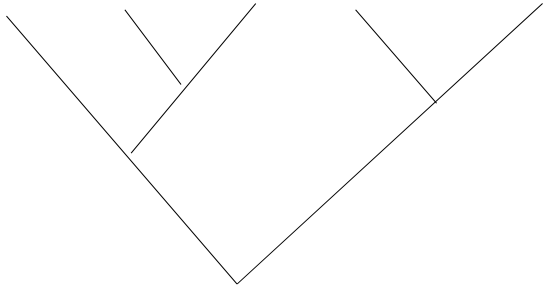
Phylogeny and GLS

Comparative studies

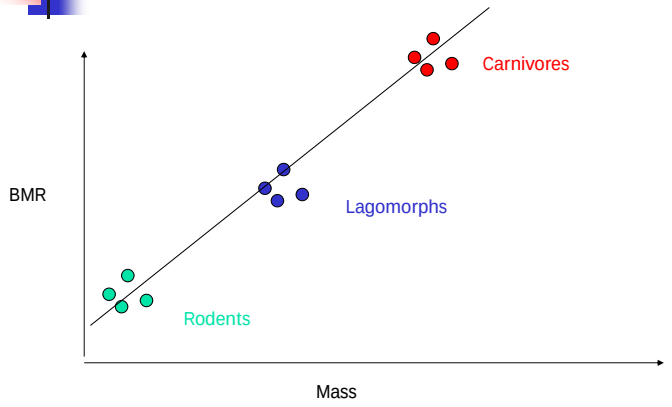
- A comparative study is when you look at data across many species
- A subclass of observational (not experimental)
- Useful for many evolutionary and ecological questions
- Harvey & Pagel
  - "The comparative method in evolutionary biology" 1991 Oxford University Press

Are these points independent?

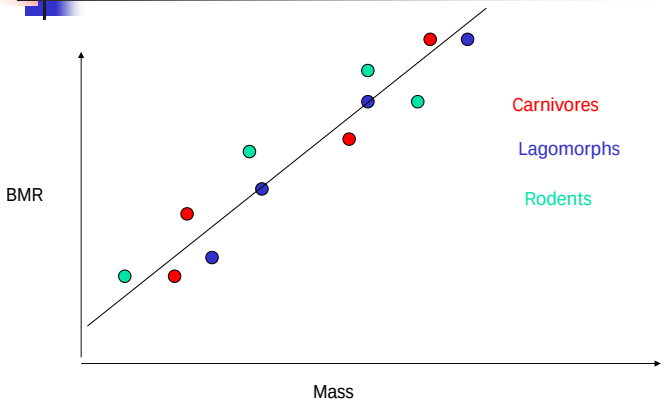
BMR	400	283	912	445	327
Mass	10	15	18	27	13



Non-independent – how many points do I really have?



Independent – how many points do I really have?



Enormous controversy

- Was a period you couldn't publish comparative analyses without correcting for phylogeny
- Very heated debate in pages of Ecology, elsewhere
  - Some said this was wrong
  - Ecology trumps evolution
- In some cases PIC is just a fancy way to lose power, but other times it is important

Importance of phylogeny

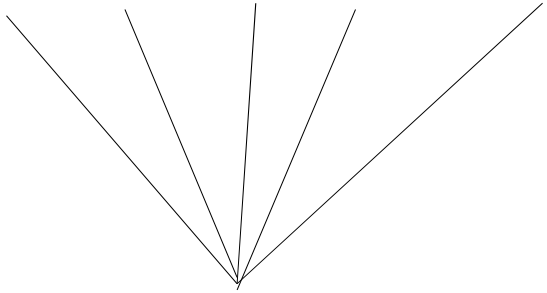
- Multiple reasons for non independence
  - Varying degrees of time of evolution apart
  - Genetic/phenotypic constraints
  - Ecological syndromes (once a grassland species, always a grassland species)
- Depends on:
  - How rapidly traits evolve relative to taxonomic levels studied
  - Only one trait is phylogenetically derived (e.g. other is environmental)
  - Emphasis on ecological or evolutionary questions

Methods of correcting

- Use taxonomy in nested analysis:
  - Family/genus/species
- Phylogenetically independent contrasts (PIC)
  - Felsenstein 1985 AmNat
- GLS
  - Generalization of GLM to allow ε_i to covary
- Garland and Ives show two methods identical

Assume independence =polytomy?

