabund package reference manual

Visualising Tasmania copepod data



Structure of GLM's

Specify 3 components of a GLM:

1. Random part
   - Prob distbn for response variable & defines mean-variance relationship
2. Systematic part
   - Form of explanatory variables = linear predictor
3. Link function
   - Links random and systematic parts; mean response related to explanatory variables

Assumptions of manyglm

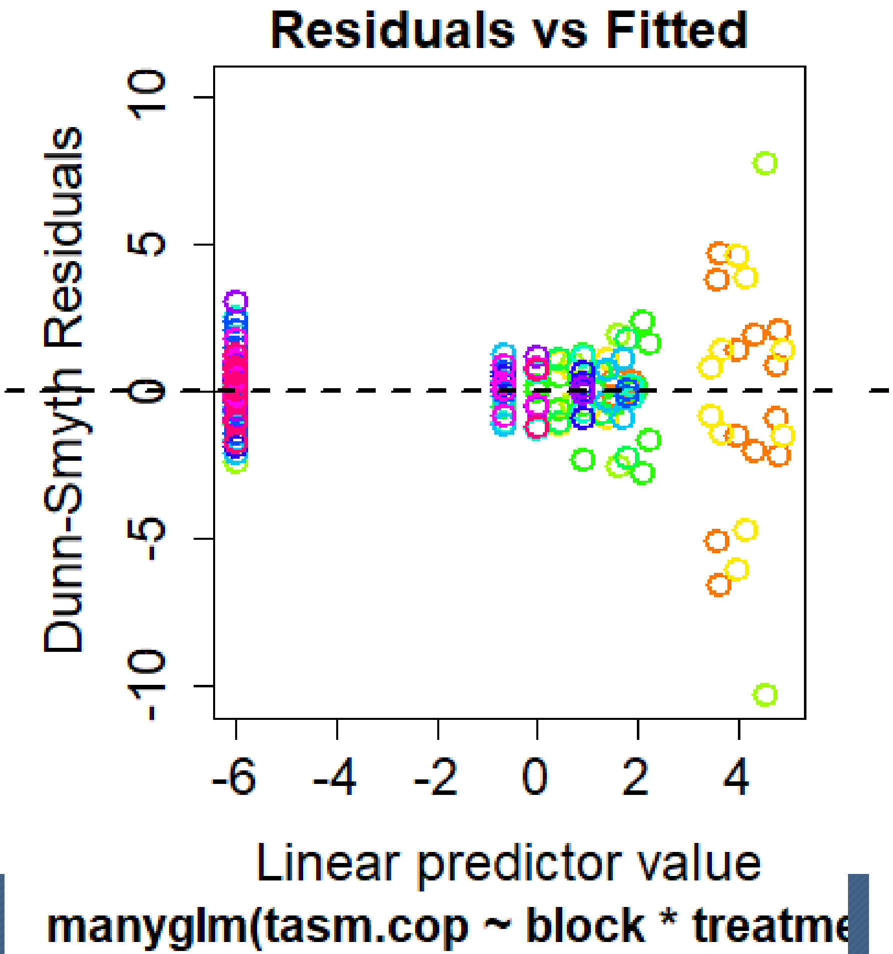
- Observed y values are independent, after conditioning on x
- Y-values come from a known distribution with known mean-variance relationship
- Straight line relationship between some known function of the mean of y and each x
- Residuals have a constant correlation matrix across observations

`manyglm` applied to Tasmanian copepods

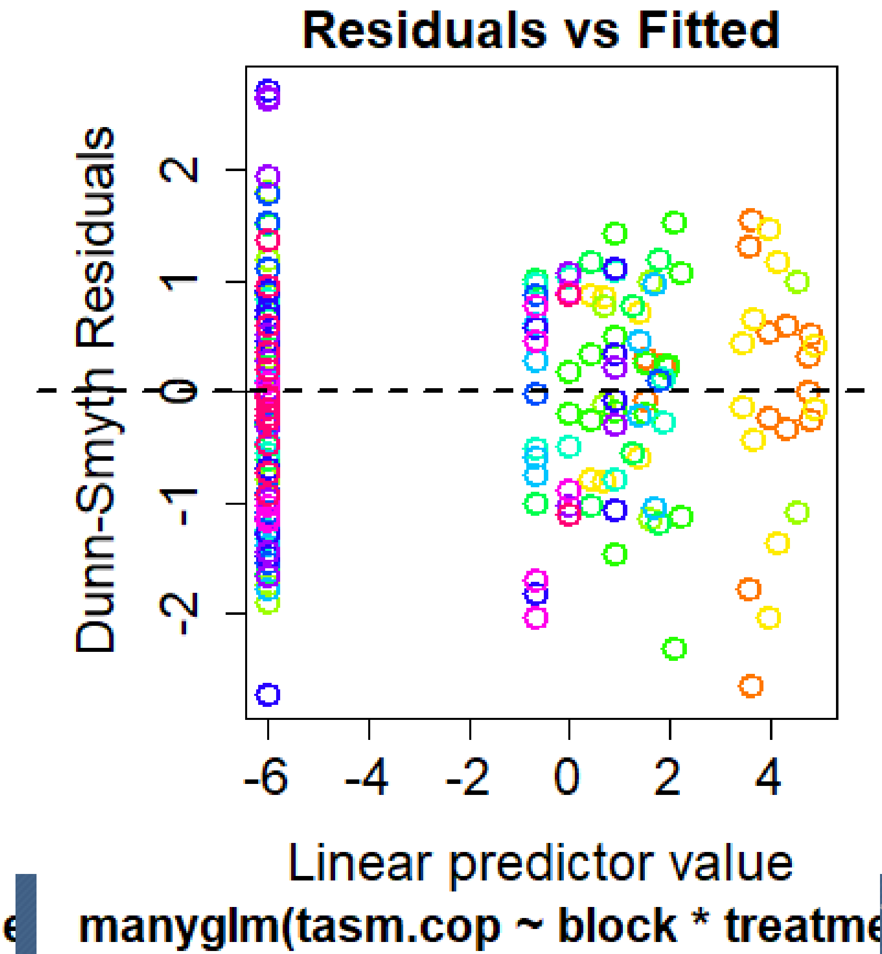
```
tasm.cop.nb <-
    manyglm(tasm.cop ~ block*treatment,
    family="negative.binomial")

tasm.cop.pois<-
    manyglm(tasm.cop ~ block*treatment,
    family=“poisson")
```

