# ANOVA & Regression, one big happy family: Introduction to General Linear models Part II

SMEE
ZOL851
Sept 20th 2012

$$y \sim N\left(\beta_0 + \beta_1 x, \sigma^2\right)$$

# goals

- How do we interpret the co-efficients (and other output) for linear models.
- Reading the tea leaves (diagnostics)
- RRRRRRRRRR!!!!!!
- Interaction terms in linear models.

## Team Tinbergen

Marta<br>
Julie
Jeff
John

## Team von Frisch

Amanda
Nora
Masoud

## Team Fisher

Byron
Anne
Angela
David

## Team Hutchinson

Jory (Jorden)
Alexandra (Alex)
Nicki (Cybil)
Brandon

## Team MacArthur

Randy
Susan
Melanie
Dustin

## Team McClintock

Cory
Jake (Jakob)
Eric
Anna

# Readings for linear models

Readings for linear models continued

(These provide very different perspectives on model checking)

Required:

The R Book pages 323-335, 339-353, 356-363

If you want to read about contrasts 368-374, 377-386.

Gelman and Hill Appendix A

Dalgaard pages 218 – 224 (background review, useful summary if necessary. I would read this before the readings in the R book).

# Optional Readings for linear models

It would also be advisable to have PDFs of The R book Chapters 10-12 to aid you as you build models (probably too much reading to do, but at least know where things are).  Chapter 11 goes through MANY ANOVA designs, while chapter 12 covers a bit on ANCOVA.

# Readings for Tuesday: Effect sizes.

- Posted in the effect size folder on ANGEL.
- The first two papers are “assigned” readings, with the others being optional, but these relate to wildlife management, forest ecology and a few others.

# Diagnostics for general linear models

- Histogram or density plot of residuals (try different numbers of bins).
- QQ-Plot (quantile-quantile plot)
- Applies to ANOVA & regression

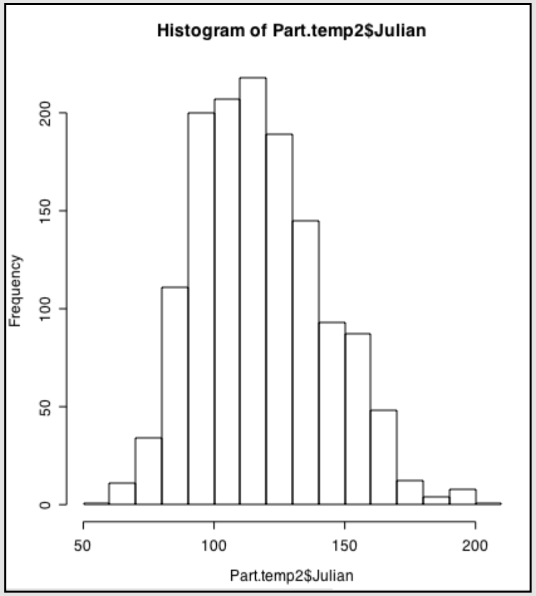
plot(lm.object) provides some basic model assessment tools, but they are not often sufficient.

# There are two separate goals that we do simultaneously

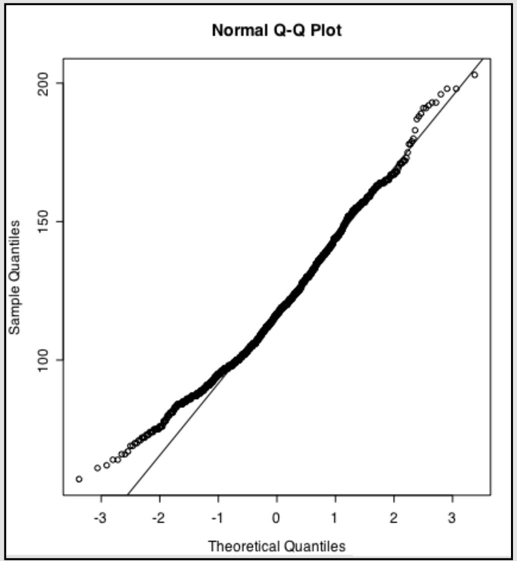
- Model criticism (and evaluating fit)
- Model checking (of assumptions).

- You will notice that Crawley (R book) and Gelman & Hill basically disagree on issues of model checking (and also on on some aspects of model checking). For Thursday tell me why!!!

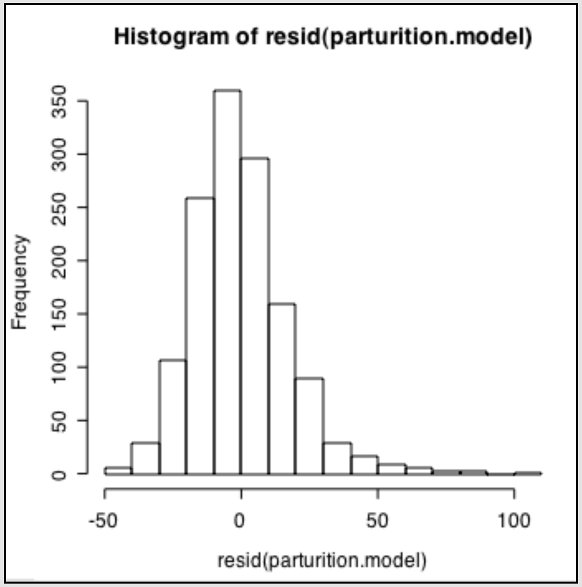
# Normality of Errors



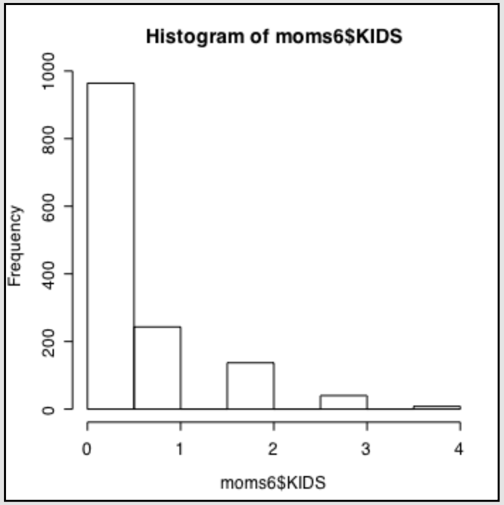
```
> hist(Part.temp2$Julian)
```

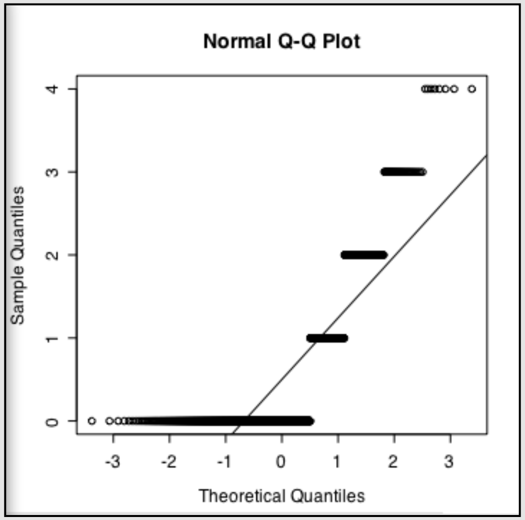


```
> qqnorm(Part.temp2$Julian)
> qqline(Part.temp2$Julian)
```



```
> hist(resid(parturition.model))
```





