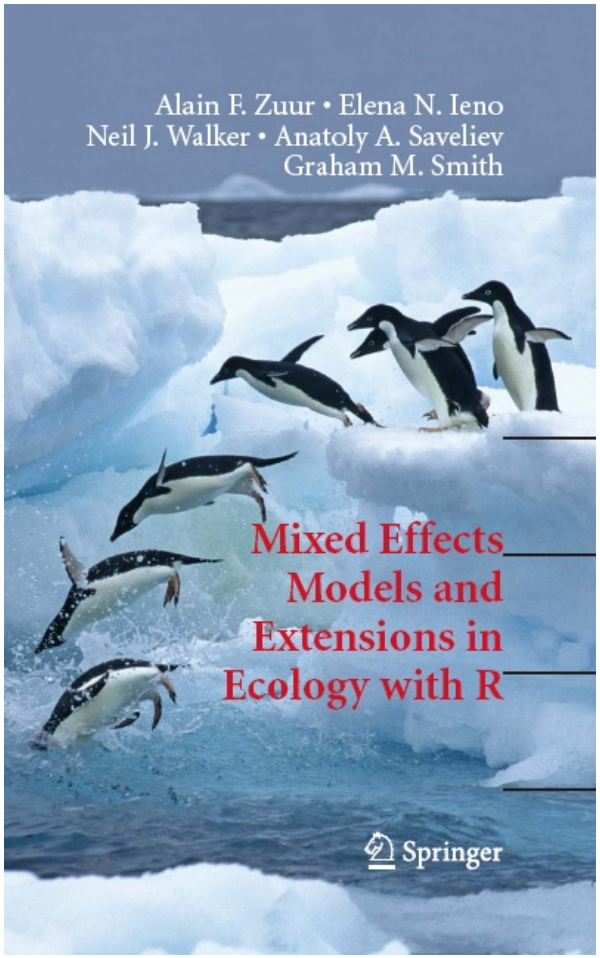
A protocol for data exploration to avoid common statistical problems

Zuur et al (2010)