`manyglm` applied to Tasmanian copepods

Poisson

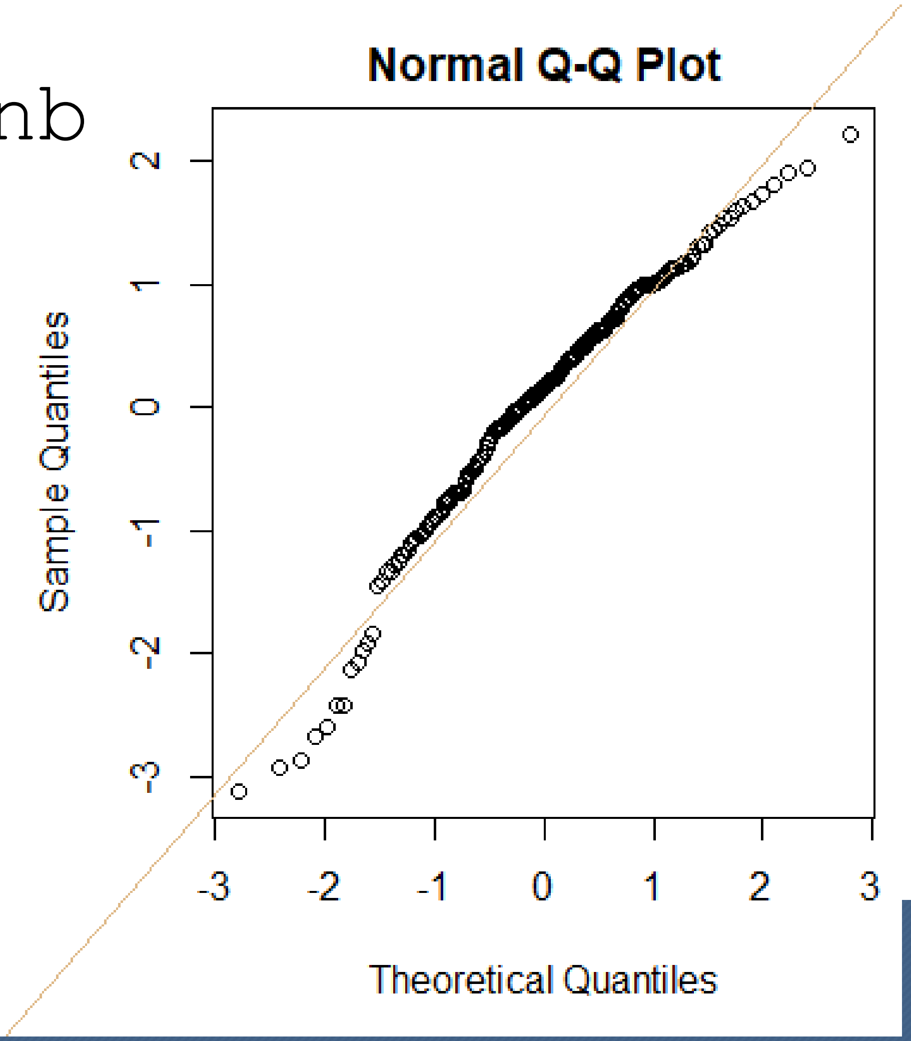


Negative binomial

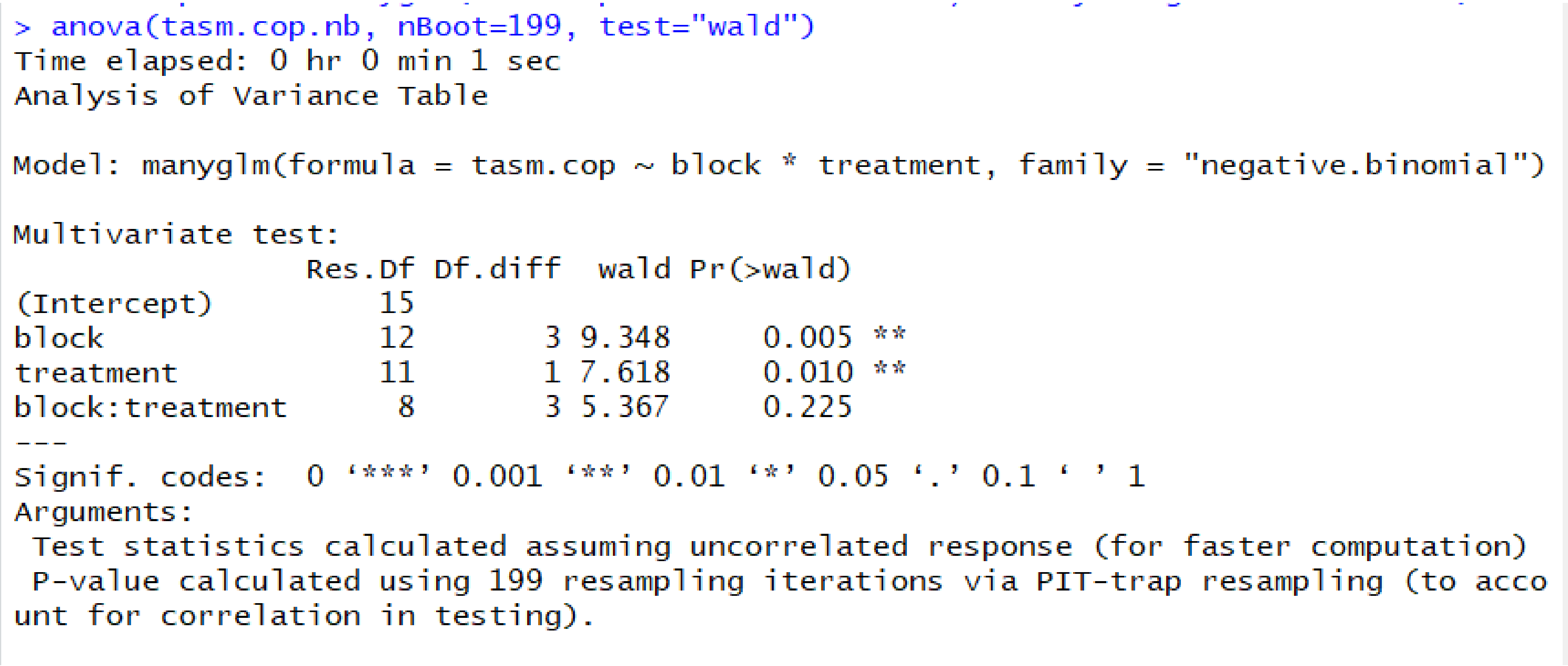


`manyglm` applied to Tasmanian copepods

`tasm.cop.nb`

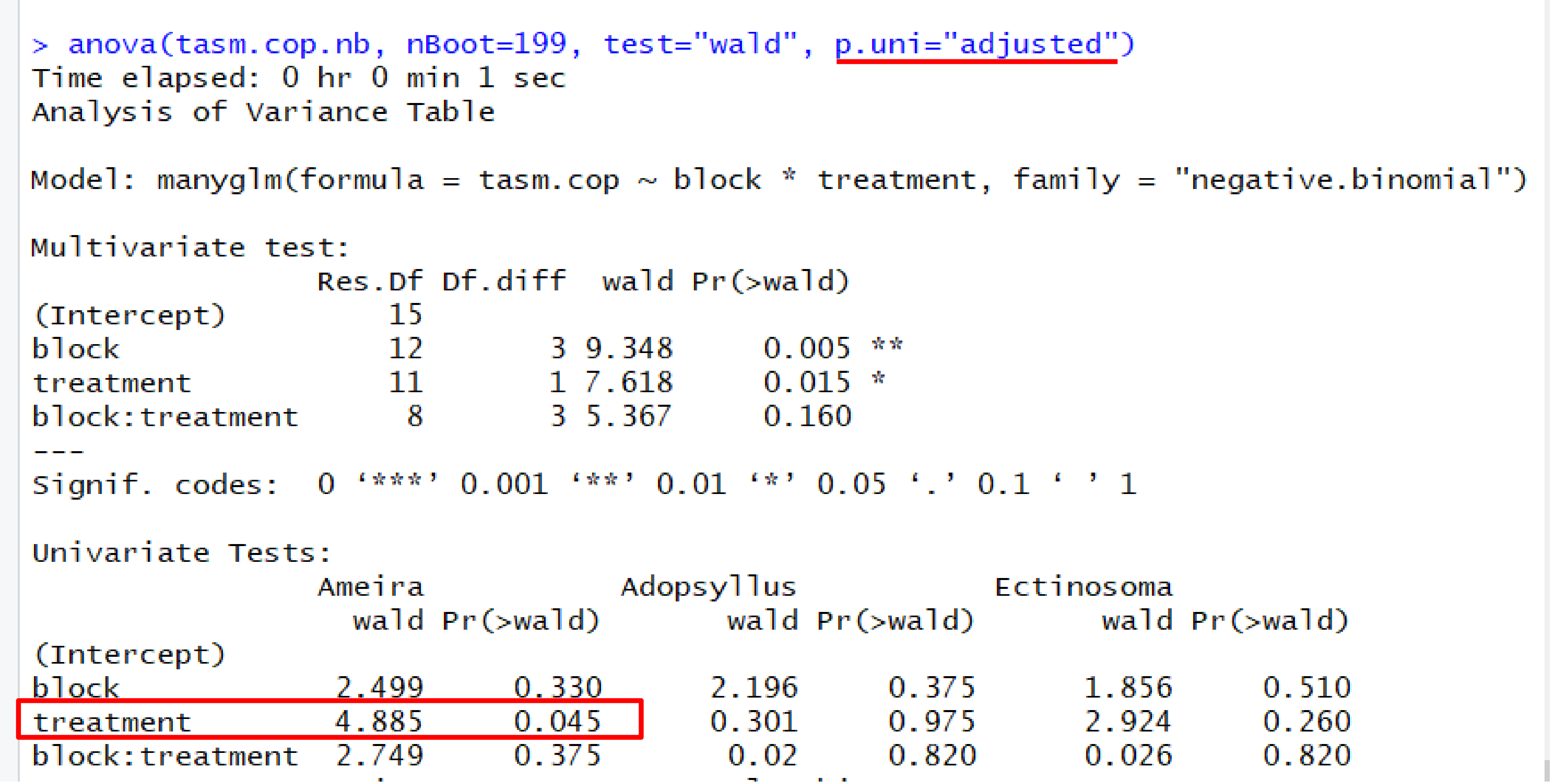


anova

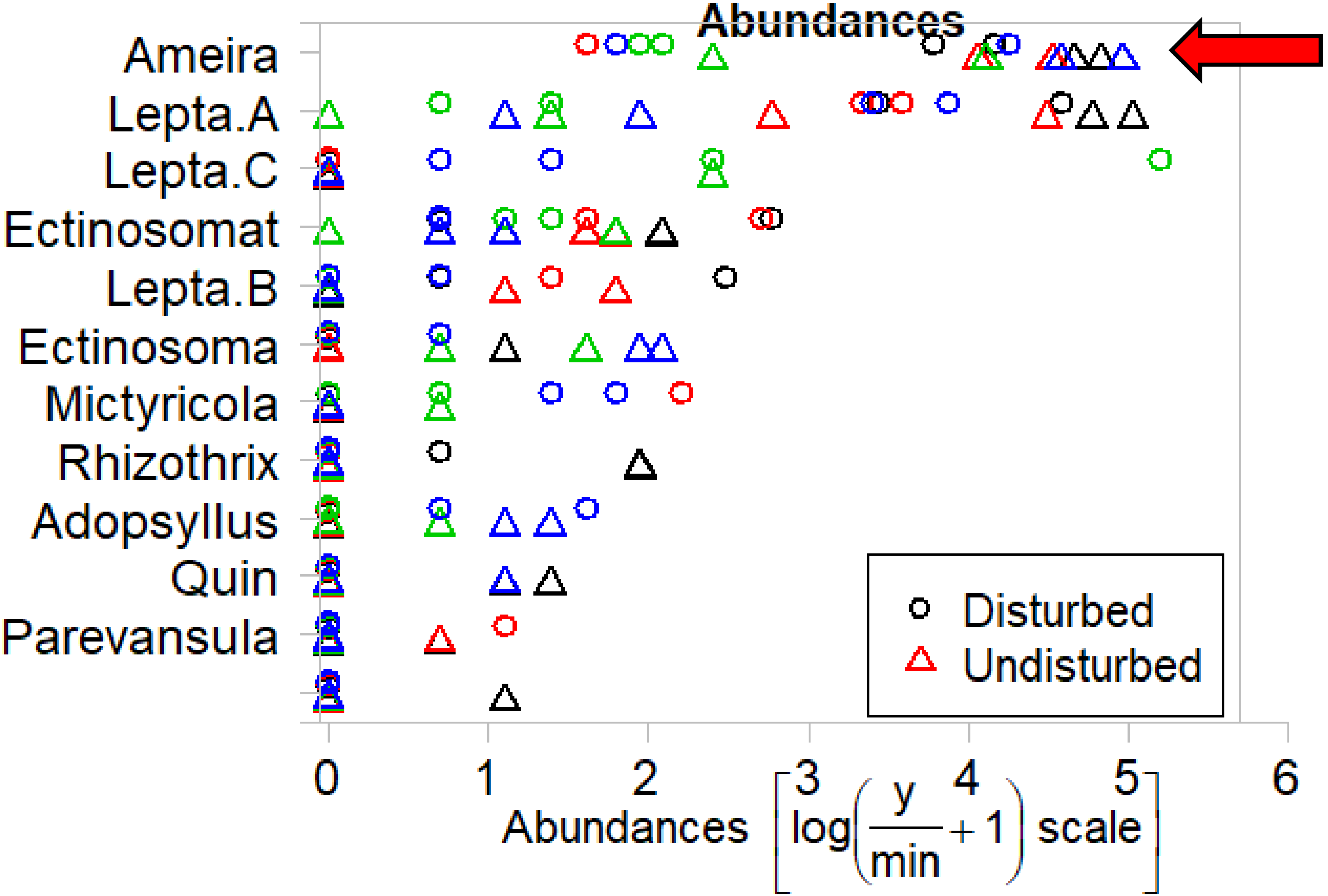


```
> anova(tasm.cop.nb, nBoot=199, test="wald")
Time elapsed: 0 hr 0 min 1 sec
Analysis of Variance Table

Model: manyglm(formula = tasm.cop ~ block * treatment, family = "negative.binomial")

Multivariate test:
                Res.Df Df.diff  wald Pr(>wald)
(Intercept)         15
block               12       3 9.348     0.005 **
treatment           11       1 7.618     0.010 **
block:treatment      8       3 5.367     0.225
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Arguments:
 Test statistics calculated assuming uncorrelated response (for faster computation)
 P-value calculated using 199 resampling iterations via PIT-trap resampling (to acco
unt for correlation in testing).
```

In which species is there an effect?