# Distributions of Errors

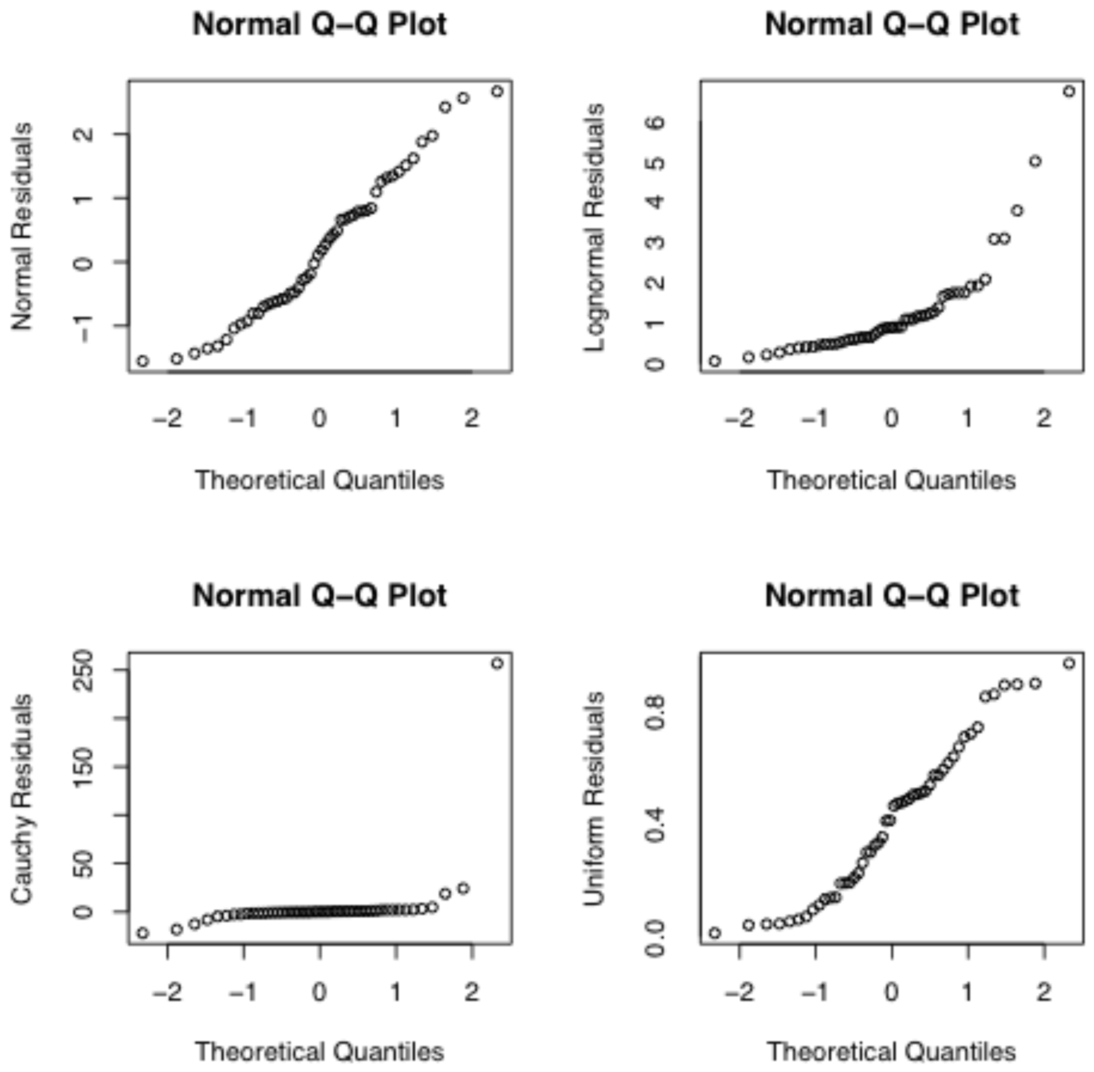


Figure 7.10: QQ plots of simulated data

Platykurtic
(very long tails)

Leptokurtic
(uniform)

- from Faraway, 2002

# If Non-Normal

- As long as you have a peak with reasonable symmetry, you’re probably OK – very robust

- If data are heavily skewed
  - Transform the data (e.g. log)
  - GLiM
  - Non-parametric tests

# Outliers

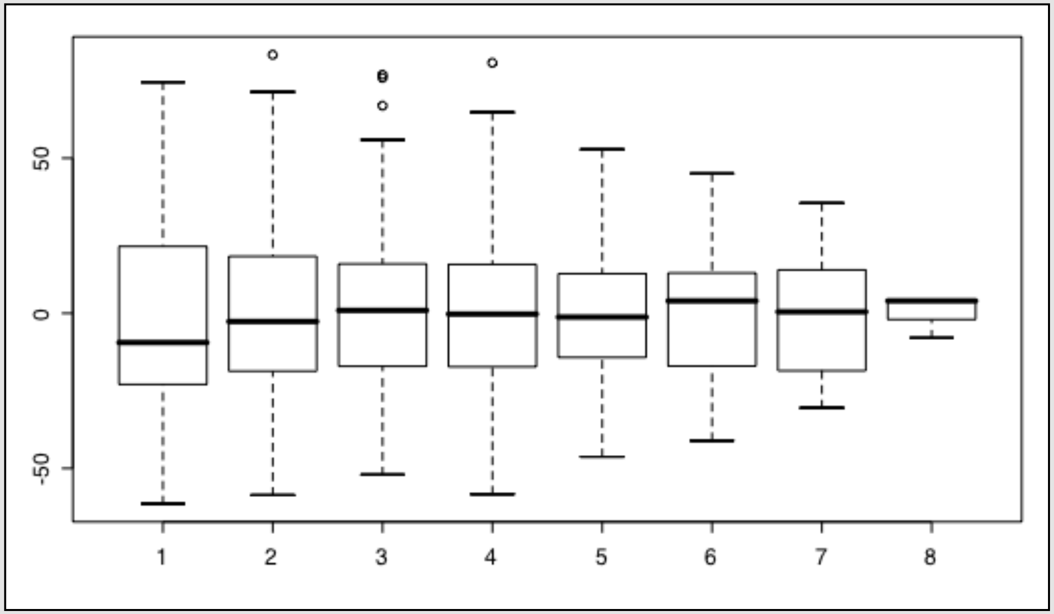
- The normal distribution assumes tails are “small”
    - extreme outliers (~3+ standard deviations out) “shouldn’t” occur
- Can heavily skew estimates
- Detecting outliers
    - Residual plots
    - Outside the whiskers in Box plots
    - Also calculate leverage or influence (Cook’s distance)
        - Degree of effect on coefficients

# What to do about outliers

- Revisit the paper trail for that data point
  - Most often a data entry or other human error
  - Simply correct
- Revisit notes about that site/experiment/data point
- Remove if:
  - Has high leverage and care about estimating parameters
  - Obvious experimental issue
- **If you remove data it is unethical to fail to report this**
  - **OK to report and explain why**

# ANOVA – other assumptions

- Only check for heteroscedasticity
- Use box plot
- Levene's test – take absolute value of ε, then do ANOVA on these (see script)