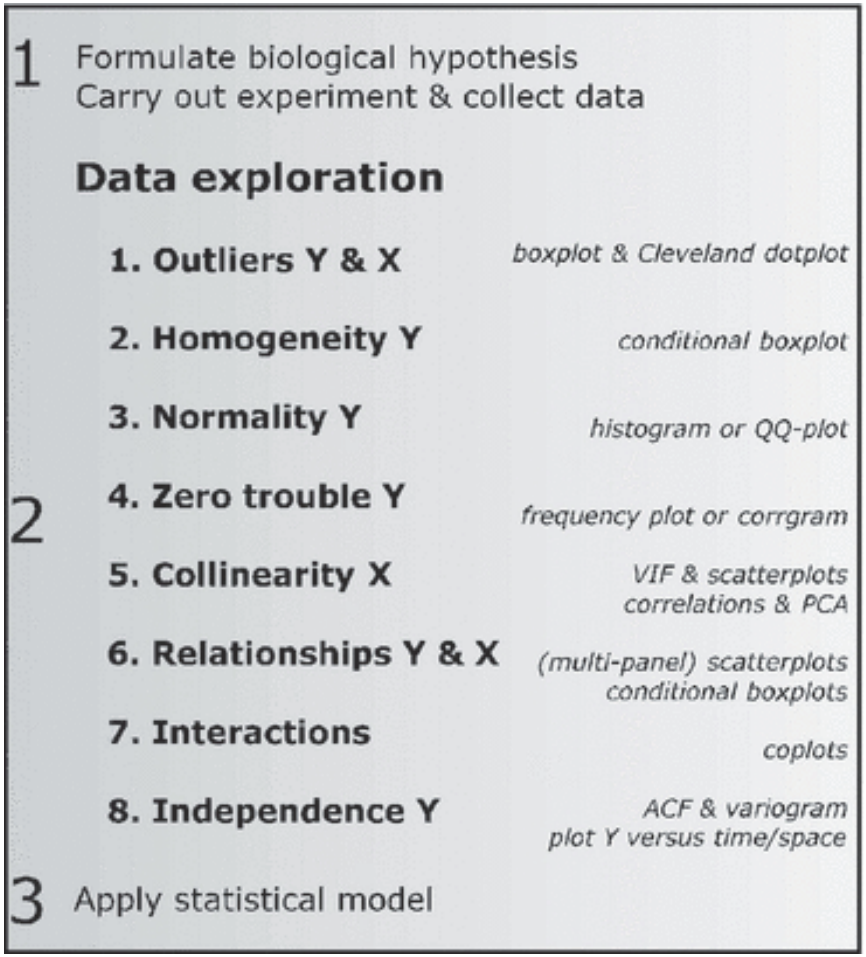
Methods in Ecology and Evolution

Zuur et al 2009 (Mixed Effects Models and Extensions in Ecology with R)

