Generalised Linear Models for data of multiple species

Bert van der Veen

Department of Mathematical Sciences, NTNU

Outline

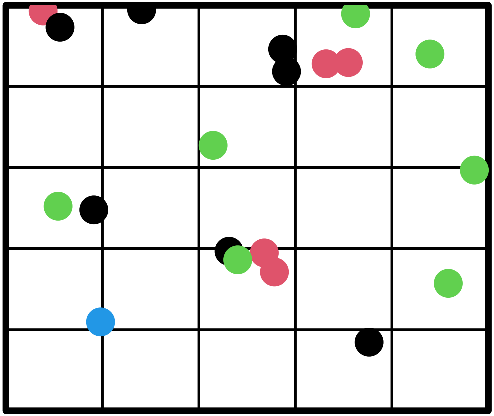
- Brief recap of sampling theory
- Generalised Linear Models background
  - Because GLLVMs are an extension
- Binomial, Poisson, Negative binomial
- Building on some material from the GLM workshop

Sampling data

Figure 1: dw.com

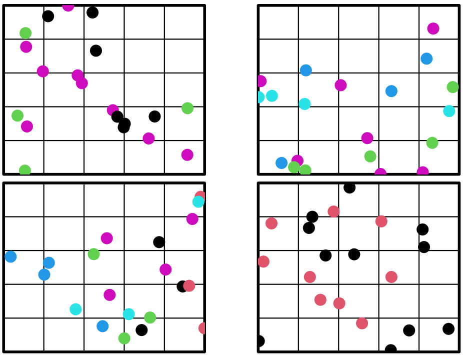
A meadow in the Dutch dunes.

How many plant species are there in this meadow?



We go into the meadow and count plants in quadrats. We find 4 species.

Resampling the dunes

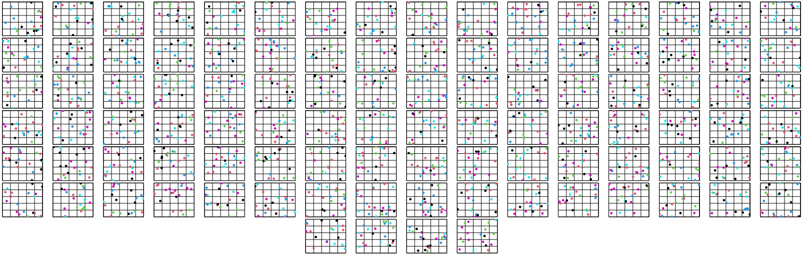


We resample the dune meadow, and find different numbers of species: 3, 3, 6, 2. And different compositions. On average we have found 3.5 species per quadrat.

Sampling variation

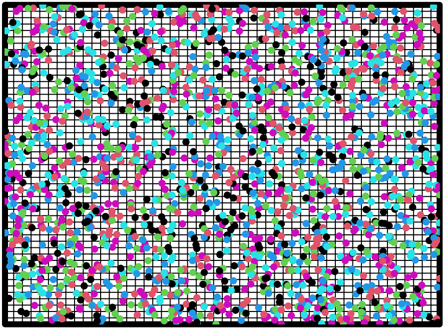
1) Each time we sample, we have a slightly different sample
2) Each time we estimate a parameter, it might be slightly different due to this sampling variation
3) The more data we sample, the better we can represent this variability of our estimate
4) And the better we can represent the true richness/cover in the meadow

Sampling variation



If we sample many times, we have an accurate picture of the whole meadow (and variability in the number of species in a quadrat that we might find).

Sampling variation

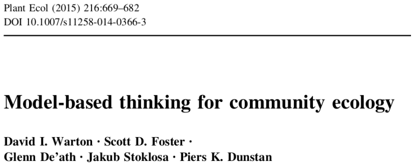


If we sample in one large (field-sized) quadrat we also have an accurate picture of the whole meadow (and variability in the number of species in a quadrat that we might find).