# Residual plots for regression

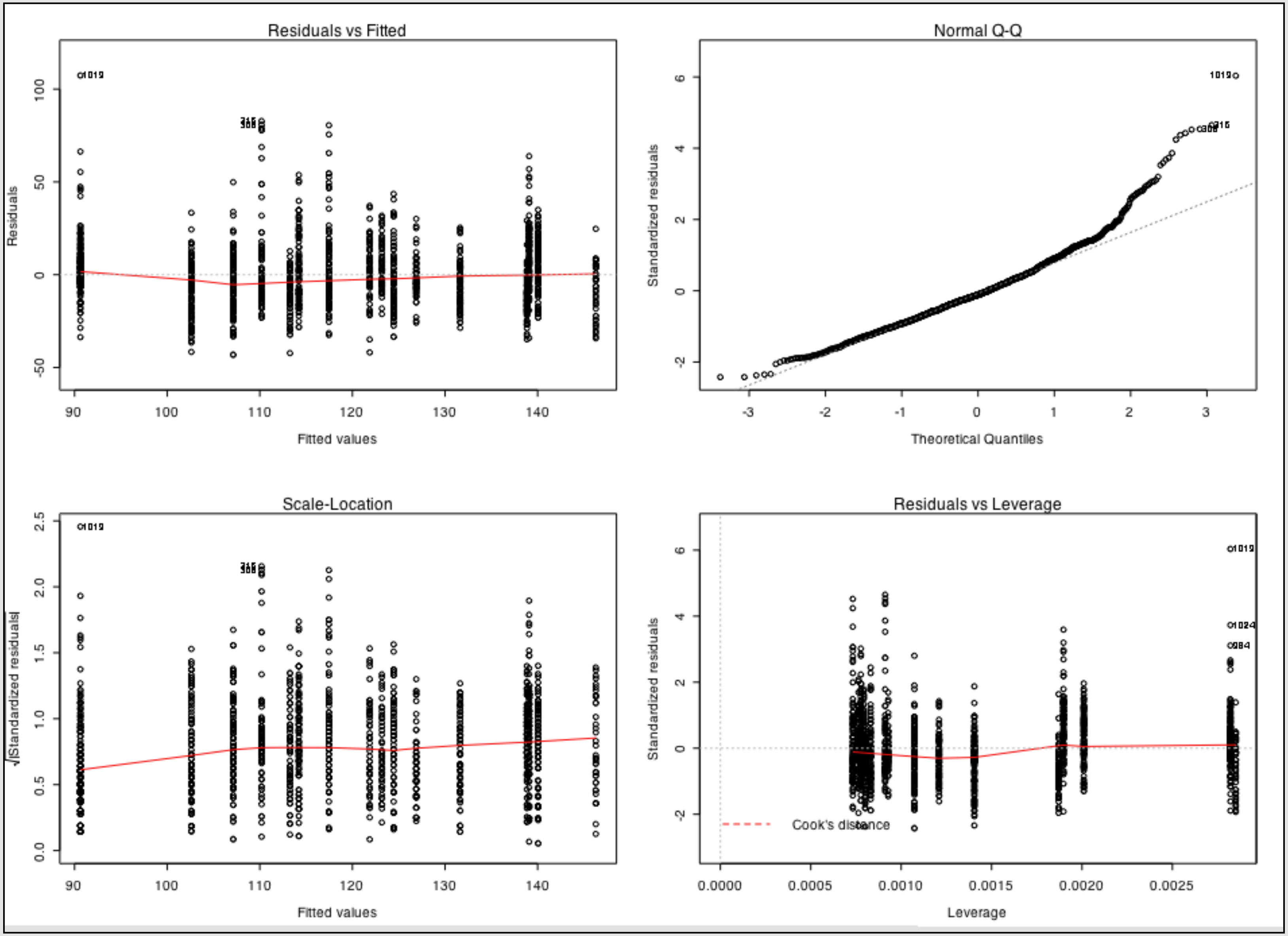
Check 3 other assumptions

- Independence of error terms
- Homoscedasticity
- Linear model appropriate

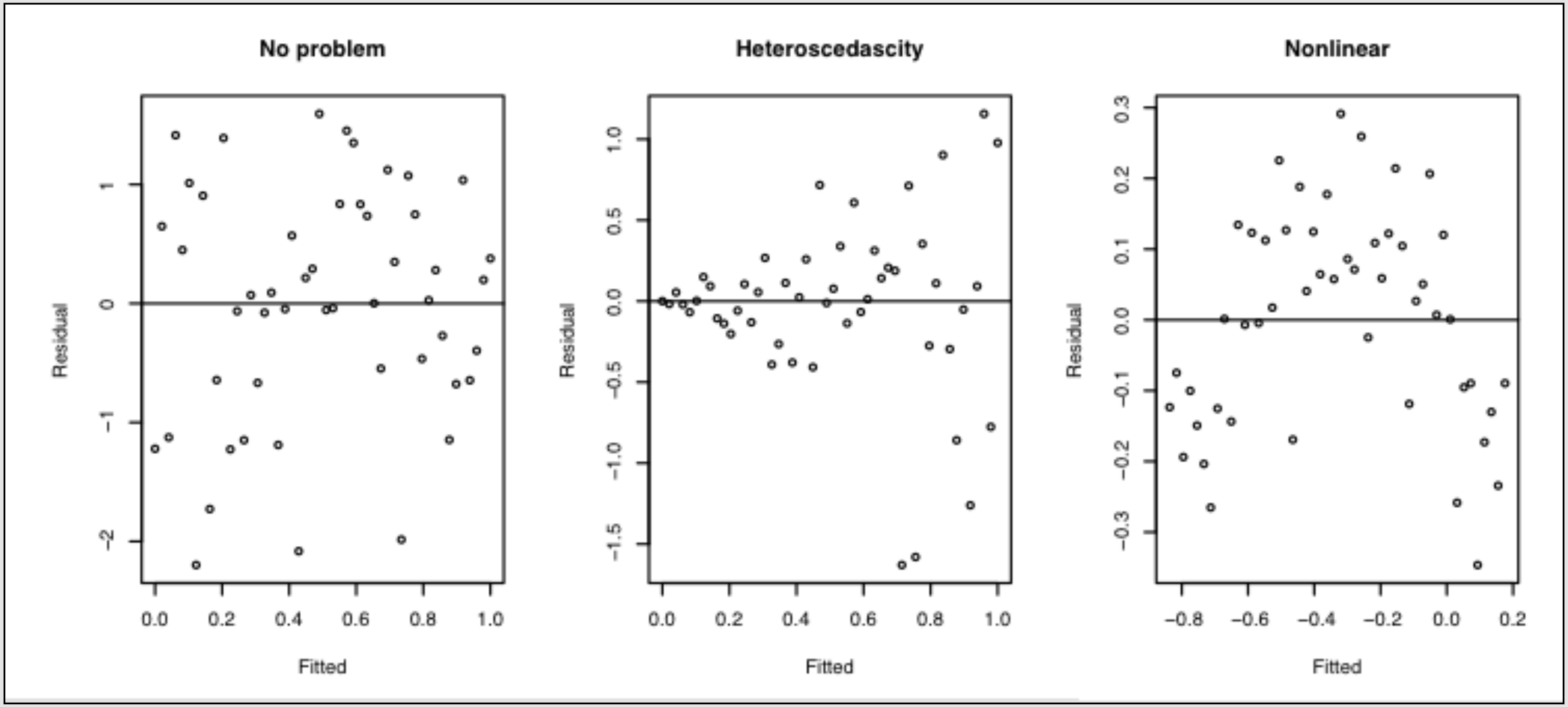
- Two main plots
  - $\varepsilon_i$ vs. predicted y
    - Detect heteroscedasticity, nonlinearity, some independence

If you have an lm or glm fit a call to plot (your.model) will produce some summary plots

```
> plot (parturition.model)
```