BMR	400	283	912	445	327
Mass	10	15	18	27	13



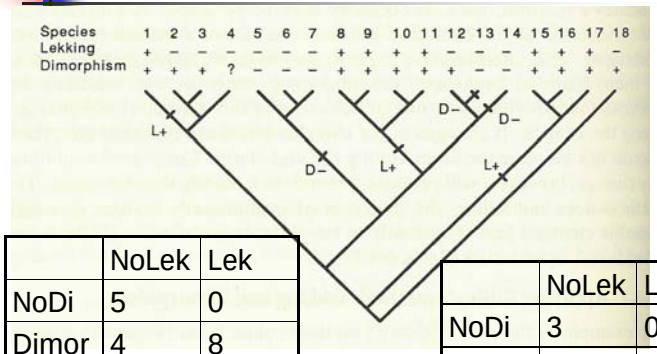
PIC on continuous

BMR	400	283	912	445	327
Mass	10	15	18	27	13

PIC continued

- N tips produces n-1 points
- These points can be plugged into a regression (with no intercept – zero forced)
- Can't use to predict (no intercept), but get correlation, slopes (same as w/o PIC) & significance

PIC and discrete data



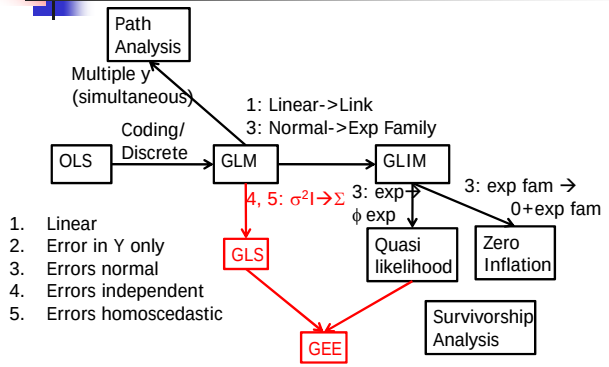
	NoLek	Lek
NoDi	5	0
Dimor	4	8

	NoLek	Lek
NoDi	3	0
Dimor	2	3

Discrete continued

- Called Ridley's method
- Maddison's method allows imputation of directionality (a causes b)
- Pagel & Harvey have a more general method that also includes branch lengths

Road map



14

Covariance matrices

- Recall $\sigma_{xy}=cov(x,y)=\Sigma(x_i-x_{mean})(y_i-y_{mean})$
  - $cov(x,y)=var(x)=\sigma_x^2$
  - $cor(x,y)=\sigma_{xy}/\sigma_x\sigma_y$
  - In matrices $\rho=(V^{1/2})^{-1}S(V^{1/2})^{-1}$
- What if $x_1,x_2,x_3,x_4,...$?
- Covariance matrix

$$\begin{bmatrix} \sigma_{11}=\sigma_1^2 & \sigma_{12} & \sigma_{13} & \sigma_{14} & \sigma_{15} \\ \sigma_{21} & \sigma_{22}=\sigma_2^2 & \sigma_{23} & \sigma_{24} & \sigma_{25} \\ \sigma_{31} & \sigma_{32} & \sigma_{33}=\sigma_3^2 & \sigma_{34} & \sigma_{35} \\ \sigma_{41} & \sigma_{42} & \sigma_{43} & \sigma_{44}=\sigma_4^2 & \sigma_{45} \\ \sigma_{51} & \sigma_{52} & \sigma_{53} & \sigma_{54} & \sigma_{55}=\sigma_5^2 \end{bmatrix}$$

Correlation matrix

$$\begin{bmatrix} 1 & \rho_{12} & \rho_{13} & \rho_{14} & \rho_{15} \\ \rho_{21} & 1 & \rho_{23} & \rho_{24} & \rho_{25} \\ \rho_{31} & \rho_{32} & 1 & \rho_{34} & \rho_{35} \\ \rho_{41} & \rho_{42} & \rho_{43} & 1 & \rho_{45} \\ \rho_{51} & \rho_{52} & \rho_{53} & \rho_{54} & 1 \end{bmatrix}$$