Classical approach

Classically:

1) Decide on a distance
2) Fit an ordination
3) Make a plot
4) Do a hypothesis test



Plant Ecol (2015) 216:669–682
DOI 10.1007/s11258-014-0366-3

Model-based thinking for community ecology

David I. Warton · Scott D. Foster ·
Glenn De'ath · Jakub Stoklosa · Piers K. Dunstan

Statistical modeling

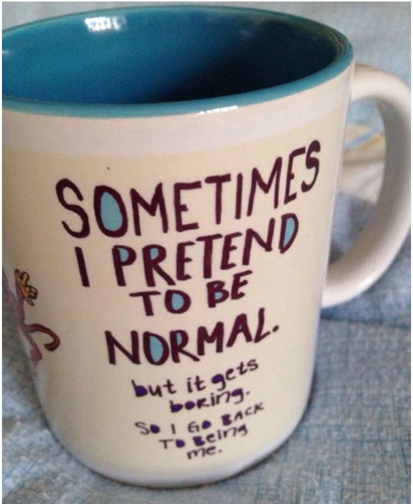
Instead of a focus on data, we consider the *data generating process*

- We collect data
- Decide on a research question for *the population*
- Learn about the variation in the data
  - Which requires formulating a model
- Work out distribution of the estimates
  - And find the “best” estimate
- Conclude if our answer is robust for the population (e.g., fields like this have more than 6 species)

Generalised Linear Models

For when the assumptions of linear regression fail.

- Linearity (straight line)
- Independence of errors
- Homoscedasticity (same variance for all errors)
- Normality (distribution of errors)



Generalised linear models (GLMs)

GLMs as a framework were introduced by Nelder and Wedderburn (1972) uniting many different models. With a special focus on teaching statistics.

- Linear regression
- Logistic regression
- Probit regression
- Complementary log-log regression
- Log-linear regression
- Gamma regression

Generalised linear models (2)

GLMs extend the linear model framework to address:

- Variance changes with the mean
- Range of **y** is bounded

The basis of many statistical models in Biology

Components of a GLM

- Systematic component: η
- Random component: data/distribution)
- The link function: connects these components
  - This is not a data transformation
- The variance function

But no explicit error term

GLM Likelihood

- We use MLE for estimation
- With a distribution in the "exponential family" (for fixed ϕ)

All GLMs have the likelihood:

$$\mathcal{L}(y_i;\Theta) = \exp\{\frac{y_i\eta_i - b(\eta_i)}{a(\phi)} + c(y_i,\phi)\} \tag{1}$$

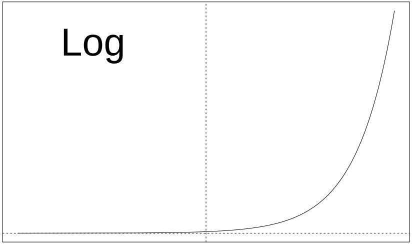
Generalised linear model

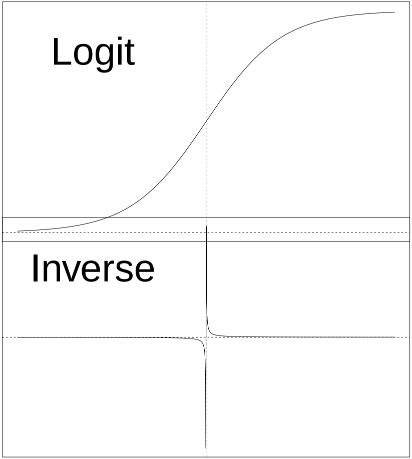
$$\begin{aligned} g\{\mathbb{E}(y_i|x_i)\} &= \eta_i = \alpha + x_i\beta \\ \mathbb{E}(y_i|x_i) &= g^{-1}(\eta_i) = g^{-1}(\alpha + x_i\beta) \end{aligned} \tag{2}$$

g$(\cdot)$ is the **link function**

The link function

- Is a smooth/monotone function
- Has an inverse $g^{-1}(\cdot)$
- Restricts the scale
- g$(\cdot)$ can be e.g.