# Inspection of Residuals

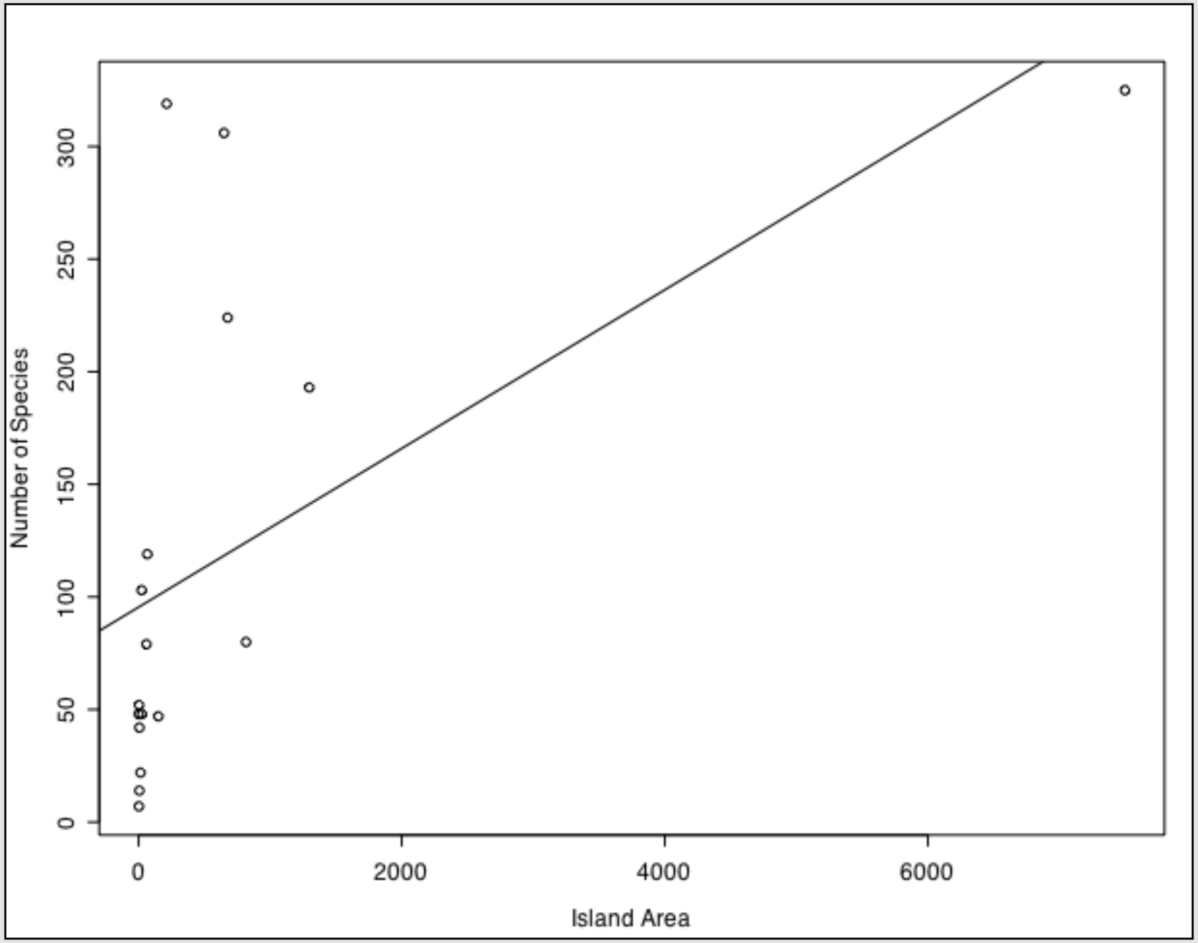


- from Faraway, 2002

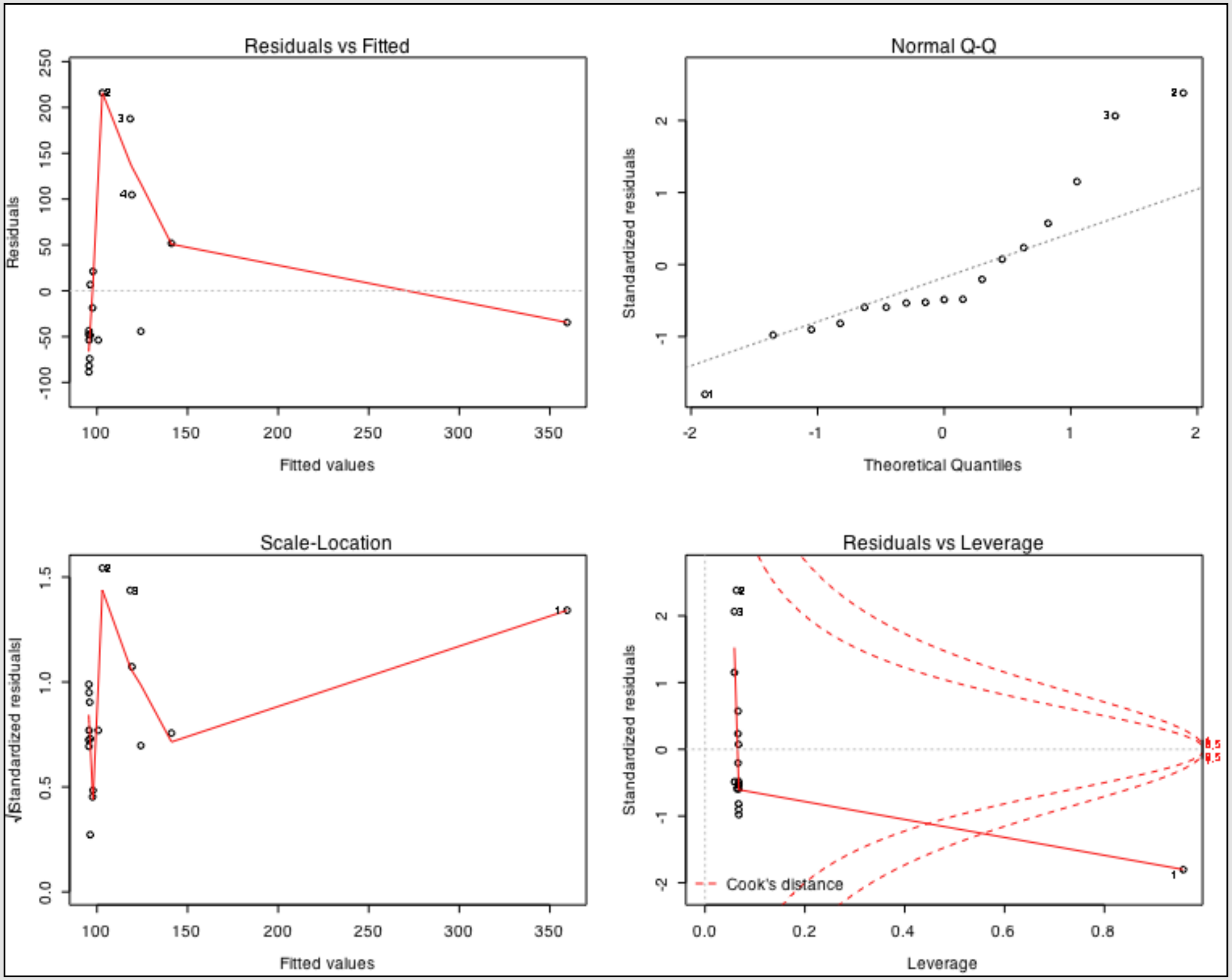
# If errors are heteroscedastic…

- ANOVA is robust *if* design is nearly balanced
- Regression moderately robust, but one side of regression figures more heavily into estimates - bias
- In some cases transformation will fix the problem

# Species-Area Relationships



Number of species on 17 Galápagos Islands (data from Preston 1962)



Residuals vs Fitted

Normal Q-Q

Scale-Location

Residuals vs Leverage

# Is this an outlier?

# Diagnostics-

- Error diagnostics ( (un)equal variances, independence, normality).

  ie. iid and ~N(mean,var)
- Model (partial regression, partial residuals)
- Unusual observations
- Predictors: errors in predictors, Collinearity amongst the predictor

# Other model checking tools

- resid ~ covariates (patterns of non-linear covariation)
- acf(resid) - (is there autocorrelation)
- rstandard() - gives standardized residuals
- rstudent() - ( leave one observation out residuals).
- dffits() – how do observations influence the fitted value
- dfbetas() – (leave one out influences on parameter estimates)
- cooks.distance() – joint measure of dfbetas

# Model criticism

- Of course none of the tools we just looked at tells you how well the model actually fits your data.
- This is a point that Gelman and Hill make a lot.
- plot( response ~ fitted(your.model) ) to look at this. Often more useful than just the $R^2$

# Sums of Squares

- Three general ways of calculating SS
- Important when design is unbalanced
  - unequal number of reps per category (almost all observational studies)
- Terminology comes from SAS originally but is wide-spread now

# Type I, II, and III

- Type-I
  - Significance is assessed sequentially
  - Effect adjusted for all previous terms but not any subsequent terms
  - Order matters
  - i.e. *y ~ a+b* is not the same as *y ~ b+a*
  - In case of y ~ a+b+c+a:b, significance of c is assessed after adjusting for effects of a and b but not a:b
  - the base anova() in R uses this approach, and is designed for testing SEQUENTIAL EFFECTS (or between series of nested models).
- Type-II
  - Significance of each term is assessed against all other terms except its higher order relatives
  - Order doesn’t matter
  - In model y ~ a+b+c+a:b+a:c, the significance of b is assessed after adjusting for a and c and a:c but not a:b
- Type-III
  - Adjusts for all other effects in the model (hard to interpret).

Need to use Anova() in the “car” library for typeII or typeIII

# A word of caution about comparing full and reduced models in R (or any stats package)

- Be careful about how you deal with NA's. You want to make sure you are using the same set of observations for both models.
- #The comparison between two or more models will only be valid if they are fitted to the same dataset. This may be a problem if there are missing values and **R**'s default of `na.action = na.omit` is used.