The GLM/GLIM assumption

1. Linear
2. Error in Y only
3. Errors normal
4. Errors independent
5. Errors homoscedastic

- #4 - $cov(\varepsilon_i,\varepsilon_j)=0$ if ij
- #5 - $cov(\varepsilon_i,\varepsilon_i)=\sigma^2$ independent of i
- or covariance matrix of errors= $\sigma^2 I$
- i.e. #3-#5 boils down to $\varepsilon_i \sim N(0, \sigma^2 I)$

$$\begin{bmatrix} \sigma^2 & 0 & 0 & 0 & 0 \\ 0 & \sigma^2 & 0 & 0 & 0 \\ 0 & 0 & \sigma^2 & 0 & 0 \\ 0 & 0 & 0 & \sigma^2 & 0 \\ 0 & 0 & 0 & 0 & \sigma^2 \end{bmatrix}$$

Heteroscedastic

- Constant CV → big measurements have more variance (e.g. what is variance of mass for 1kg organism vs. 100 kg organism)
- Variances depend on the x-value (i.e. independent variable)
  - var(y)=$\sigma^2 f(x)$
  - f(x)=exp(δx) #exponential
  - f(x)=$|x|^\delta$ #power
  - F(x)=δ_i #by discrete variable
- More generally WLS (weighted least squares)
  - Weight each point by $1/\sigma_i^2$
  - compare with scaled regression
  - EG some points have more confidence in
- Covariance:

	ε_1	ε_2	ε_3	ε_4	ε_5
ε_1	$\sigma_1^2=\sigma^2 f(x_1)$	0	0	0	0
ε_2	0	$\sigma_2^2=\sigma^2 f(x_2)$	0	0	0
ε_3	0	0	$\sigma_3^2=\sigma^2 f(x_3)$	0	0
ε_4	0	0	0	$\sigma_4^2=\sigma^2 f(x_4)$	0
ε_5	0	0	0	0	$\sigma_5^2=\sigma^2 f(x_5)$

Timeseries/repeated measures

- Observe one variable over time (say population size): $y_1, y_2, y_3, \ldots y_t$
  - Want $y \sim temp_t + precip_t$ but ε autocorrelated
  - Assume $cov(\varepsilon_t, \varepsilon_s)=\rho^{|t-s|}$
  - AR1
  - Use correlation
  - 2 parameters: σ, ρ

	ε_1	ε_2	ε_3	ε_4	ε_5
ε_1	1	ρ	ρ^2	ρ^3	ρ^4
ε_2	ρ	1	ρ	ρ^2	ρ^3
ε_3	ρ^2	ρ	1	ρ	ρ^2
ε_4	ρ^3	ρ^2	ρ	1	ρ
ε_5	ρ^4	ρ^3	ρ^2	ρ	1

Blocking – discrete spatial

- Can treat as a factor
- But can also treat as a source of correlation in errors
- Let τ be within block correlation, σ be measurement standard deviation
  - $\rho=\tau^2/(\tau^2+\sigma^2)$
- EG – 2 sites, 2 measurements in 1st, 3 in 2nd

	$\varepsilon_{1,1}$	$\varepsilon_{1,2}$	$\varepsilon_{2,1}$	$\varepsilon_{2,2}$	$\varepsilon_{2,3}$
$\varepsilon_{1,1}$	1	ρ	0	0	0
$\varepsilon_{1,2}$	ρ	1	0	0	0
$\varepsilon_{2,1}$	0	0	1	ρ	ρ
$\varepsilon_{2,2}$	0	0	ρ	1	ρ
$\varepsilon_{2,3}$	0	0	ρ	ρ	1

	$\varepsilon_{1,1}$	$\varepsilon_{1,2}$	$\varepsilon_{2,1}$	$\varepsilon_{2,2}$	$\varepsilon_{2,3}$
$\varepsilon_{1,1}$	$\tau^2+\sigma^2$	τ^2	0	0	0
$\varepsilon_{1,2}$	τ^2	$\tau^2+\sigma^2$	0	0	0
$\varepsilon_{2,1}$	0	0	$\tau^2+\sigma^2$	τ^2	τ^2
$\varepsilon_{2,2}$	0	0	τ^2	$\tau^2+\sigma^2$	τ^2
$\varepsilon_{2,3}$	0	0	τ^2	τ^2	$\tau^2+\sigma^2$

Continuous Spatial

- Assume $cov(\varepsilon_i, \varepsilon_j)=d_{ij}$ – only a function of distance
- Typically $\rho_{ij}=\exp(-kd_{ij})$
- So estimate σ, k