Variance function

Perhaps most critically, variance changes with the mean:

$$\mathsf{var}(y_i; \mu_i, \phi) = \frac{\partial^2 g(\eta_i)}{\partial \eta_i^2} a(\phi)$$

- ϕ: the dispersion parameter, constant over observations
    - Fixed for some response distributions
- $a(\phi)$ is a function of the form ϕ / w_i (McCullagh and Nelder 1989)

Assumptions

- No outliers
- Independence
- Correct distribution
- Correct link function
- Correct variance function (implied by previous two)

We can check these assumptions with residual plots.

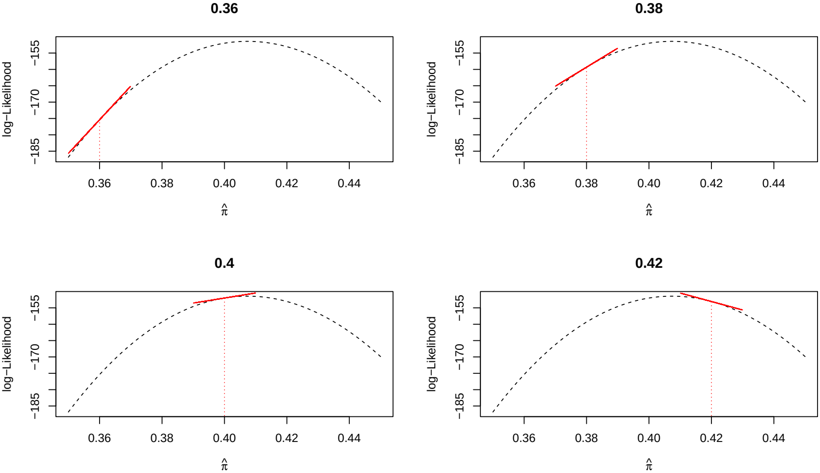
Fitting GLMs

Parameters in GLMs need to be estimated **iteratively**.

- More difficult to fit
- Requires numerical optimisation
- Susceptible to local convergence

Holds for GLLVMs too

Estimating GLMs



We need a good algorithm to find the maximum!

Why is this important?

1) A basic (mathematical) understanding helps apply methods correctly.
2) GLMMs/GLLVMs may not always converge to the MLE. Then, you will get warnings/errors.
3) If you understand them, you might know what to do!

Often used distributions in ecology

- Binomial: occurrence/counts. Presence of species, number of germinated seeds out of a total
- Poisson: counts. Abundance
- Negative binomial (fixed dispersion): counts. Number of species or abundance
- Gamma: (positive) continuous. Body size or biomass
- Ordinal (cumulative link). Cover classes
- Beta (logit link). Cover (note: not a GLM)

Binomial GLM use

- When a linear regression is not appropriate :)
- For binary data or counts of successes/failures

In ecology

- Predicting species' distributions
- Number of germinated plant seeds
- Prevalence of disease in a population
- Probability of observing a behavior
- Proportion of orchids 🙃

The binomial GLM

Link functions:

- Logit: $\log(\frac{\pi_i}{1-\pi_i})$ and inverse $\frac{\exp(\eta_i)}{1+\exp(\eta_i)}$ - *the canonical link*
- Probit: $\Phi^{-1}(\pi_i)$ and inverse $\Phi(\eta_i)$
- Complementary log-log: $\log(-\log(1-\pi_i)$ and inverse $1-\exp(-\exp(\eta_i))$
- ~~Log-log~~

- Logit is canonical and easier to interpret
- Probit is sometimes easier mathematically than Logit
- Complementary log-log for counts

Example: count of wolfspiders in Dutch dunes

- Data originally by van der Aart et al. (1975)
- Counts of 12 species, caught with pitfall traps