# Interactions

- Until now we have considered only additive effects.
- Linear models can include nonadditive terms - interactions
- Interaction: the effect of one covariate is not constant but depends on the level of some other covariate

# Types of Interactions (A:B)

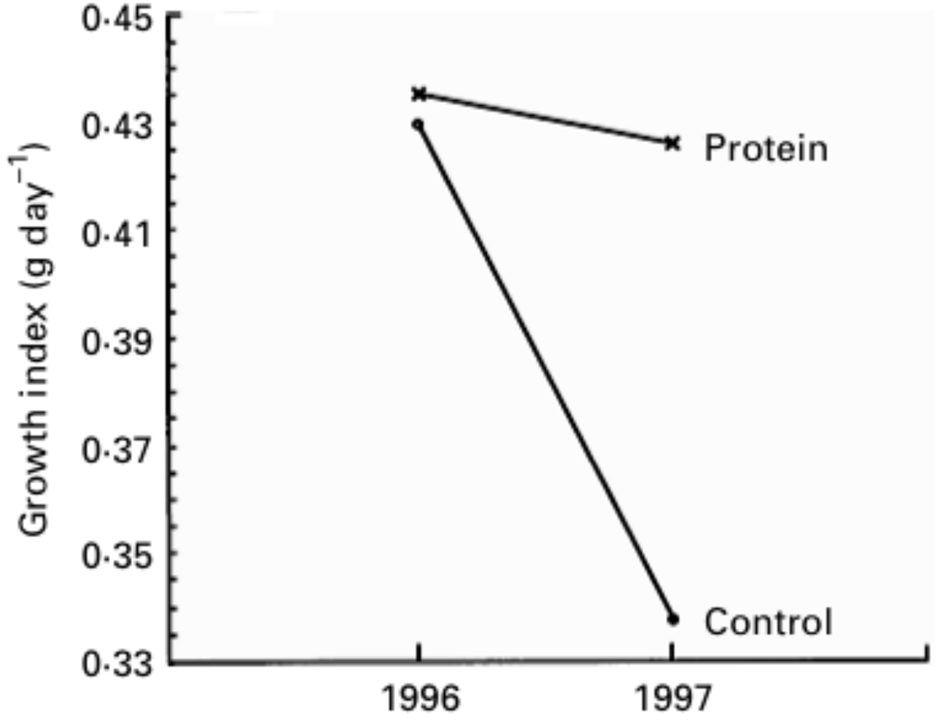


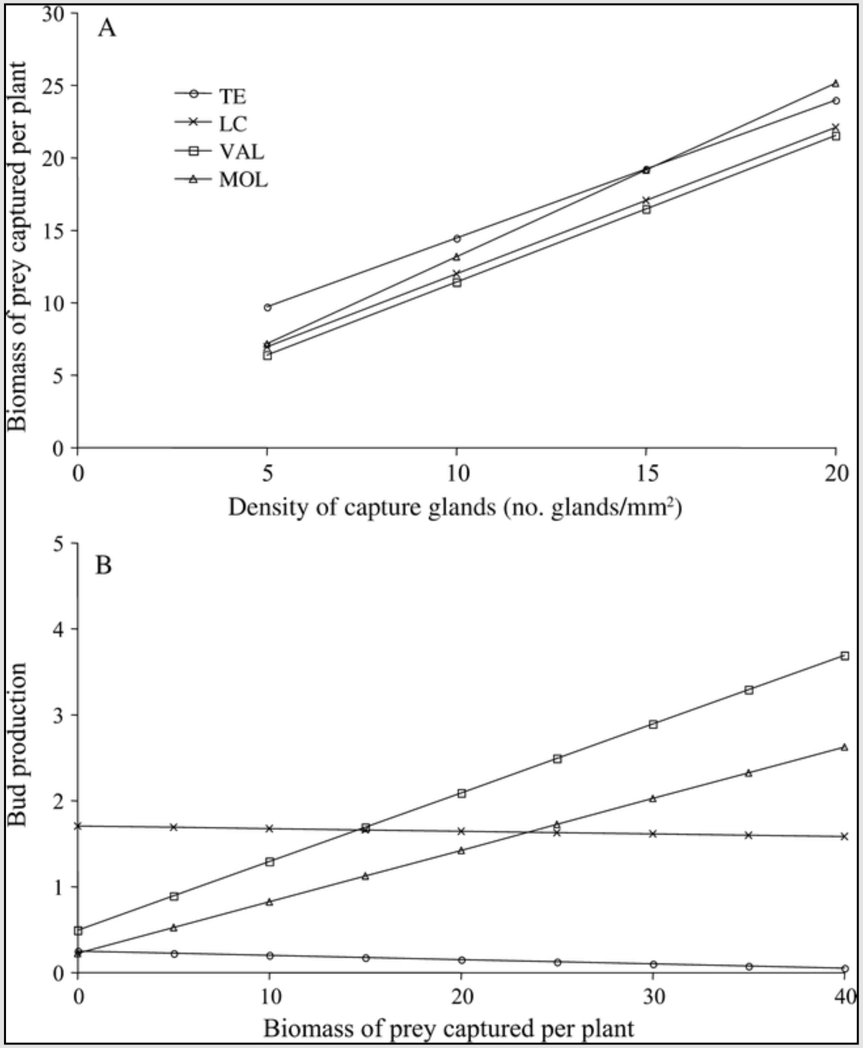
**Fig. 1.** Significant two–way interaction ($F_{1,5} = 9{\cdot}12$, $P = 0{\cdot}029$) of nestling growth indices (g day$^{-1}$) for young-of-the-year *Peromyscus maniculatus* in protein supplemented (×) and control populations (o) during the breeding seasons of 1996 and 1997.