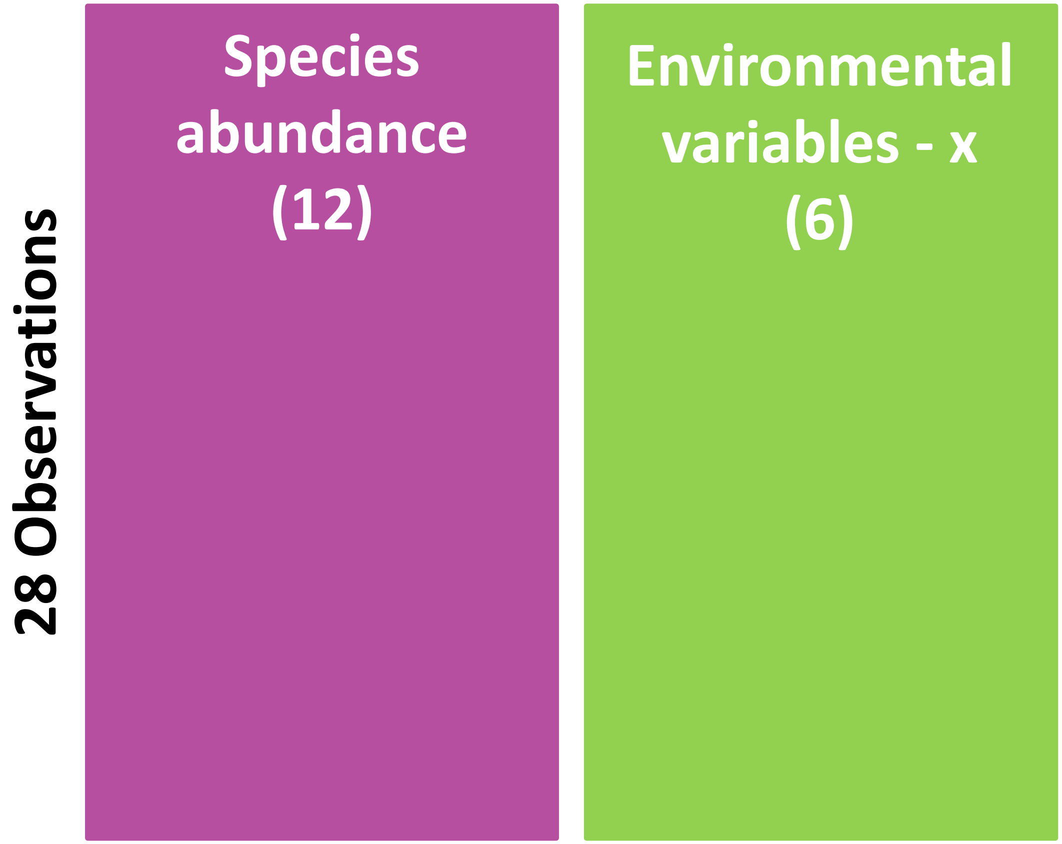
```
> anova(tasm.cop.nb, nBoot=199, test="wald", p.uni="adjusted")
Time elapsed: 0 hr 0 min 1 sec
Analysis of Variance Table

Model: manyglm(formula = tasm.cop ~ block * treatment, family = "negative.binomial")

Multivariate test:
                Res.Df Df.diff  wald Pr(>wald)
(Intercept)         15
block               12       3 9.348     0.005 **
treatment           11       1 7.618     0.015 *
block:treatment      8       3 5.367     0.160
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Univariate Tests:
                Ameira           Adopsyllus           Ectinosoma
                  wald Pr(>wald)       wald Pr(>wald)       wald Pr(>wald)
(Intercept)
block            2.499     0.330      2.196     0.375      1.856     0.510
treatment        4.885     0.045      0.301     0.975      2.924     0.260
block:treatment  2.749     0.375       0.02     0.820      0.026     0.820
```


Abundances
Ameira
Lepta.A
Lepta.C
Ectinosomat
Lepta.B
Ectinosoma
Mictyricola
Rhizothrix
Adopsyllus
Quin
Parevansula
○ Disturbed
△ Undisturbed
0
1
2
3
4
5
6
Abundances $\left[\log\left(\frac{y}{\min}+1\right)\text{ scale}\right]$

What do we want to know?

- Does treatment have an effect on assemblage? ✓
  - discrete explanatory variables
- What are the indicator species? ✓
- Which environmental variables are most strongly associated with an assemblage?
  - Continuous explanatory variables

Hunting Spider data

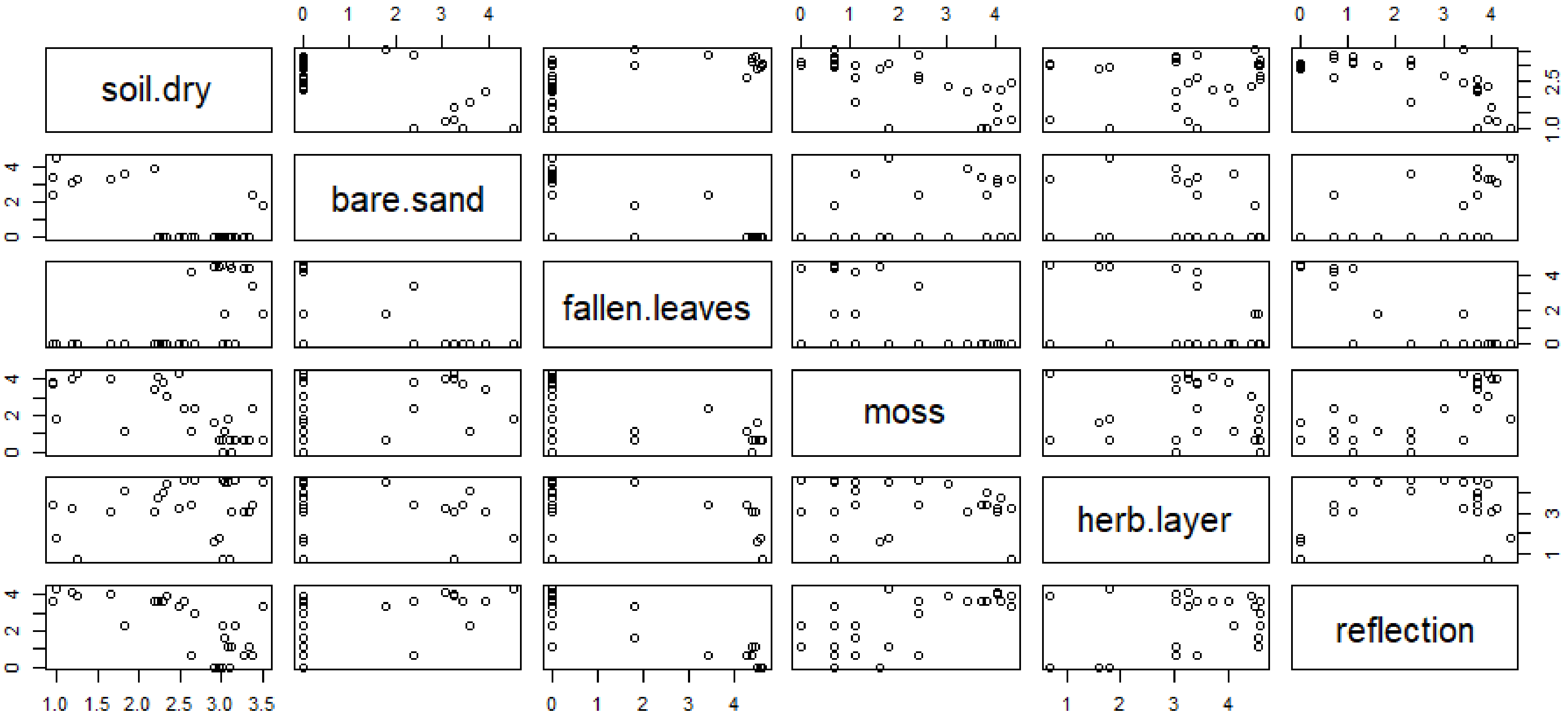


28 Observations

Species abundance (12)

Environmental variables - x (6)