Spiders: the data

A classical dataset that is commonly re-analyzed

Alopacce	Alopcune	Alopfabr	Arctlute	Arctperi	Auloalbi	Pardlugu	Pardmont	Pardnigr	Pardp
25	10	0	0	0	4	0	60	12	
0	2	0	0	0	30	1	1	15	
15	20	2	2	0	9	1	29	18	
2	6	0	1	0	24	1	7	29	
1	20	0	2	0	9	1	2	135	
0	6	0	6	0	6	0	11	27	
2	7	0	12	0	16	1	30	89	1
0	11	0	0	0	7	55	2	2	
1	1	0	0	0	0	0	26	1	
3	0	1	0	0	0	0	22	0	
15	1	2	0	0	1	0	95	0	
16	13	0	0	0	0	0	96	1	
3	43	1	2	0	18	1	24	53	
0	2	0	1	0	4	3	14	15	
0	0	0	0	0	0	6	0	0	
0	3	0	0	0	0	6	0	2	
0	0	0	0	0	0	2	0	0	
0	1	0	0	0	0	5	0	0	
0	1	0	0	0	0	12	0	1	
0	2	0	0	0	0	13	0	0	
0	1	0	0	0	0	16	1	0	

Spiders: the environmental data

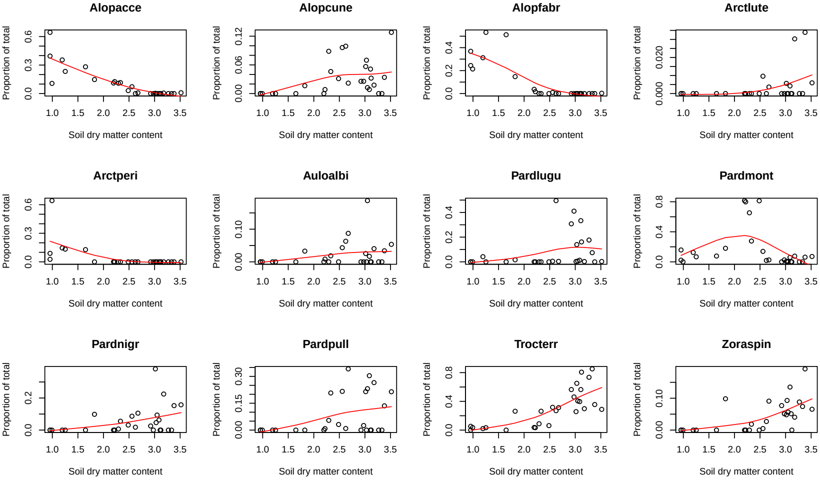
soil.dry	bare.sand	fallen.leaves	moss	herb.layer	reflection
2.3321	0.0000	0.0000	3.0445	4.4543	3.9120
3.0493	0.0000	1.7918	1.0986	4.5643	1.6094
2.5572	0.0000	0.0000	2.3979	4.6052	3.6889
2.6741	0.0000	0.0000	2.3979	4.6151	2.9957
3.0155	0.0000	0.0000	0.0000	4.6151	2.3026
3.3810	2.3979	3.4340	2.3979	3.4340	0.6931
3.1781	0.0000	0.0000	0.6931	4.6151	2.3026
2.6247	0.0000	4.2627	1.0986	3.4340	0.6931
2.4849	0.0000	0.0000	4.3307	3.2581	3.4012
2.1972	3.9318	0.0000	3.4340	3.0445	3.6889
2.2192	0.0000	0.0000	4.1109	3.7136	3.6889
2.2925	0.0000	0.0000	3.8286	4.0254	3.6889
3.5175	1.7918	1.7918	0.6931	4.5109	3.4012
3.0865	0.0000	0.0000	1.7918	4.5643	1.0986
3.2696	0.0000	4.3944	0.6931	3.0445	0.6931
3.0301	0.0000	4.6052	0.6931	0.6931	0.0000
3.3322	0.0000	4.4543	0.6931	3.0445	1.0986
3.1224	0.0000	4.3944	0.0000	3.0445	1.0986
2.9232	0.0000	4.5109	1.6094	1.6094	0.0000
3.1091	0.0000	4.5951	0.6931	0.6931	0.0000
2.9755	0.0000	4.5643	0.6931	1.7918	0.0000
1.2528	3.2581	0.0000	4.3307	0.6931	3.9120
1.1939	3.0445	0.0000	4.0254	3.2581	4.0943
1.6487	3.2581	0.0000	4.0254	3.0445	4.0073

Spiders: to long format

```
data <- data.frame(spider$abund, spider$x, total = rowSums(spider$abund))
datalong <- reshape(data,
                    varying = colnames(spider$abund),
                    v.names = "Count",
                    idvar = "Site",
                    timevar = "Species",
                    direction = "long")

datalong$Species <- factor(datalong$Species,
                           labels = colnames(spider$abund))
```

Spiders: visually inspect the data



Is there a community-level trend?