$A_{Factor}$:$B_{Factor}$

- Effect of level 1 for Factor A depends on level of Factor B

- from McAdam & Millar, 1999

# Types of Interactions (A:B)

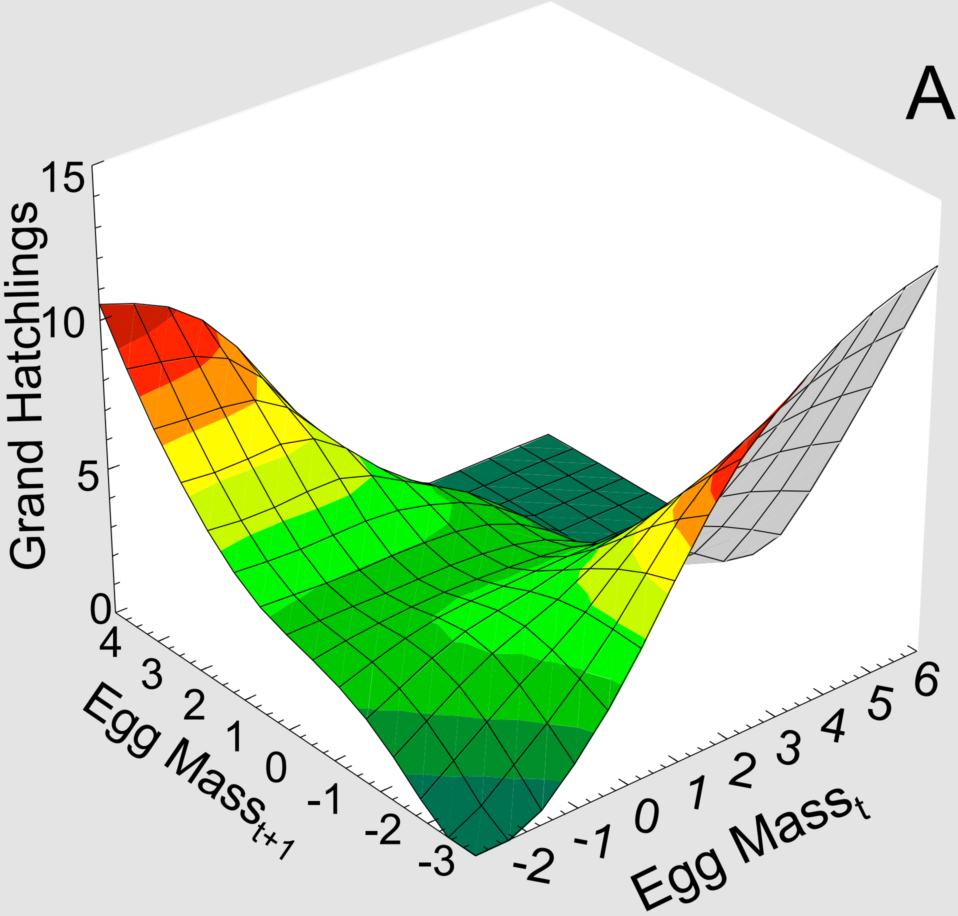


$A_{\text{Continuous}}:B_{\text{Factor}}$

- Slope for continuous variable A depends on level of factor B

- from Alcala & Dominguez, 2005 Ecology, 86:9

# Types of Interactions (A:B)



$A_{\text{Continuous}}$:$B_{\text{Continuous}}$

– Effect of continuous variable A if variable B held constant…

  but slope of relationship for A depends on the value for variable B

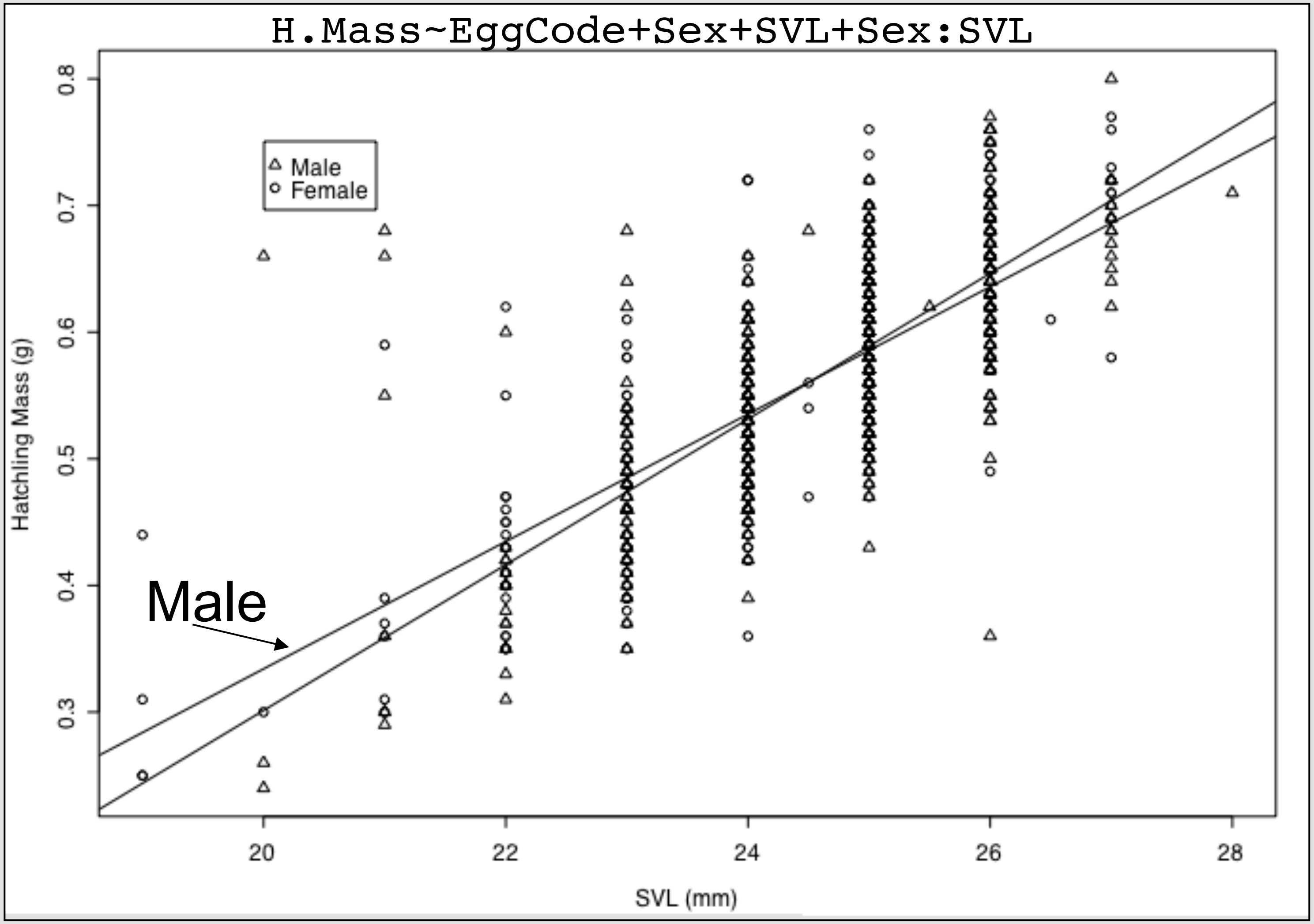
# Interactions in R

- Interactions are coded in R using :
- The notation is the same for factors and continuous predictors

```
object<-lm(y~a+b+a:b, data=data.temp, na.action=na.omit)
# The Same model fit
object<-lm(y~a*b, data=data.temp, na.action=na.omit)
```

a*b is shorthard for the model a+b+a:b

# Example: Effects of sex and body size (mm) on hatchling mass (g)

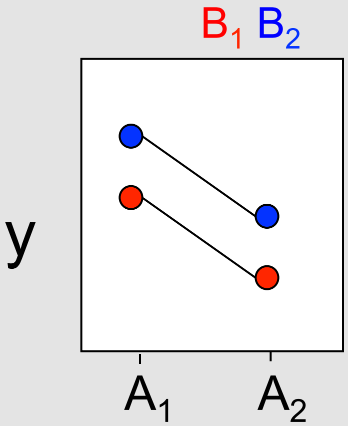


# Interpreting Interactions

- When interactions are significant, then difficult to interpret main effects or lower order interactions
    - e.g. Difficult to interpret significance of a in:

      y = a + b + a:b when a:b is significant
- Interpreting interactions when a component main effect has been removed can be misleading
    - i.e y = a + a:b is dangerous
- Always difficult to interpret higher order interactions biologically
    - Try breaking into categories to make sense of the results

# No Interaction



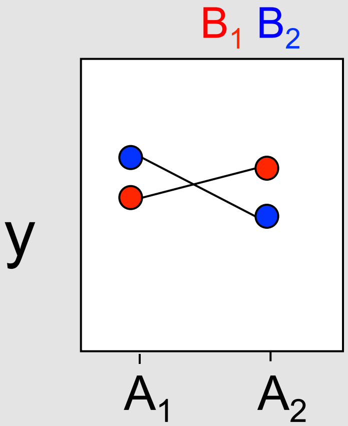
A = *

B = *

A:B = ns

Interactions are fundamentally about having different slopes. The lines don’t have to cross, as long as slope different, interaction.