1	$\exp(-kd_{12})$	$\exp(-kd_{13})$	$\exp(-kd_{14})$	$\exp(-kd_{15})$
$\exp(-kd_{21})$	1	$\exp(-kd_{23})$	$\exp(-kd_{24})$	$\exp(-kd_{25})$
$\exp(-kd_{31})$	$\exp(-kd_{32})$	1	$\exp(-kd_{34})$	$\exp(-kd_{35})$
$\exp(-kd_{41})$	$\exp(-kd_{42})$	$\exp(-kd_{43})$	1	$\exp(-kd_{45})$
$\exp(-kd_{51})$	$\exp(-kd_{52})$	$\exp(-kd_{53})$	$\exp(-kd_{54})$	1

Phylogenetic

- Brownian motion (Felsenstein)
  - $\text{cov}(x,y)=\gamma t_a$
  - t_a is shared time (sum of branch lengths from root to most recent common ancestor)
- Stabilizing selection (Martin)
  - $\text{cov}(x,y)=\gamma \exp(-\alpha t_a)$
  - t_a is shared time (sum of branch lengths from root to most recent common ancestor)
- Two parameters: σ, γ
  - Correlation matrix

γt_{11}	γt_{12}	γt_{13}	γt_{14}	γt_{15}
γt_{21}	γt_{22}	γt_{23}	γt_{24}	γt_{25}
γt_{31}	γt_{32}	γt_{33}	γt_{34}	γt_{35}
γt_{41}	γt_{42}	γt_{43}	γt_{44}	γt_{45}
γt_{51}	γt_{52}	γt_{53}	γt_{54}	γt_{55}

Mixed – e.g. Blocking +AR1

- Gives block diagonal or toeplitz form

	$\varepsilon_{1,1}$	$\varepsilon_{1,2}$	$\varepsilon_{1,3}$	$\varepsilon_{2,1}$	$\varepsilon_{2,2}$	$\varepsilon_{2,3}$	$\varepsilon_{2,4}$
$\varepsilon_{1,1}$	1	ρ	ρ^2	0	0	0	0
$\varepsilon_{1,2}$	ρ	1	ρ	0	0	0	0
$\varepsilon_{1,3}$	ρ^2	ρ	1	0	0	0	0
$\varepsilon_{2,1}$	0	0	0	1	ρ	ρ^2	ρ^3
$\varepsilon_{2,2}$	0	0	0	ρ	1	ρ	ρ^2
$\varepsilon_{2,3}$	0	0	0	ρ^2	ρ	1	ρ
$\varepsilon_{2,4}$	0	0	0	ρ^3	ρ^2	ρ	1

Fortin et al

$$y_{ij} = A_i(1-e^{B_i(t_{ij}+D_i)})C_i + \varepsilon_{ij})$$

- Where i=plot, j=time period/measurement
- Correct covariance for
  - Repeated measures $\text{var}(\varepsilon_{ij})\text{var}(\varepsilon_{ij'})\rho^{|t-t'|}$
    - Nearby times correlated errors
  - Where $\text{var}(\varepsilon_{ij}) = \sigma^2|f(x_{ij},\beta)|^\theta$
    - Variance goes up monotonically with predicted y=f(x)
- $\text{cov}(\varepsilon_{ij}, \varepsilon_{ij'}) = \sigma^2 \quad |f(x_{ij},\beta)|^\theta \quad |f(x_{ij'},\beta)|^\theta \rho^{|t-t'|}$
- var heterosced repeated meas?

Fortin et al

- Covariance matrix - Block diagonal
  - 3 time periods/repeated measures/plot
  - 2 plots

