MB590-012
Microbiome Analysis
Regression Analysis

Dr. Christine Hawkes

NC STATE UNIVERSITY

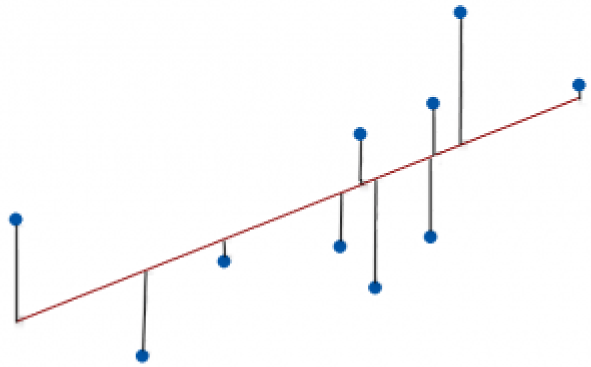
Regression analysis

- Used for prediction of dependent variables based on a set of independent variables

Regression type	Dependent Vars	Independent Vars
Simple	1	1
Multiple	1	>1
Multivariate multiple	>1	>1

Ordinary least squares (OLS) linear regression

- Estimates parameters in regression model by minimizing the sums of the squared residuals
  - i.e., draws a line through the data to minimize sum of squared differences between observed and fitted values



Defining linear models for regression

- General linear model format:
    - Y = B0 + B1*X1 + B2*X2+…+Bn*Xn + error
- Specification in R is

Main effects model	Y ~ factor1 + factor2 + … + factorN
Main effects plus interactions	Y ~ factor1*factor2
Interaction only	Y ~ factor1:factor2

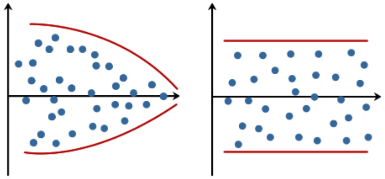
- Can fit a specific model or use variable selection methods

OLS linear regression

- Best subsets = identifies best models from all possible combinations of predictors
  - Uses multiple criteria: P, R2, AIC, BIC, etc.

- Stepwise = “locally optimal” version of best subsets
  - updates included variables by one variable at each step, instead of re-optimizing over all possible subsets
  - Uses P or AIC criteria for entry and removal

OLS linear regression assumptions

Assumption	How to test?	What to do if violated?
Linearity	Graph data or residuals	Transform or use non-linear regression
Normality (symmetric around the mean)	Plot residuals Shapiro-Wilk test Anderson-Darling test	Transform or use non-linear regression R: gnm, nlme
Homoscedasticity (equal variances) 	Breusch-Pagan χ^2 test	Transform or use generalized least squares (GLS), which also minimize residual SS R: MASS::lm.gls, nlme::gls

No multicollinearity (independent vars are not highly correlated)	Variance inflation factors
	vif = 1/(1-R^2)
	vif = 1 not correlated (ideal)
	vif = 1 to 2.5-4, moderately correlated (not ideal)
	vif > 4 high to very high (violated!)
		Drop, combine two vars, or use Ridge or Lasso regression to penalize/shrink coefficients and limit included variables
		R: glmnet