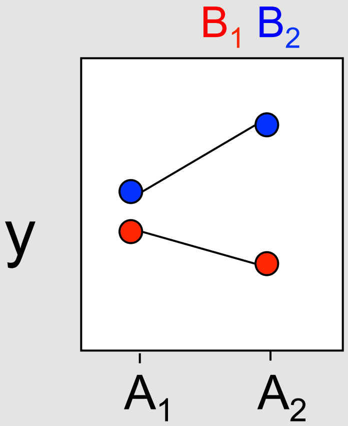
# Interaction



A = ns

B = ns

A:B = *

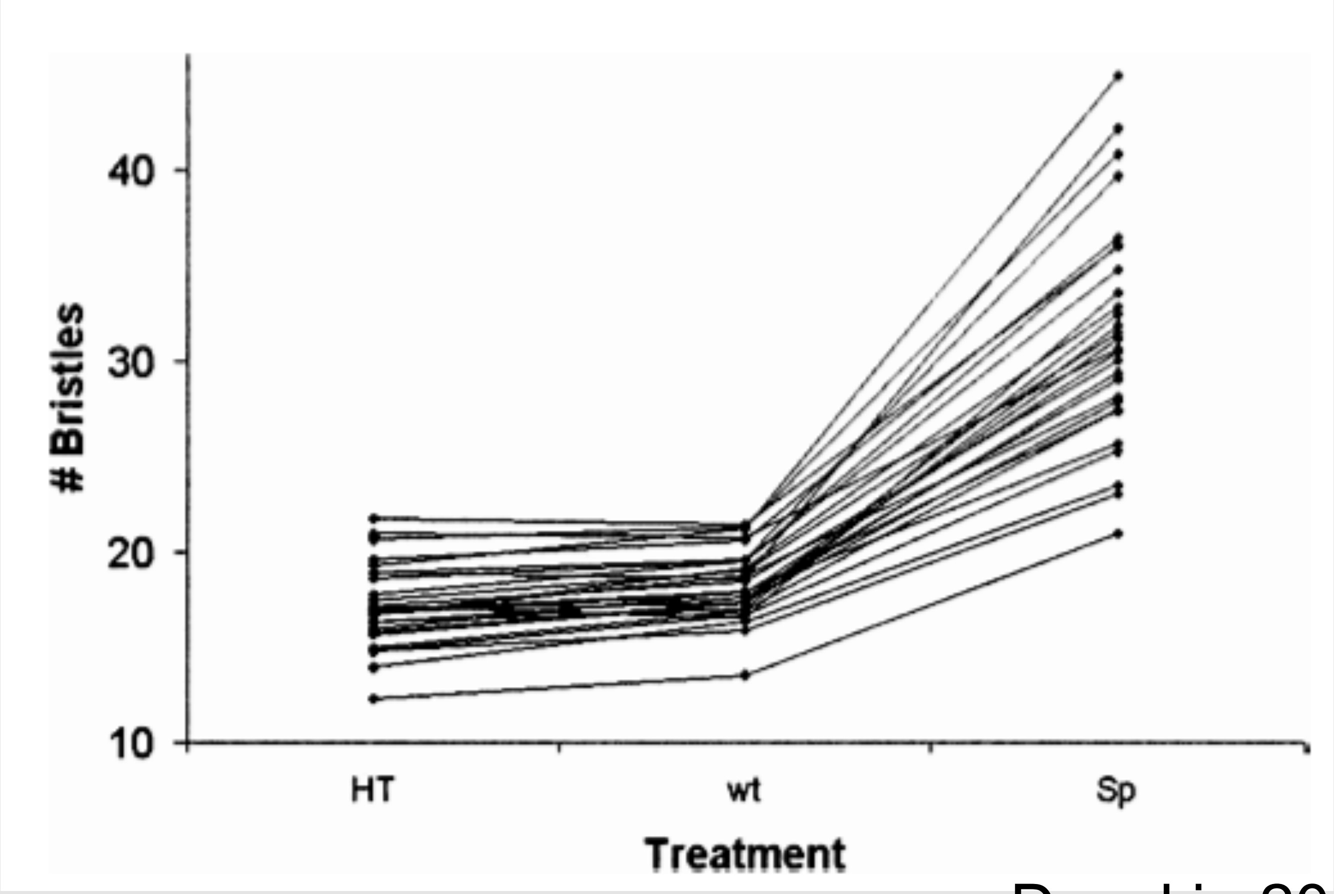


A = ns

B = *

A:B = *

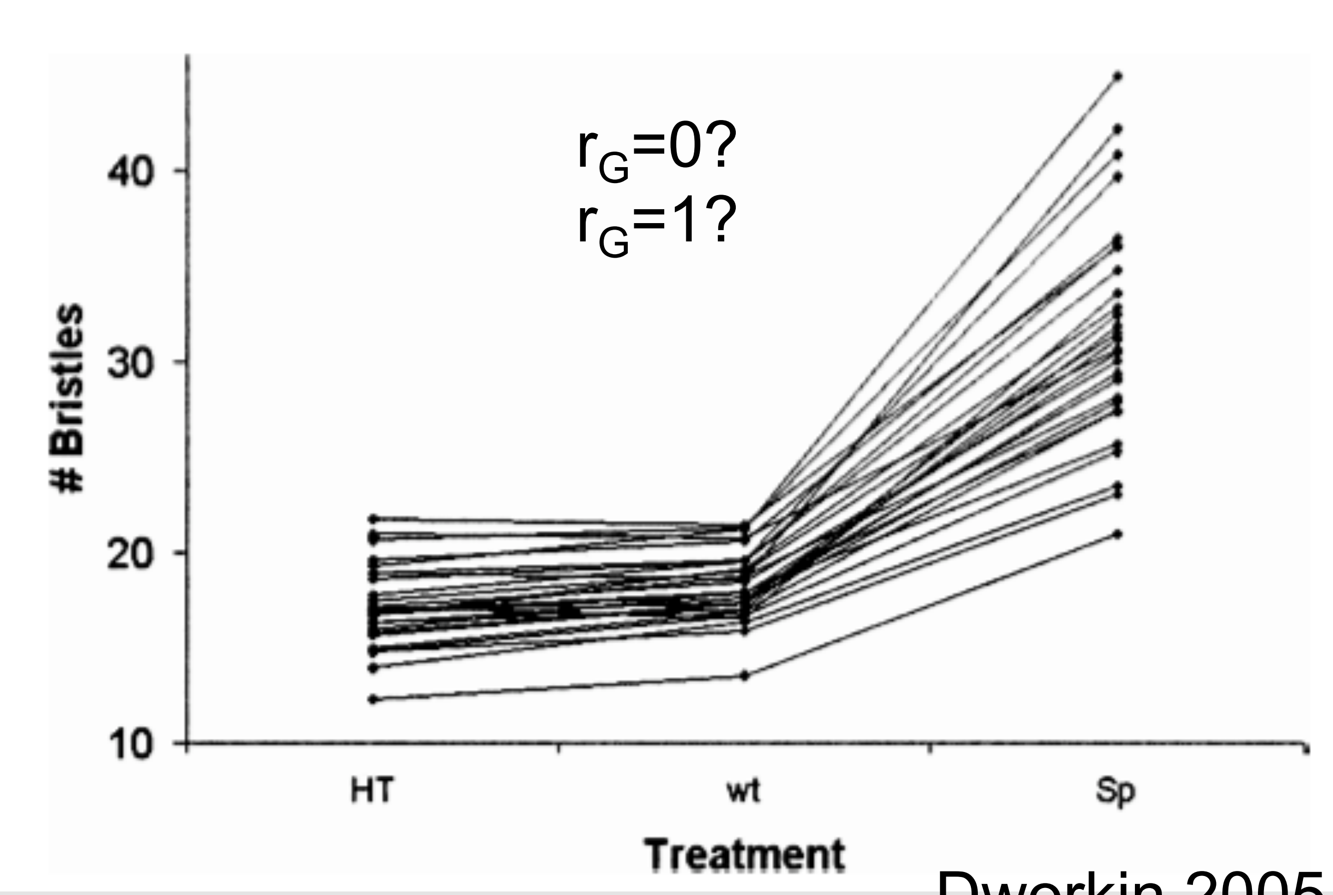
Making sense of interactions: What do you do when you have some crossing of reaction norms and some changes in scale? This type of data appears all of the time in population ecology, evolutionary ecology and quantitative genetics.