Check enviro variables for colinearity

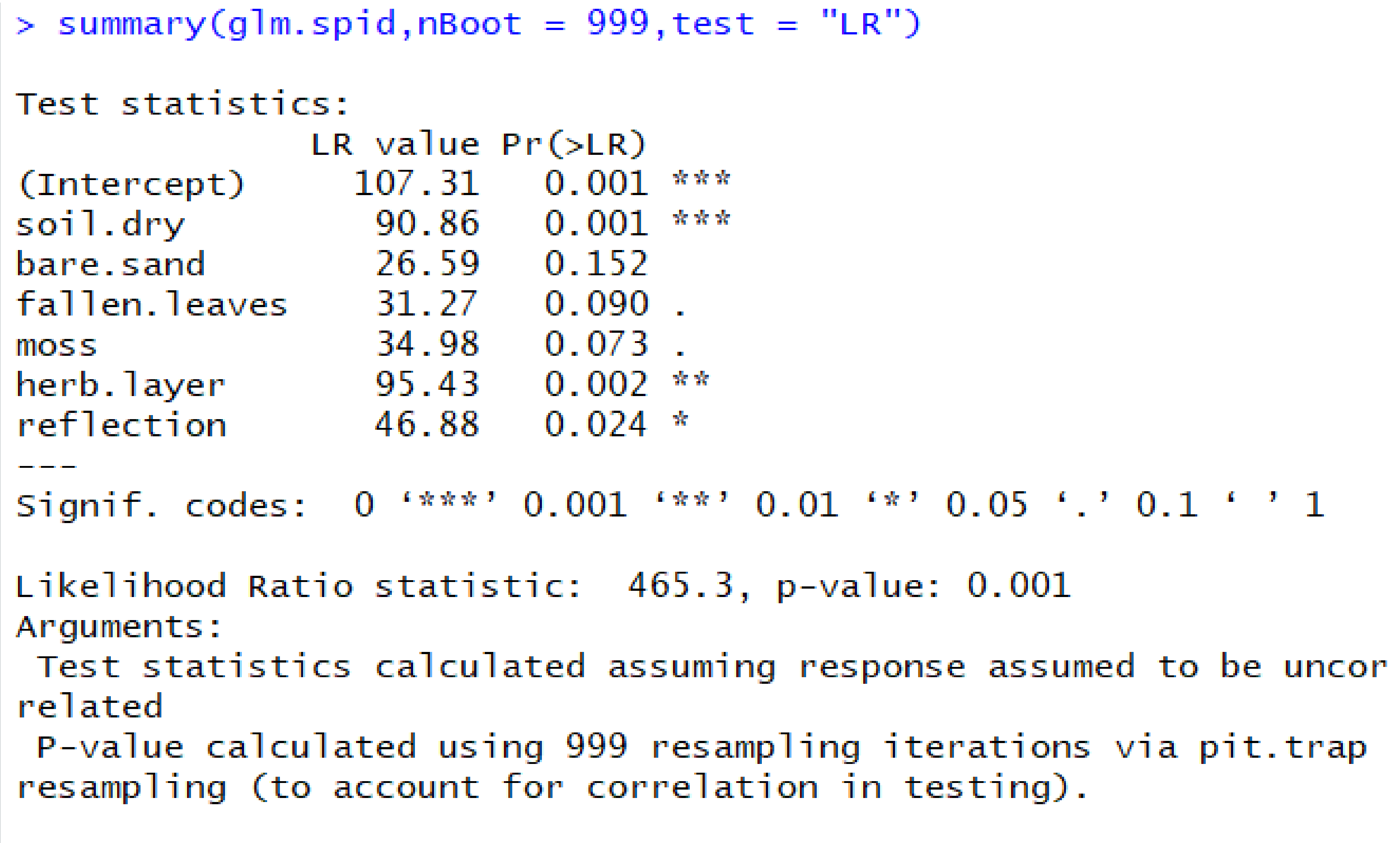


`manyglm` applied to Hunting spiders

```
> spiddat <- mvabund(spider$abund)
> X<-data.frame(spider$x)

>glm.spid <- manyglm(spiddat ~.,data=X)
```