Spiders: fit a model

```
model1 <- glm(cbind(Count, total)~soil.dry,
              data = datalong, family="binomial")
coef(model1)
```

```
##   (Intercept)      soil.dry
## -2.484907e+00 -9.683925e-10
```

- cbind: combines two vectors as columns
- total is the site sum (total spiders at a site)
- Canonical link is used by default
- Intercept and Soil.dry effects are the same for all species

Modeling spiders

1) Is the same effect for all species realistic?
2) Is the same (average) probability of occurrence for all species realistic?

Spiders: species-specific effects

```
model2 <- glm(cbind(Count, total)~soil.dry*Species,
           data = datalong, family="binomial")
```

- One intercept per species, relative to the first species
- One soil.dry effect per species, relative to that of the first species
- one dispersion parameter for all species

Binomial regression: interpreting coefficients

```r
coef(model2)
```

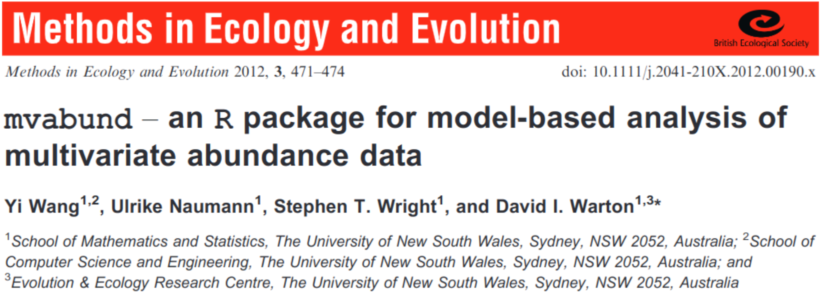
```
##              (Intercept)                 soil.dry          SpeciesAlopcune
##                1.1321838               -1.7508946               -5.9521931
##          SpeciesAlopfabr          SpeciesArctlute          SpeciesArctperi
##                0.6982423              -11.8871608                0.8371418
##          SpeciesAuloalbi          SpeciesPardlugu          SpeciesPardmont
##               -6.3163244               -4.7672488               -0.9719283
##          SpeciesPardnigr          SpeciesPardpull          SpeciesTrocterr
##               -6.6485015               -4.9776248               -4.4773503
##          SpeciesZoraspin soil.dry:SpeciesAlopcune soil.dry:SpeciesAlopfabr
##               -7.5680861                2.3504791               -0.8279564
## soil.dry:SpeciesArctlute soil.dry:SpeciesArctperi soil.dry:SpeciesAuloalbi
##                3.7084779               -1.7215942                2.4230221
## soil.dry:SpeciesPardlugu soil.dry:SpeciesPardmont soil.dry:SpeciesPardnigr
##                1.8815112                0.9258855                2.9111995
## soil.dry:SpeciesPardpull soil.dry:SpeciesTrocterr soil.dry:SpeciesZoraspin
##                2.4776391                2.4815951                2.9537234
```

- Odds for species 1 at soil.dry 0 is `exp(1.13) = 3.10`
- This decreases by `exp(-1.7509)` for every unit of soil dry matter content `3.10*exp(soil.dry*-1.7509) = 3.10*0.17 = 0.527`

Multispecies modeling

1) Fundamental niche: species have their own preferred environmental conditions (“gleasonian”)
2) Some species might still like similar conditions
3) We might be able to separate a community-level effect (e.g., with GLMMs)

mvabund



Methods in Ecology and Evolution

British Ecological Society

Methods in Ecology and Evolution 2012, **3**, 471–474 doi: 10.1111/j.2041-210X.2012.00190.x

mvabund – an R package for model-based analysis of multivariate abundance data

Yi Wang[1,2], Ulrike Naumann[1], Stephen T. Wright[1], and David I. Warton[1,3*]