Alternatives for complex multivariate data with more variables than samples

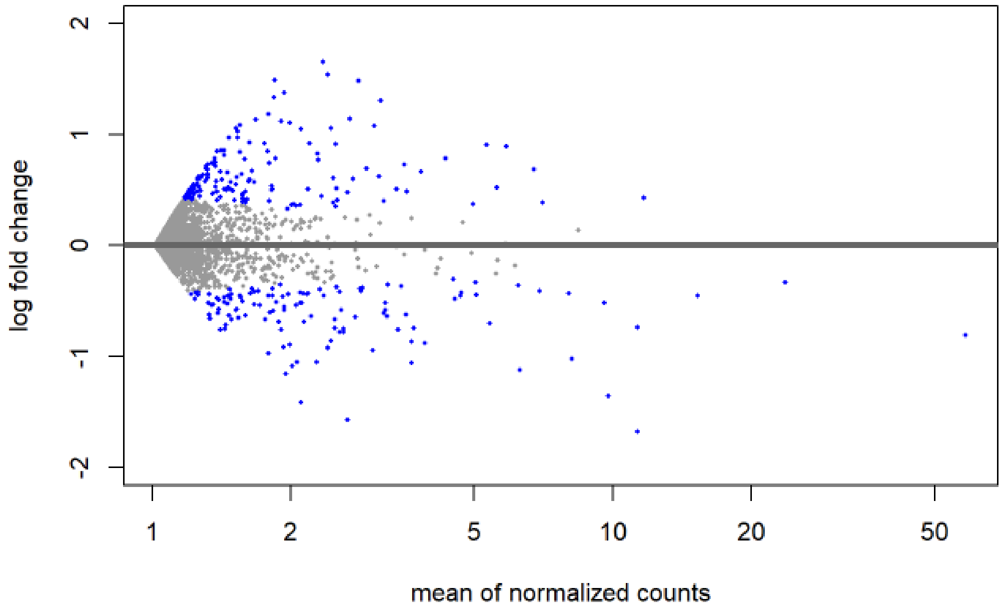
- Ridge or Lasso regression

Alternatives for complex multivariate data with more variables than samples

- Ridge or Lasso regression
  - Shrink regression coefficients for variables with minor contributions to outcome
  - Penalize adding terms to the model
    - Ridge uses L2-norm, which is the sum of squared coefficients, can be tuned by λ term
    - Lasso uses L1-norm, which is the sum of absolute coefficients, also tuned by λ term
    - When λ=0, ridge and lambda are same as OLS
    - As λ increases, ridge coefficients approach zero and lasso coefficients are zero (and zero variables are discarded)
  - Ridge preferred when outcome is function of many predictors that all have similar coefficient values
  - Lasso performs better when some predictors have large and others small coefficients

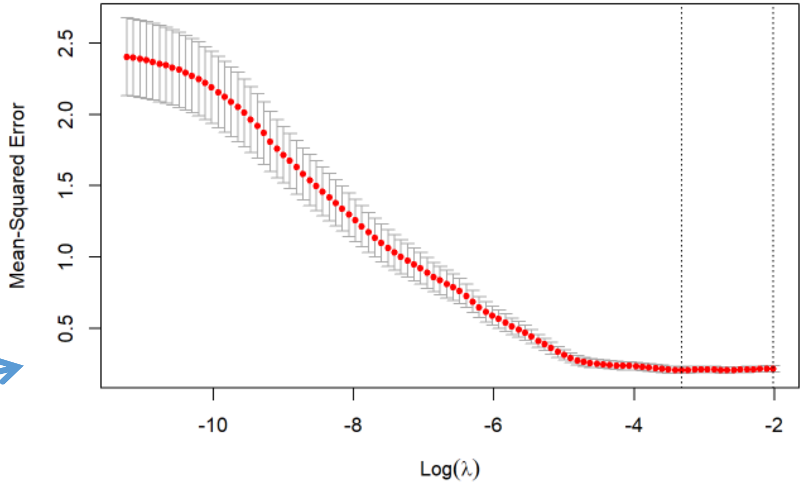
Lasso regression handles large, sparse data

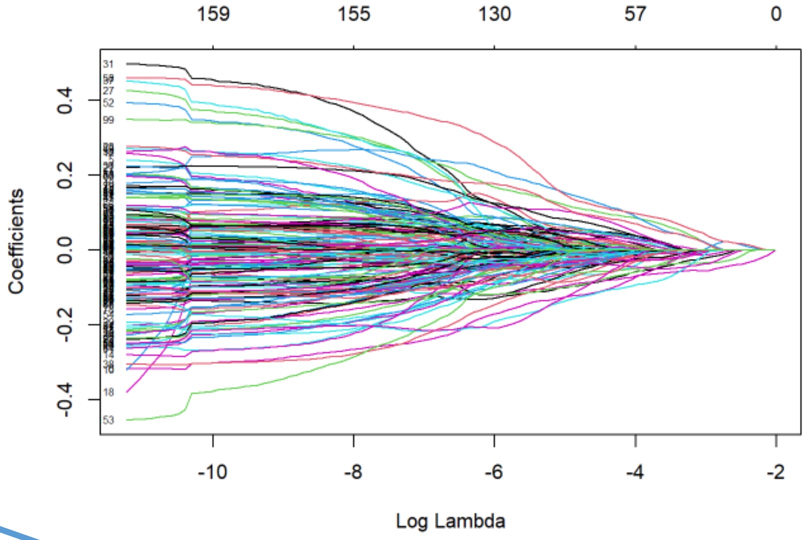
- But computationally slow
- Can first reduce dataset by differential abundance of taxa between groups
    - Treatment and control
    - Two treatments
- DeSeq2



Lasso regression - glmnet

- Use “training” data (~75-80% of data) to identify
  - lambda that minimizes error (tune L1Norm penalty)
    - K-fold cross validation splits data into K “folds”, considers training on all but kth part, then iterates
    - Provides variance estimate for MSE at each lambda
  - predictive features (taxa) given penalty associated with min lambda
- Check model fit and MSE with “test” data





```
## 160 x 1 sparse Matrix of class "dgCMatrix"
##                        s1
## (Intercept)  1.0888593419
## OTU_22        .
## OTU_87       -0.0097569159
## OTU_37        .
## OTU_845       .
## OTU_188       .
## OTU_130      -0.0174419494
## OTU_441      -0.0302163968
## OTU_7         .
```

```
##   Df  %Dev  Lambda
## 1 21 28.63 0.03604
```

```
## $mse
##        s0
## 0.247926
## attr(,"measure")
## [1] "Mean-Squared Error"
```