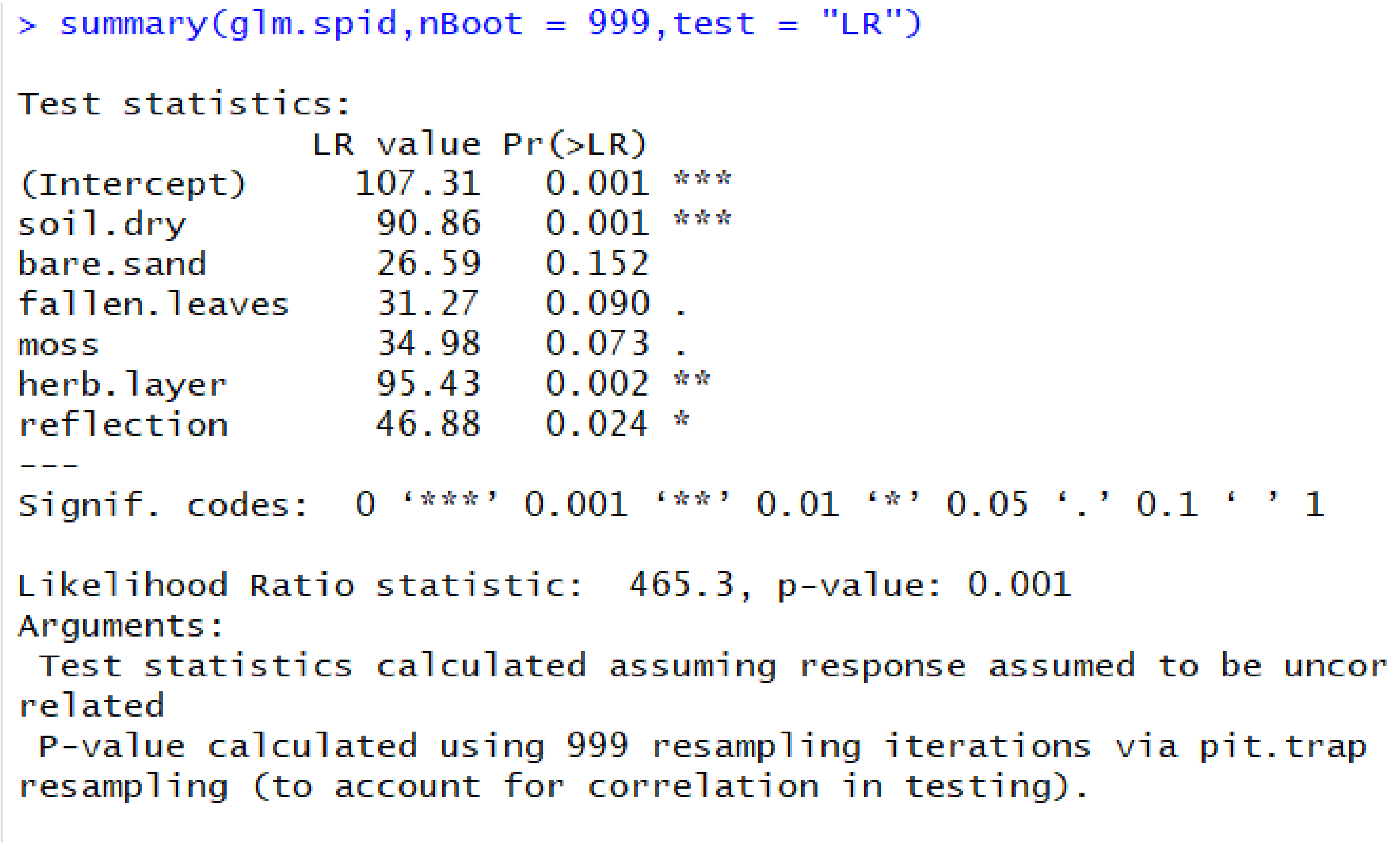
Default = negative binomial

summary fn → conditional effects



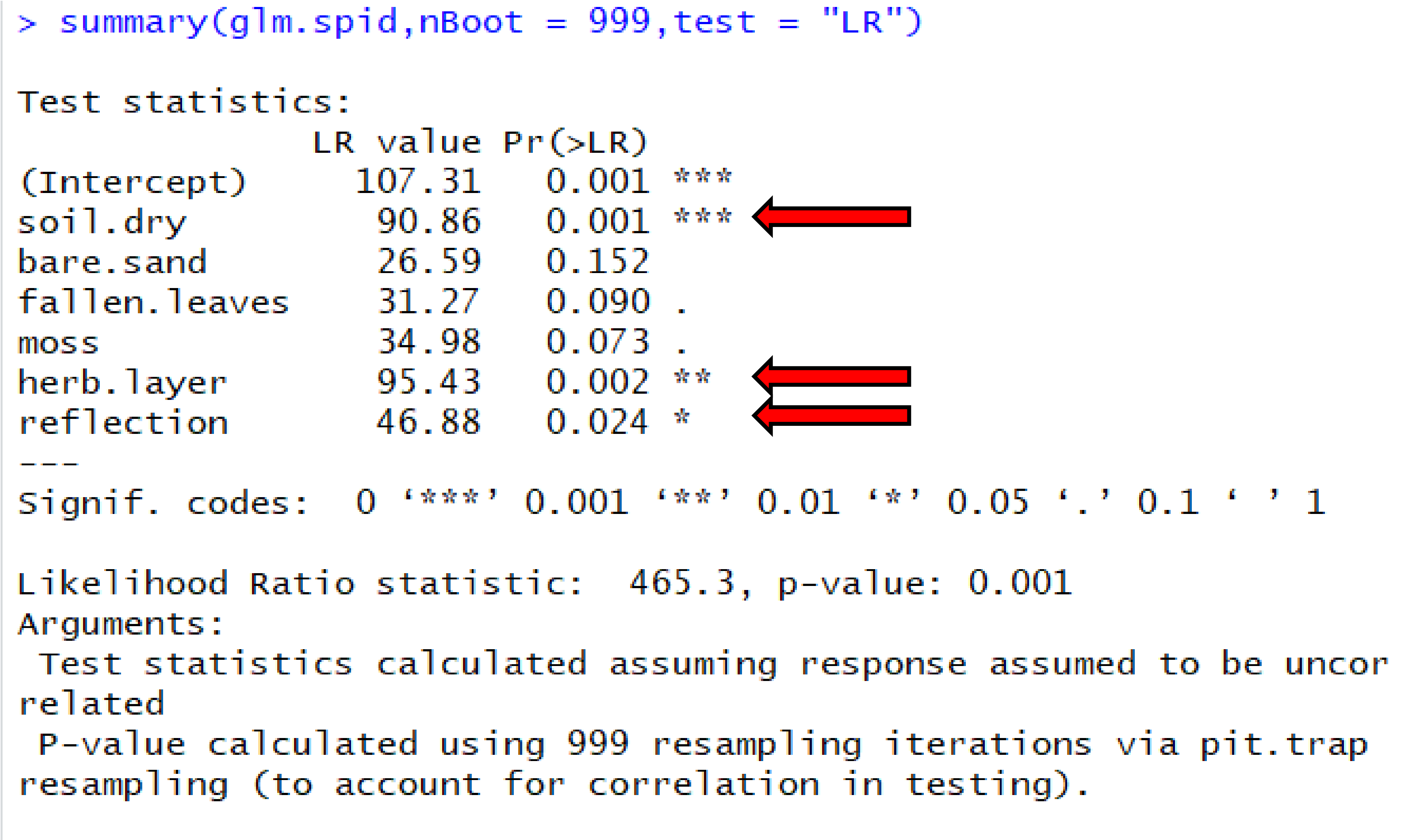
```
> summary(glm.spid,nBoot = 999,test = "LR")

Test statistics:
              LR value Pr(>LR)    
(Intercept)     107.31   0.001 ***
soil.dry         90.86   0.001 ***
bare.sand        26.59   0.152    
fallen.leaves    31.27   0.090 .  
moss             34.98   0.073 .  
herb.layer       95.43   0.002 ** 
reflection       46.88   0.024 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Likelihood Ratio statistic:  465.3, p-value: 0.001
Arguments:
 Test statistics calculated assuming response assumed to be uncor
related
 P-value calculated using 999 resampling iterations via pit.trap
resampling (to account for correlation in testing).
```

summary fn → conditional effects



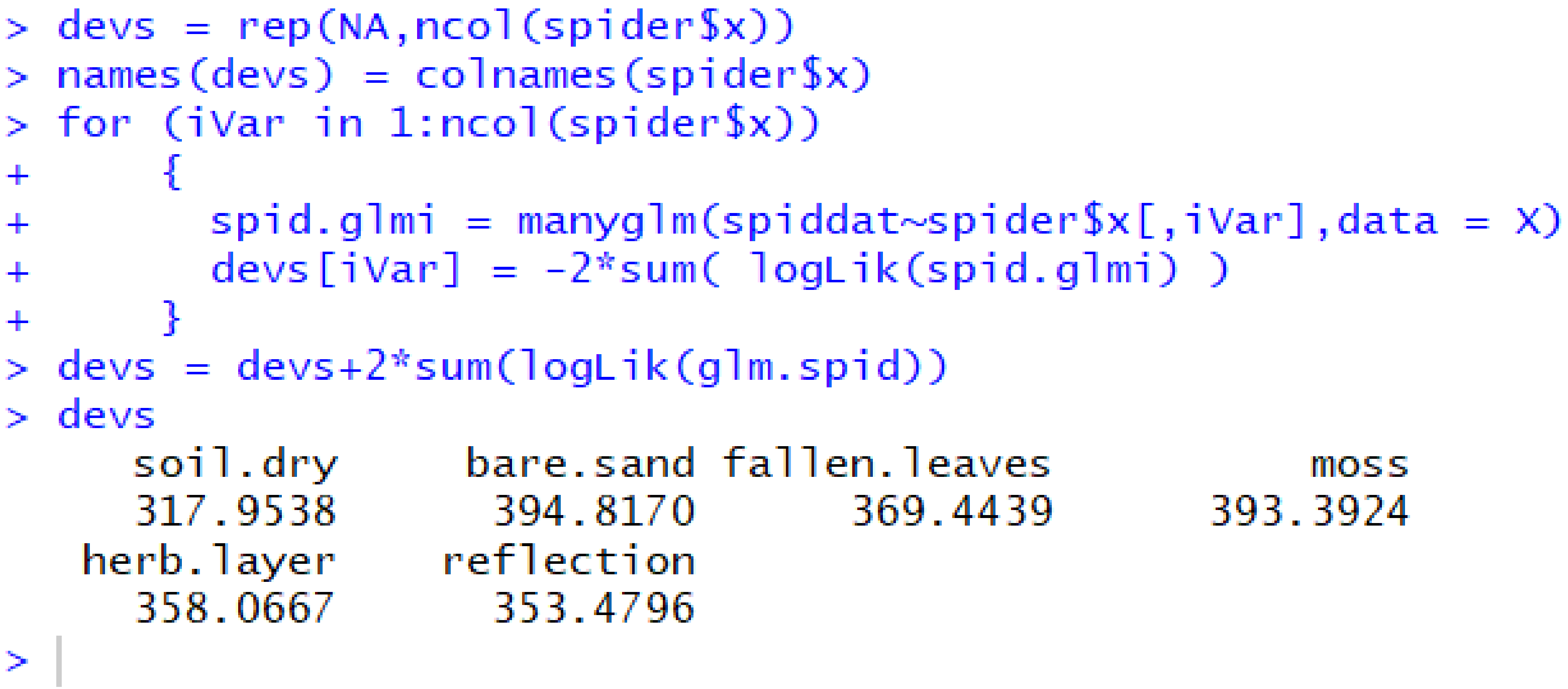
```
> summary(glm.spid,nBoot = 999,test = "LR")

Test statistics:
              LR value Pr(>LR)
(Intercept)     107.31   0.001 ***
soil.dry         90.86   0.001 ***
bare.sand        26.59   0.152
fallen.leaves    31.27   0.090 .
moss             34.98   0.073 .
herb.layer       95.43   0.002 **
reflection       46.88   0.024 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Likelihood Ratio statistic:  465.3, p-value: 0.001
Arguments:
 Test statistics calculated assuming response assumed to be uncor
related
 P-value calculated using 999 resampling iterations via pit.trap
resampling (to account for correlation in testing).
```