[1]School of Mathematics and Statistics, The University of New South Wales, Sydney, NSW 2052, Australia; [2]School of Computer Science and Engineering, The University of New South Wales, Sydney, NSW 2052, Australia; and [3]Evolution & Ecology Research Centre, The University of New South Wales, Sydney, NSW 2052, Australia

GLMs for multiple species (vector GLMs)

Typical count cases

When the maximum count is not fixed:

- Number of species
- Individuals at a location
  - Pidgeons in a city
- Bigfoot reports
- Wrongful convictions
- Stars in the night sky

Counts: a multiplicative process

Say that we have the model:

$$\log(\lambda) = \alpha + x_i\beta \tag{3}$$

- with $\alpha = 1$ and $\beta = \log(2) \approx 0.693$
- x_i is either 0 or 1: either I was fishing or you were
- `exp(1) = 2.71828` the average number of fish I caught
- `exp(1+log(2)) = exp(1)*2 = 5.437` the average number of fish you caught

So, you caught twice as many fish!

Poisson assumptions

- An event can occur $0 \dots \infty$ times
- Events are independent
- The rate of events is constant
- Events cannot occur simultaneously
- **Variance equals the mean**

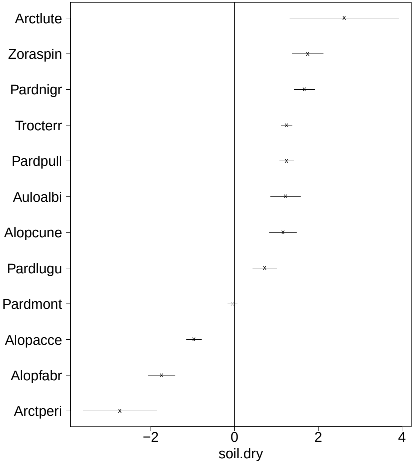
Example: spiders with mvabund

```r
Y <- mvabund::mvabund(spider$abund)
model3 <- mvabund::manyglm(Y ~ soil.dry,
             family = "poisson", data = spider$x)
```

Here, we assume that the total site abundance is not fixed

Log-linear regression: interpreting parameters

- A multiplicative process
- (Intercept) = species **a** log-abundance at soil.dry 0
- Soil dry: `exp(-0.97) =` `0.38`, so half the count by `log(0.5)/-0.97 =` `0.7145847` increase in soil dry matter content



Overdispersion

Our assumption: $\lambda = \text{var}(\mathbf{y})$
Reality: $\lambda \geq \text{var}(\mathbf{y}) = \mu + \mu^2\phi$

- Mean = variance
- If there is more variation, this assumption fails
- Consequences: CIs underestimate, biased parameter estimates, inflation in model selection

For our example: spiders generally occur at low abundances, but some places have an extrodinarily large amount of spiders

Dealing with dispersion: options

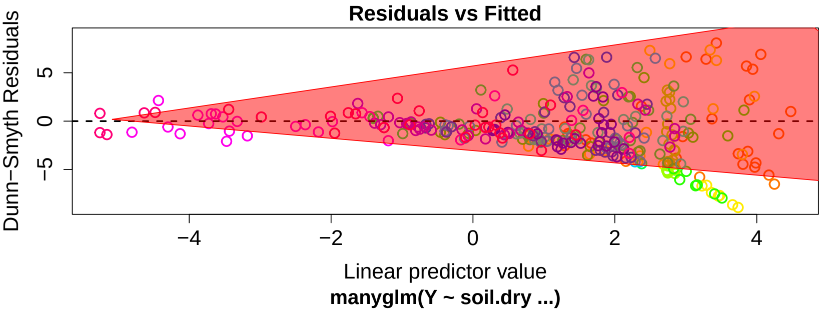
- Correct for it (calculate dispersion)
- Fit a different model
    - Negative binomial (overdispersion, `MASS package`)
    - Conway-Maxwell Poisson
    - Generalized Poisson
    - Quasi-likelihood models (not covered here)
    - Mixed models