Common regressions with microbiome data

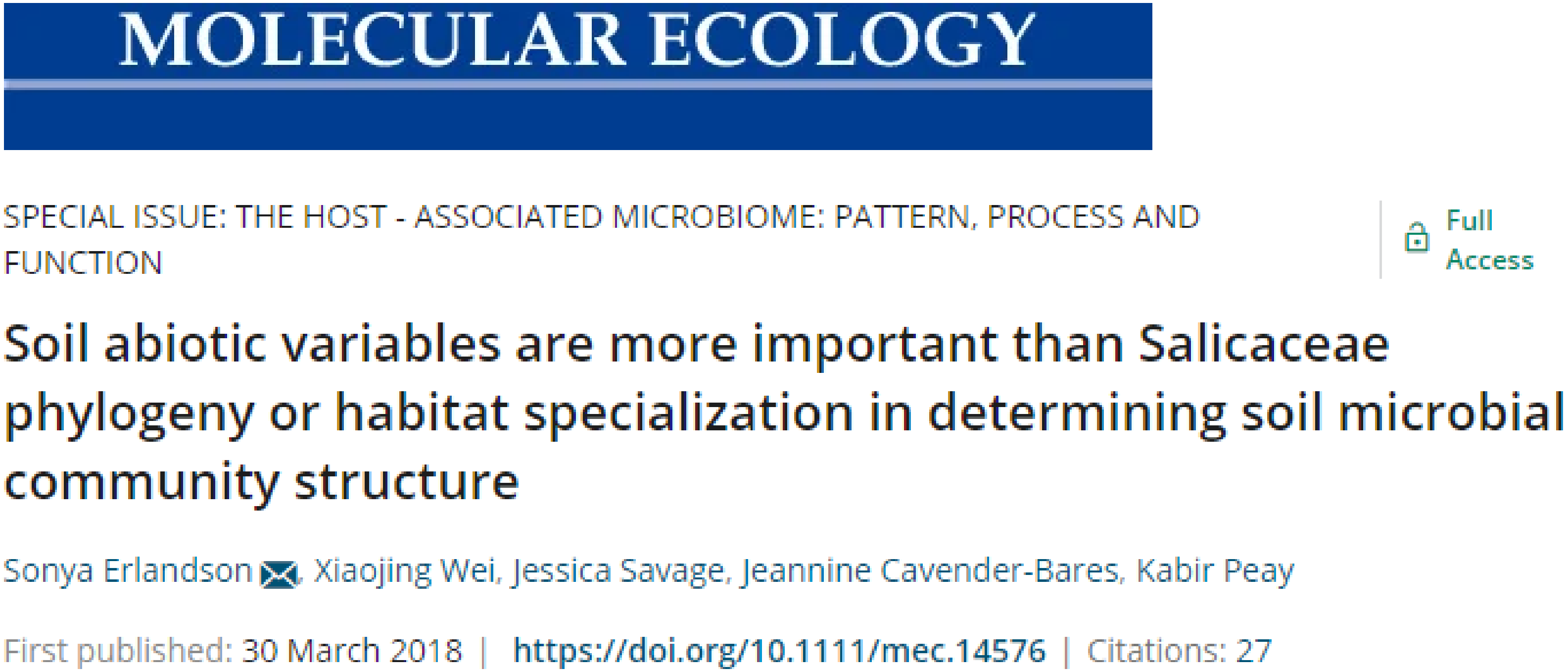
- Alpha diversity as dependent variable
- Ordination axis summarizing beta-diversity as dependent variable

OLS or non-linear regression

- ASV matrix as independent variable
    - Often used to identify taxa that predict plant or ecosystem traits, metabolites, etc

LASSO or Ridge regression

The data



MOLECULAR ECOLOGY

SPECIAL ISSUE: THE HOST - ASSOCIATED MICROBIOME: PATTERN, PROCESS AND FUNCTION

Full Access

Soil abiotic variables are more important than Salicaceae phylogeny or habitat specialization in determining soil microbial community structure

Sonya Erlandson, Xiaojing Wei, Jessica Savage, Jeannine Cavender-Bares, Kabir Peay

First published: 30 March 2018 | https://doi.org/10.1111/mec.14576 | Citations: 27

- Dryad data archive:

  https://datadryad.org/stash/dataset/doi:10.5061/dryad.5f24ks4

The data

- Reduced to only bacteria this week
- Environmental variables
    - Soil properties (water table depth, Nmin, NO3, NH4, pH)
    - Plant properties (height plus many traits)
- Spatial variables (lat, lon)
- Factors
    - Treatment → common garden sites
        - D = upland, W=wetland
    - Ecology → habitat specialization of willow species
        - u = upland specialist, w= wetland specialist, i = generalist
    - Plant species (willows) and genotypes