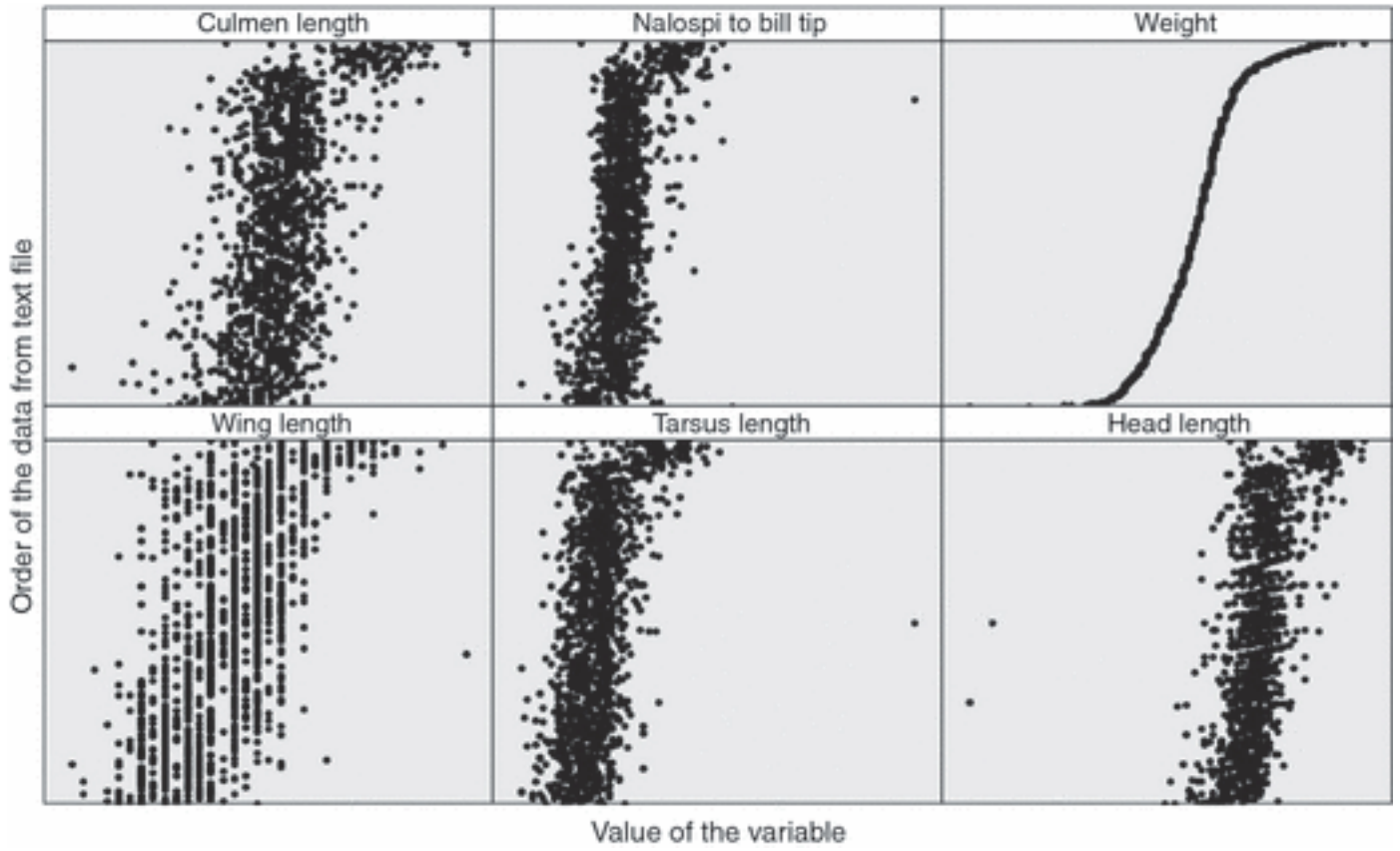


- Appendix A
  - A.1 The data