Dworkin 2005 Evolution

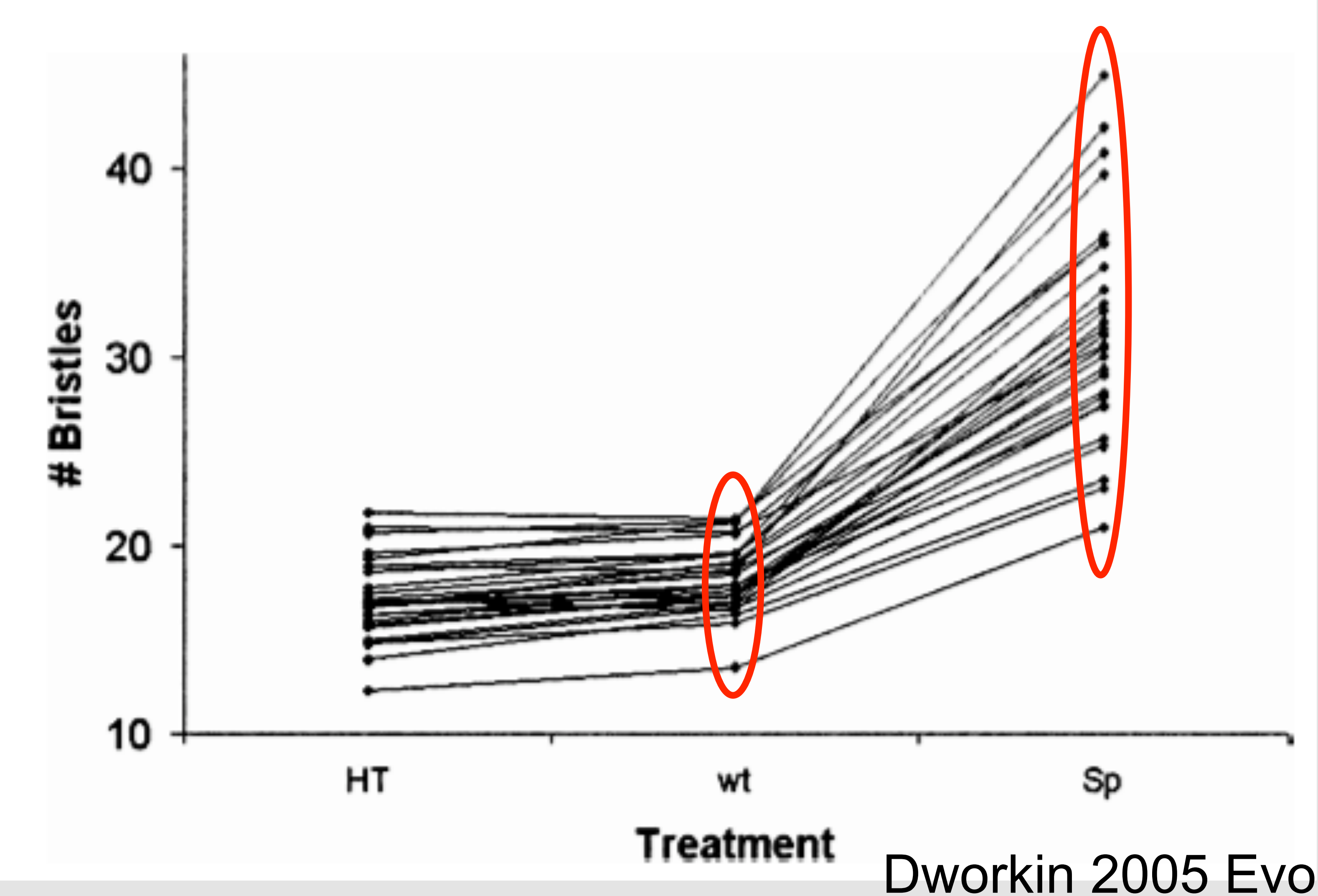
Two aspects of the interaction should be noted. First, how do the estimates across treatment levels co-vary?

How correlated are these points? Do the observations tend to be correlated by the environment? What does it mean to have a correlation = 1?



Dworkin 2005 Evolution

Second, Is there a change in the amount of variance in the estimates at different treatment levels.



Dworkin 2005 Evolution

# Issues in Regression

- Collinearity in predictors
- Model II Regression
  - Regression when there is error in X

# Multiple Regression

- The central complication of linear models is that the independent/predictor/covariates variables are correlated with each other
  - Less noticed in ANOVA because usually factors are "orthogonal" – i.e. uncorrelated
    - Not always true - unbalanced designs **ARE** a type of collinearity
- Also called "multicollinearity"
- This correlation creates many challenges
- If the variables were not correlated at all, then doing a multiple regression would give the same results as a series of simple regressions onto each independent variable.

# Bouncing $\beta$'s

- Extreme case – two identical columns
  - Sometimes shows up – mass, volume*density
- There is no correct answer for $\beta_1$ vs $\beta_2$
- Could be: 3/0, 2/1, 1.5/1.5, 0/3, etc
- Slight changes in error terms cause drastic shifts in betas

| … | $x_1$ | $x_2$ | … |
|---|---|---|---|
| … | 2 | 2 | … |
| … | 4 | 4 | … |
| … | 3 | 3 | … |
| … | 1 | 1 | … |

# Real World

- Rarely this extreme, weaker correlations (r=0.3-0.9) very, very common
  - Temperature, growing season length
  - Body size, speed
  - % forested landscape, % spruce

- Also multicollinearity where:
  - Total trees = # oak + # walnut + # maple + # ash
  - Covary negatively and more than just simple correlations

# Common Symptoms

- Large correlations in predictors
- Theoretically or previously important predictors have high p-values
- Coefficients have the wrong sign
- Individual betas not significant but overall F of regression model is significant (and high $R^2$).
- Deletion of one column or row causes the betas to change drastically (i.e. removing one observation changes the estimates wildly).

# Why you should care about collinearity

- Instability in estimates of βs
- Increases the Standard errors for individual βs.

# How to inspect for collinearity between independent variables.

- Print/plot the covariance matrix

```
plot(data.frame)
cor(data.frame)
```

This helps you look for collinearity, but it is hard to  visually inspect for multi-collinearity.

# More Rigorous Tests

- The real question is whether $X^TX$ is singular, where X is the design matrix (which may be a set of continuous variables, not just dummy variables).
- When $X^TX$ is singular it indicates 'exact' (multi)collinearity.
- You can examine the eigenvalues of the design matrix

```
eigen(t(X) %*% X)
```

Where small eigenvalues, close to 0 suggest a problem. As do very unequal eigenvalues.

# More Rigorous Tests

Alternatively, the condition number ($\kappa$) is defined as

$$\kappa_P = \sqrt{\frac{\lambda_1}{\lambda_P}}$$

Where $\lambda_1$ is the dominant (1st eigenvalue), and P represents the Pth (final) eigenvalue. $\kappa > 30$ is considered large, indicating multicollinearity.

```r
mod.X <- model.matrix(lm.3)
eigen.x <- eigen(t(mod.X) %*%mod.X)
eigen.x$val # eigenvalues from the design matrix
sqrt(eigen.x$val[1]/eigen.x$val) # condition numbers
```

# More Rigorous Tests

- Calculate tolerance or variance inflation factors (VIF)
  - $\text{Tolerance}_k = 1 - R^2_k$
    - Where $R^2_k$ is the $R^2$ of $x_k \sim b_0 + x_1 + x_2 + \ldots$
  - $\text{VIF}_k = 1/\text{tolerance}_k$
    - One per column, large values indicate multicollinearity of that column
    - sqrt(VIF) is degree that $\beta$se is inflated by collinearity
- Estimates the "multiplier" for the standard error of $\text{predictor}_j$ relative to the same model with no collinearity between the $\text{predictor}_j$ and the other predictors.
- Rules of thumb: Collinearity an issue when individual tolerance < 0.1 or VIF >10.
- Also an issue if average VIF > 6.

# More Rigorous Tests

- There is a *vif()* function in the *car* package
  ```
  vif(lm.object)
  ```

- Also a *vif()* function in faraway package that functions on the design matrix

# Solutions

- Eliminate similar variables using *a priori* thinking:
  - Mean temp & growing season are obviously correlated, which one do you expect to matter more, throw the other out
  - In extreme case variable might be entirely redundant so no information is lost
- Embrace it
  - Explore the structure with the correlation matrix
  - Use ordination/principle components analysis to combine related sets of variables

# Summary of the Problem

1. Computation -parameters difficult or impossible
2. Assessment - coefficients can sometimes appear NS when they are important
3. Reliability - small changes in dataset can result in large changes in parameters
4. Inference - parameters represent the effect of x while holding other variables constant.
   - When x1 and x2 covary, the effect of x1 while x2 is held constant does not exists - requires extrapolation to region where we have no data!