summary fn → conditional effects



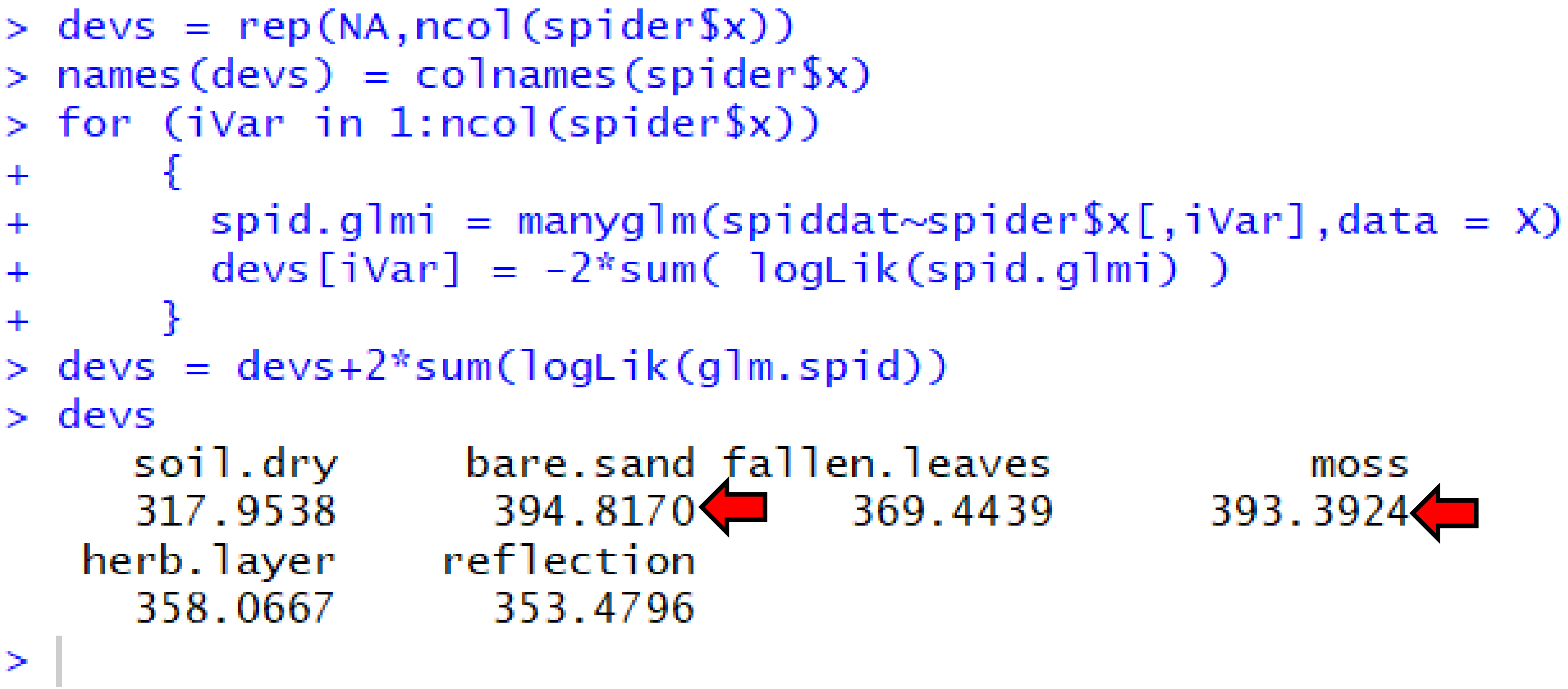
```
> summary(glm.spid,nBoot = 999,test = "LR")

Test statistics:
              LR value Pr(>LR)
(Intercept)     107.31   0.001 ***
soil.dry         90.86   0.001 ***
bare.sand        26.59   0.152
fallen.leaves    31.27   0.090 .
moss             34.98   0.073 .
herb.layer       95.43   0.002 **
reflection       46.88   0.024 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Likelihood Ratio statistic:  465.3, p-value: 0.001
Arguments:
 Test statistics calculated assuming response assumed to be uncor
related
 P-value calculated using 999 resampling iterations via pit.trap
resampling (to account for correlation in testing).
```

Considering marginal effects



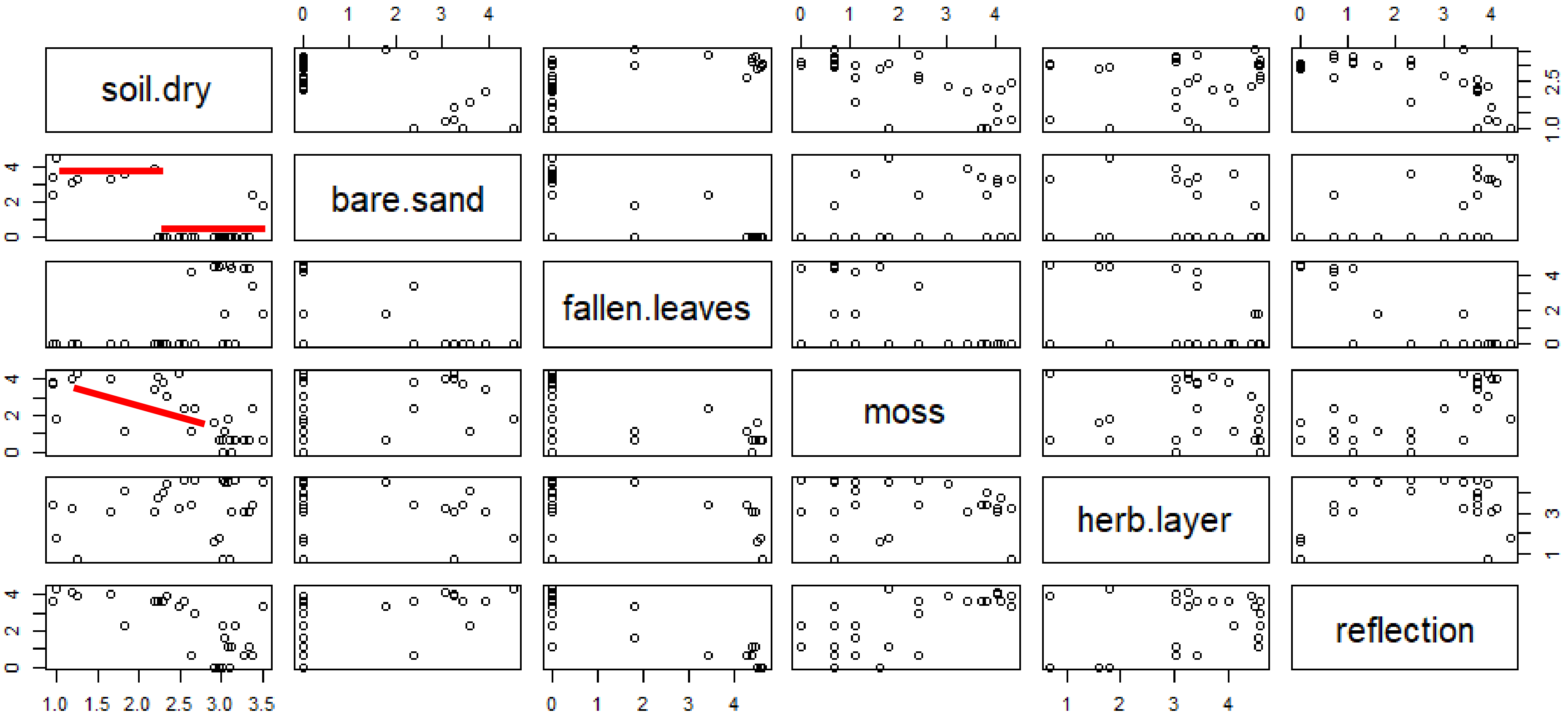
```
> devs = rep(NA,ncol(spider$x))
> names(devs) = colnames(spider$x)
> for (iVar in 1:ncol(spider$x))
+     {
+       spid.glmi = manyglm(spiddat~spider$x[,iVar],data = X)
+       devs[iVar] = -2*sum( logLik(spid.glmi) )
+     }
> devs = devs+2*sum(logLik(glm.spid))
> devs
     soil.dry     bare.sand fallen.leaves          moss
     317.9538      394.8170      369.4439      393.3924
   herb.layer    reflection
     358.0667      353.4796
> 
```