Negative binomial distribution

$$\mathcal{L}(y_i;\Theta) = \frac{\Gamma(y_i+\phi)}{\Gamma(\phi)y_i!}\left(\frac{\phi}{\mu_i+\phi}\right)^{\phi}\left(\frac{\mu_i}{\mu_i+\phi}\right)^{y_i} \tag{4}$$

- $\text{var}(\mathbf{y}) = \boldsymbol{\mu} + \frac{\boldsymbol{\mu}^2}{\phi}$
- For large ϕ Poisson!
- Requires more data/information due to extra parameter

Spiders: residuals vs. fitted

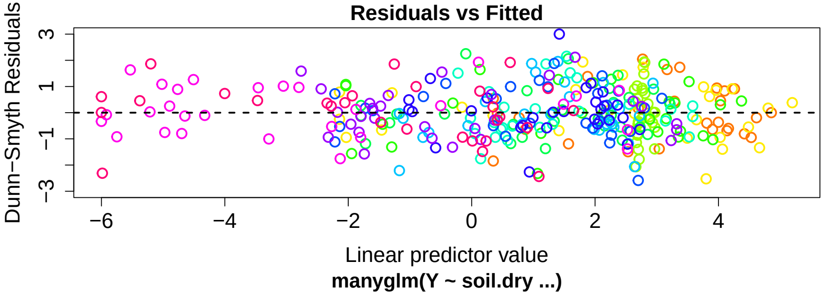


- Residuals fan out: variance increases with the mean
- Count data usually have quadratic mean-variance
- Needs to be accounted for

Spiders: negative-binomial regression

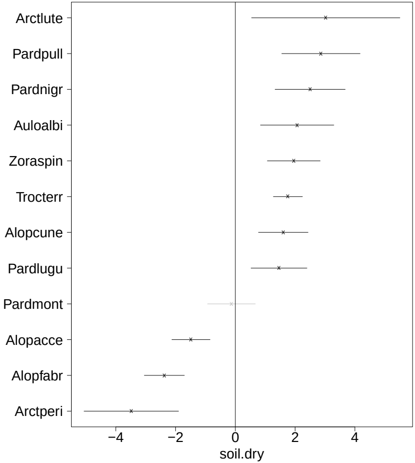
```r
model4 <- model3 <- mvabund::manyglm(Y ~ soil.dry,
            family = "negative.binomial", data = spider$x)
```

```r
plot(model4, which = 1)
```



Spiders: negative-binomial regression

- Uncertainty has become much larger
- Coefficients have changed



Spiders: hypothesis testing

```
## Time elapsed: 0 hr 0 min 4 sec
```

```
## Analysis of Deviance Table
##
## Model: Y ~ soil.dry
##
## Multivariate test:
##             Res.Df Df.diff   Dev Pr(>Dev)
## (Intercept)     27
## soil.dry        26       1 147.3    0.001 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
## Arguments:
##  Test statistics calculated assuming uncorrelated response
```

Spiders: hypothesis testing

```
## Warning in anova.manyglm(model4, cor.type = "R"): The likelihood rat
## only be used if correlation matrix of the abundances is is assumed t
## Identity matrix. The Wald Test will be used.
```

```
## Time elapsed: 0 hr 0 min 2 sec
```

```
## Analysis of Variance Table
##
## Model: Y ~ soil.dry
##
## Multivariate test:
##             Res.Df Df.diff  wald Pr(>wald)
## (Intercept)     27
## soil.dry        26       1 20.37     0.016 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## Arguments:
##  Test statistics calculated assuming unconstrained correlation response
```

Summary

- We use statistical modeling to find an answer to our question, and incorporate uncertainty
- Always start by assessing properties of your data, and formulate a question
- Most ecological data exhibit strong mean-variance relationships
- Accounting for such properties of data is important
  - To make sure we get a valid result
- Multispecies data requires formulating a hypothesis for community/species trends
  - More data → more reflection