  - A.2 Data Exploration

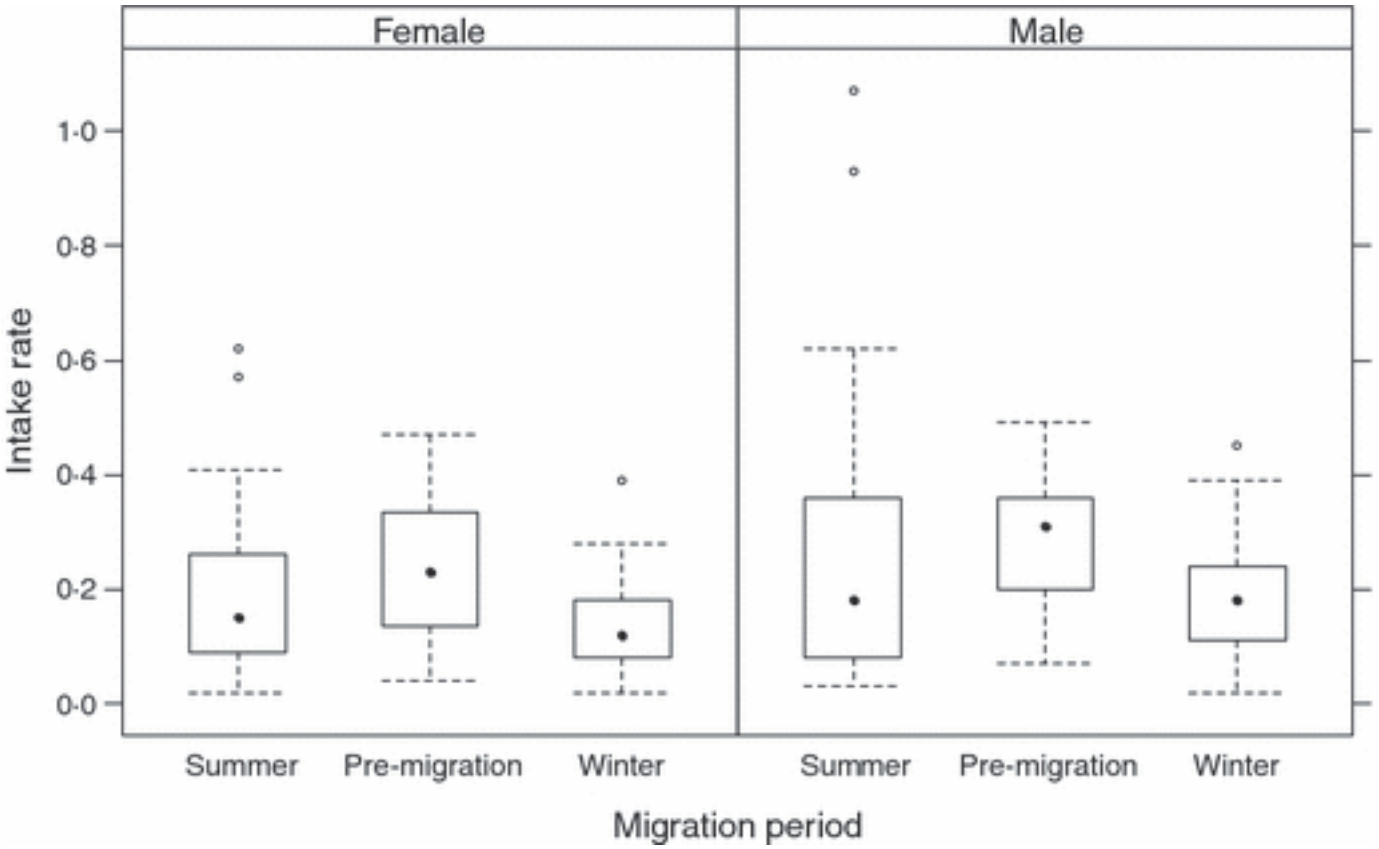
Protocol for data exploration