Considering marginal effects



```
> devs = rep(NA,ncol(spider$x))
> names(devs) = colnames(spider$x)
> for (iVar in 1:ncol(spider$x))
+     {
+        spid.glmi = manyglm(spiddat~spider$x[,iVar],data = X)
+        devs[iVar] = -2*sum( logLik(spid.glmi) )
+     }
> devs = devs+2*sum(logLik(glm.spid))
> devs
    soil.dry     bare.sand fallen.leaves          moss
    317.9538      394.8170      369.4439      393.3924
  herb.layer    reflection
    358.0667      353.4796
> 
```

Look at the data again



What do we want to know?

- Does treatment have an effect on assemblage? ✓
  - discrete explanatory variables
- What are the indicator species? ✓
- Which environmental variables are most strongly associated with an assemblage? ✓
  - Continuous explanatory variables

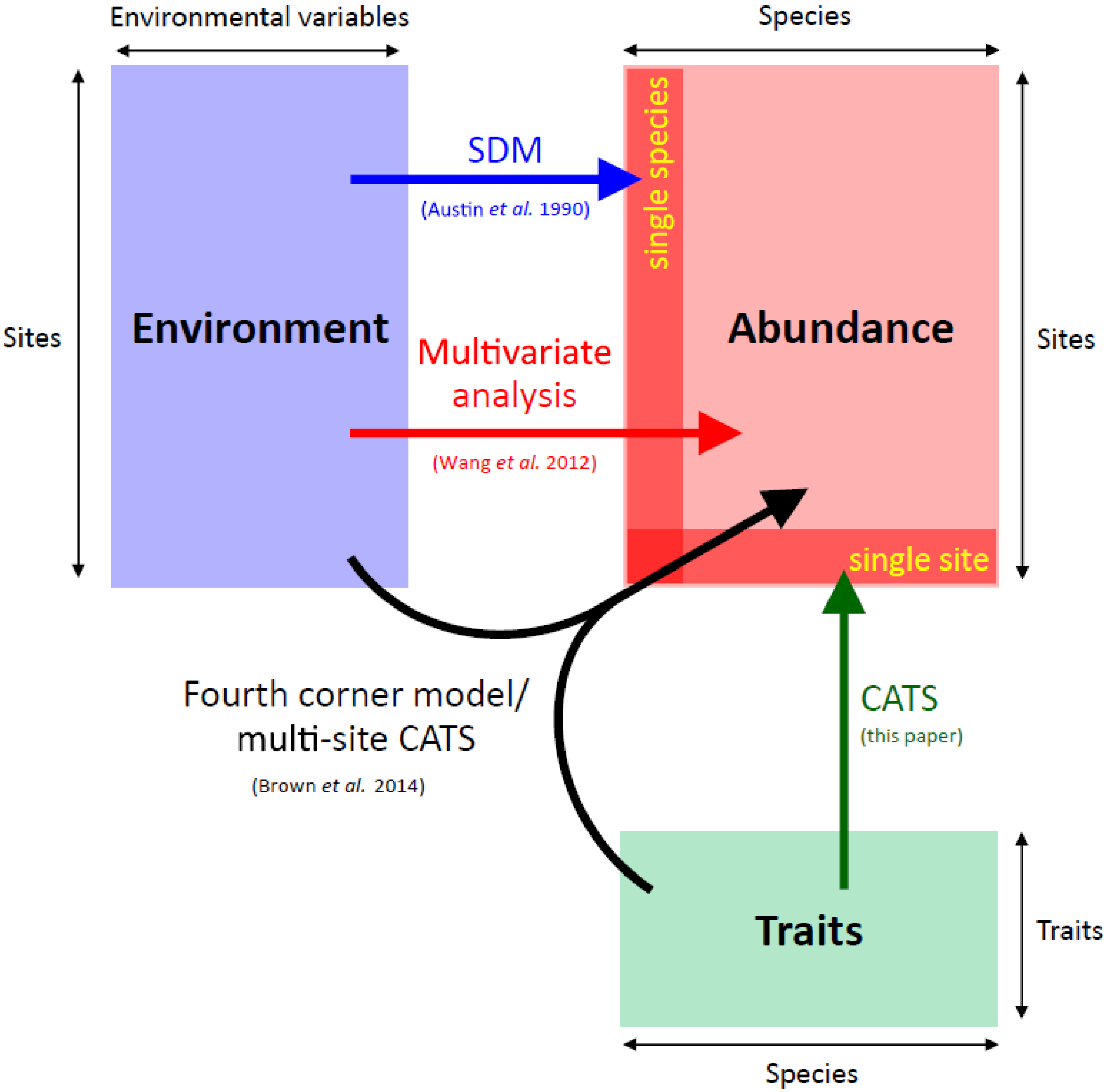
The `manyglm` function handles some of the most common families needed in ecology:

- `"negative.binomial"` for overdispersed counts (using a log link)
- `"poisson", poisson()` or similar for counts that are not overdispersed (using a log link)
- `"binomial", binomial()` or similar, for presence-absence data (using a logit-link)
- `"cloglog", binomial("cloglog")` or similar, for presence-absence data using a complementary log-log link (recommended for presence/absence data)

`manyany` function

- Compositional change
- Relative abundance
- More flexible than manyglm
- Can use many different models, e.g. GLM, GAM, etc
- Adds tweedie family – biomass or more obscure measures of abundance

`traitglm` function



Environmental variables

Species

Sites

Environment

SDM

(Austin *et al.* 1990)

single species

Abundance

Sites

Multivariate analysis

(Wang *et al.* 2012)

single site

Fourth corner model/
multi-site CATS

(Brown *et al.* 2014)

CATS

(this paper)

Traits

Traits

Species

Further info

- Wang, Y. , Naumann, U. , Wright, S. T. and Warton, D. I. (2012), mvabund– an R package for model-based analysis of multivariate abundance data. Methods in Ecology and Evolution, 3: 471-474. doi:10.1111/j.2041-210X.2012.00190.x
- mvabund vignette https://cran.r-project.org/web/packages/mvabund/index.html
- http://eco-stats.blogspot.com/2012/03/introducing-mvabund-package-and-why.html
- http://environmentalcomputing.net/introduction-to-mvabund/
- http://rpubs.com/dwarton/68823
- Warton, D. I., Wright, S. T. and Wang, Y. (2012), Distance-based multivariate analyses confound location and dispersion effects. Methods in Ecology and Evolution, 3: 89-101. doi:10.1111/j.2041-210X.2011.00127.x
- PIT-trap bootstrapping https://doi.org/10.1371/journal.pone.0181790
- Warton, D. I., Shipley, B. and Hastie, T. (2015), CATS regression – a model-based approach to studying trait-based community assembly. Methods Ecol Evol, 6: 389-398. doi:10.1111/2041-210X.12280