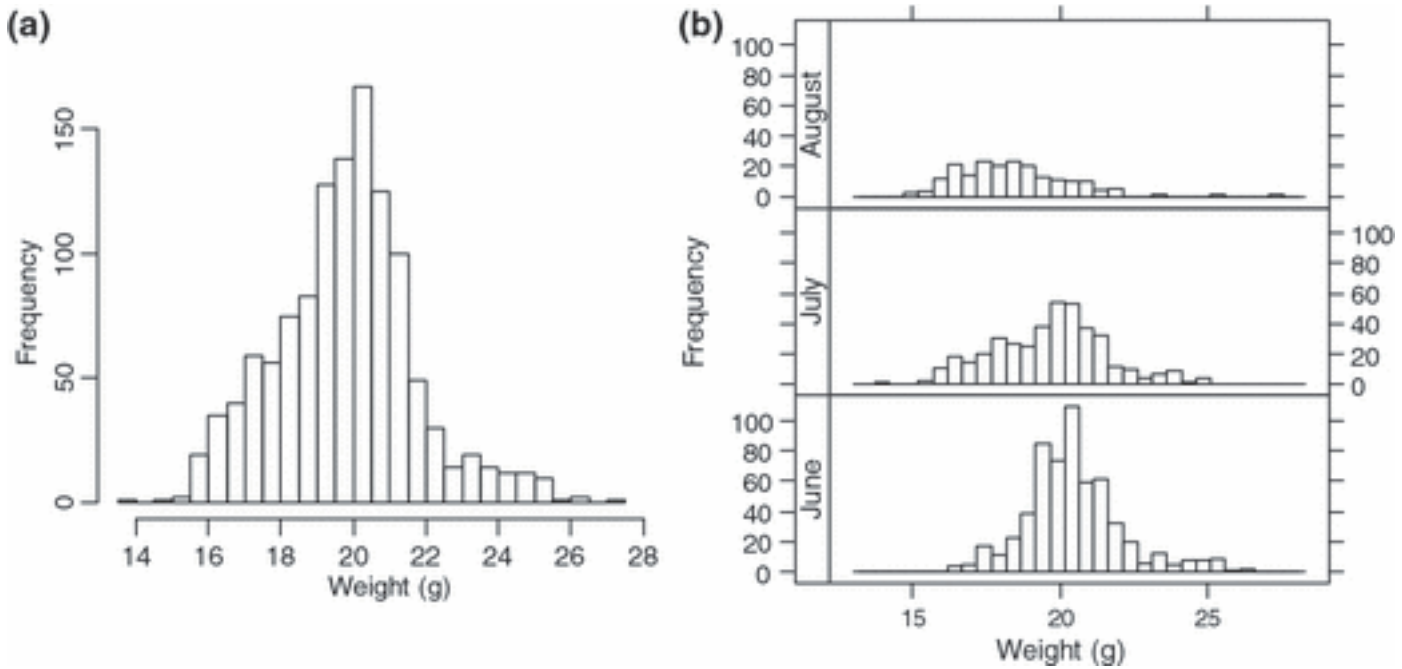
Step 1: Are there outliers in Y and X?



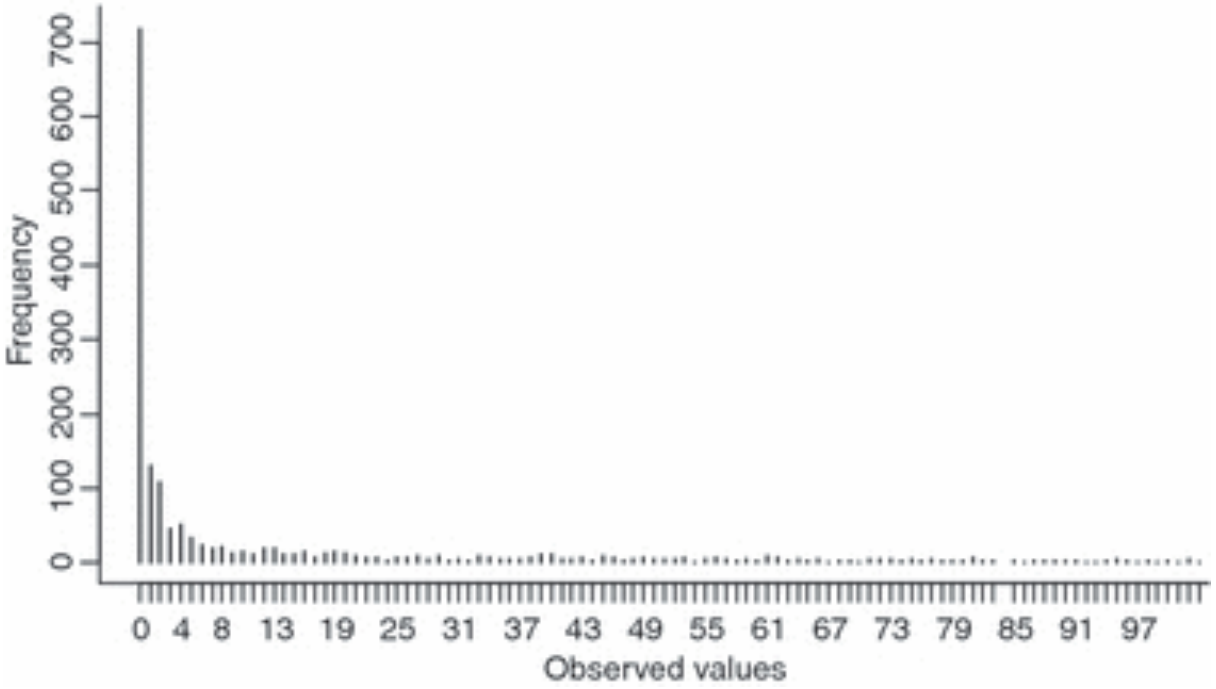
Step 2: Do we have homogeneity of variance?



Step 3: Are the data normally distributed?



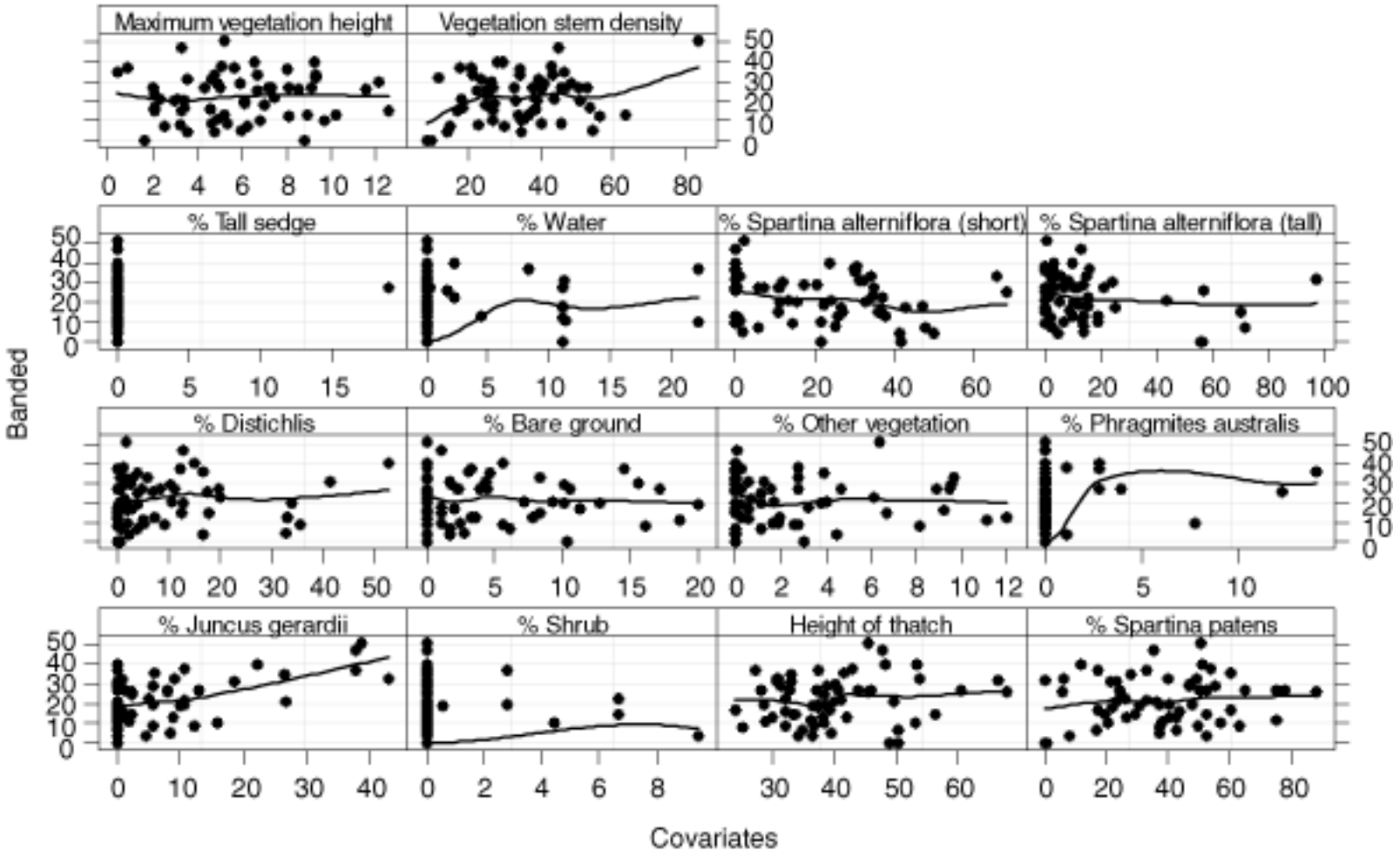
Step 4: Are there lots of zeros in the data?



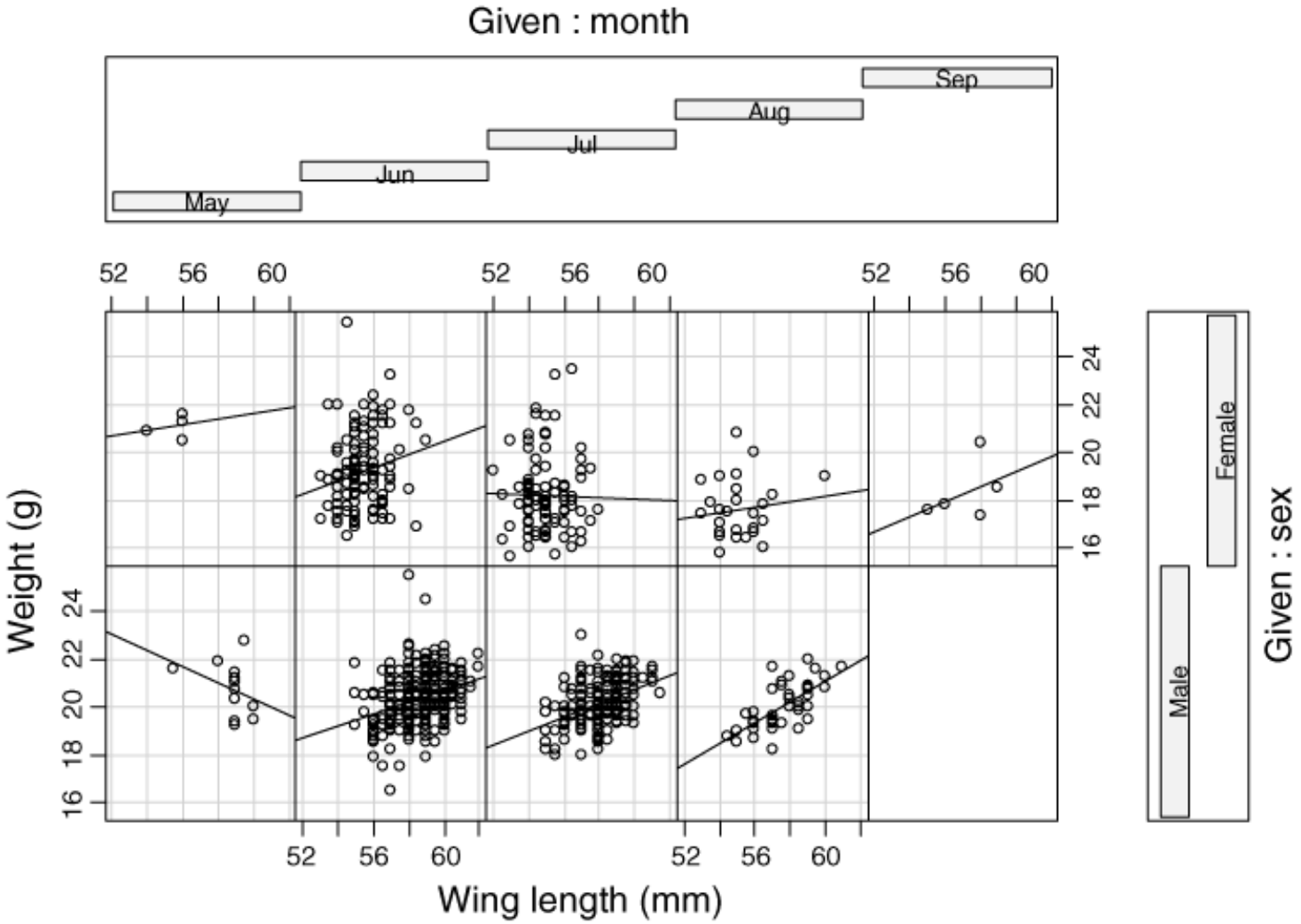
Step 5: Is there collinearity among the covariates?

Covariate	*P*-value (full model)	VIF	*P*-value (collinearity removed)	*P*-value (reduced model)
% *Juncus gerardii*	0·0203	44·9953	0·0001	0·00004
% Shrub	0·9600	2·7818	0·0568	0·0727
Height of thatch	0·9989	1·6712	0·8263	
% *Spartina patens*	0·0640	159·3506	0·3312	
% *Distichlis spicata*	0·0527	53·7545	0·2538	
% Bare ground	0·0666	12·0586	0·8908	
% Other vegetation	0·0730	5·8170	0·9462	
% *Phragmites australis*	0·0715	3·7490	0·2734	
% Tall sedge	0·2160	4·4093	0·4313	
% Water	0·0568	17·0677	0·6942	
% *Spartina alterniflora* (short)	0·0549	121·4637	0·2949	
% *Spartina alterniflora* (tall)	0·0960	159·3828		
Maximum vegetation height	0·2432	6·1200		
Vegetation stem density	0·7219	3·2064		

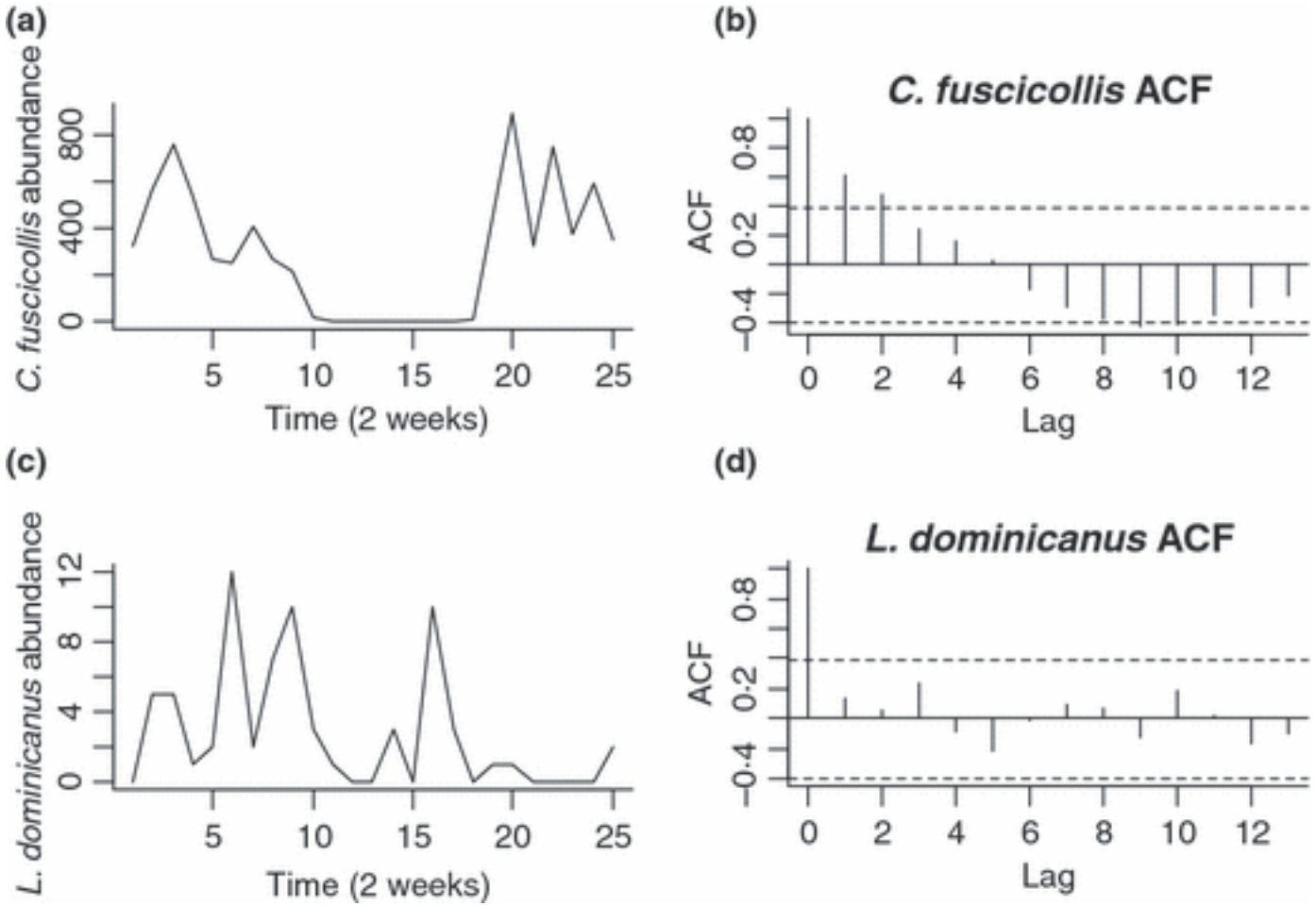
Step 6: What are the relationships between Y and X variables?



Step 7: Should we consider interactions?



Step 8: Are observations of the response variable independent?



What about the code?

- All the code needed to make all the figures that are presented here is part of the supplementary material of the paper!

The next step: Model selection

Review TRENDS in Ecology and Evolution Vol.19 No.2 February 2004

Model selection in ecology and evolution

Jerald B. Johnson[1] and Kristian S. Omland[2]

[1]Conservation Biology Division, National Marine Fisheries Service, 2725 Montlake Boulevard East, Seattle, WA 98112, USA
[2]Vermont Cooperative Fish & Wildlife Research Unit, School of Natural Resources, University of Vermont, Burlington, VT 05405, USA

Alternatives

Model selection method	Calculation[a]	Elements	Refs
Adjusted R^2	$R^2_{adj} = 1 - \frac{RSS/n-p-1}{\sum(y_i-\bar{y})^2/n-1}$	Fit	[7]
Likelihood ratio test	$\mathrm{LRT} = -2\{\ln[L(\hat{\theta}_p\|y] - \ln[L(\hat{\theta}_{p+q}\|y)]\} \sim \chi^2_q$	Fit and complexity	[7]
Akaike information criterion (AIC)	$\mathrm{AIC} = -2\ln[L(\hat{\theta}_p\|y] + 2p$	Fit and complexity	[3]
Small sample unbiased AIC (AIC_c)	$\mathrm{AIC}_c = -2\ln[L(\hat{\theta}_p\|y] + 2p\left(\frac{n}{n-p-1}\right)$	Fit and complexity (with bias correction term for small sample size)	[3]
Schwarz criterion	$\mathrm{SC} = -2\ln[L(\hat{\theta}_p\|y] + p \cdot \ln(n)$	Fit, complexity, and sample size	[10]