	$\varepsilon_{1,1}$	$\varepsilon_{1,2}$	$\varepsilon_{1,3}$	$\varepsilon_{2,1}$	$\varepsilon_{2,2}$	$\varepsilon_{2,3}$
$\varepsilon_{1,1}$	$\sigma^2\|f(x_{1,1})\ f(x_{1,1})\|^\theta$	$\sigma^2\rho\|f(x_{1,1})\ f(x_{1,2})\|^\theta$	$\sigma^2\rho^2\|f(x_{1,1})\ f(x_{1,3})\|^\theta$	0	0	0
$\varepsilon_{1,2}$	$\sigma^2\rho\|f(x_{1,2})\ f(x_{1,1})\|^\theta$	$\sigma^2\|f(x_{1,2})\ f(x_{1,2})\|^\theta$	$\sigma^2\rho\|f(x_{1,2})\ f(x_{1,3})\|^\theta$	0	0	0
$\varepsilon_{1,3}$	$\sigma^2\rho^2\|f(x_{1,3})\ f(x_{1,1})\|^\theta$	$\sigma^2\rho\|f(x_{1,2})\ f(x_{1,3})\|^\theta$	$\sigma^2\|f(x_{1,3})\ f(x_{1,3})\|^\theta$	0	0	0
$\varepsilon_{2,1}$	0	0	0	$\sigma^2\|f(x_{2,1})\ f(x_{2,1})\|^\theta$	$\sigma^2\rho\|f(x_{2,1})\ f(x_{2,2})\|^\theta$	$\sigma^2\rho^2\|f(x_{2,1})\ f(x_{2,3})\|^\theta$
$\varepsilon_{2,2}$	0	0	0	$\sigma^2\rho\|f(x_{2,2})\ f(x_{2,1})\|^\theta$	$\sigma^2\|f(x_{2,2})\ f(x_{2,2})\|^\theta$	$\sigma^2\rho\|f(x_{2,2})\ f(x_{2,3})\|^\theta$
$\varepsilon_{2,3}$	0	0	0	$\sigma^2\rho^2\|f(x_{2,3})\ f(x_{2,1})\|^\theta$	$\sigma^2\rho\|f(x_{2,3})\ f(x_{2,2})\|^\theta$	$\sigma^2\|f(x_{2,3})\ f(x_{2,3})\|^\theta$

The solution: GLS

- GLS=OLS/GLM with a specification of covariance matrix Σ
- In practice, for n data points, Σ has nxn parameters.
  - Must specify a structure of Σ that has only a few parameters
- Two main versions of GLS
  - $\sigma^2\Sigma$ where Σ must be constant (only estimate σ^2)
    - Directly solvable
  - $\sigma^2\Sigma(\theta)$ where Σ is a function of parameters (estimate σ^2 and θ)
    - Requires an iterative solution
    - Estimate β with independence
    - Estimate $\Sigma|\beta$
    - Estimate $\beta|\Sigma$
    - Estimate $\Sigma|\beta$
    - ....

GEE

- GLS is great but don't throw away the power of GLIM (link functions, nonnormal errors)
- Generalized estimating equations
  - Combines GLS idea on errors + GLIM
  - Cannot be fit through MLE
  - Uses quasi-likelihood
  - To compare models use Wald test (not LR, F or χ^2)

Variance components analysis

- How much variance at each level
- E.g. within between genera
- Treat genus like a blocking factor with 1 measurement per species within
- Variance at species=σ^2 $\sigma^2/(\sigma^2+\tau^2)$ as %
- Variance at genus=τ^2 $\tau^2/(\sigma^2+\tau^2)$ as %

	$\varepsilon_{1,1}$	$\varepsilon_{1,2}$	$\varepsilon_{2,1}$	$\varepsilon_{2,2}$	$\varepsilon_{2,3}$
$\varepsilon_{1,1}$	$\tau^2+\sigma^2$	τ^2	0	0	0
$\varepsilon_{1,2}$	τ^2	$\tau^2+\sigma^2$	0	0	0
$\varepsilon_{2,1}$	0	0	$\tau^2+\sigma^2$	τ^2	τ^2
$\varepsilon_{2,2}$	0	0	τ^2	$\tau^2+\sigma^2$	τ^2
$\varepsilon_{2,3}$	0	0	τ^2	τ^2	$\tau^2+\sigma^2$

Variance partitioning

Table 1: Results of partitioning of variance analysis

Between	Within	Abundance % variance	Average abundance % variance	Maximum abundance % variance	Mass % variance	Trophic level % variance
Replicates /year	Routes	22.1	...	...	...	...
Routes	Species	37.1	...	...	...	...
Species	Genera	11.8	32.0	39.4	2.4	...
Genera	Family	6.5	14.5	12.2	9.1	...
Families	Order	8.0	2.1	1.7	6.1	12.4
Orders	Class	14.5	51.4	46.6	82.4	87.6
Total		100.0	100.0	100.0	100.0	100.0

Note: This table shows the percentage of variance in five variables (columns 3–7) explained by each level of taxonomic hierarchy that is relevant. Columns total to 100% except for rounding errors in the last decimal place. Variation between routes was analyzed only for abundance. Abundance was then reanalyzed with routes removed by an averaging and a maximum for abundance as described in "Hierarchical Partitioning of Variance" in "Methods." Trophic level was coded at the family level and does not vary within families for this reason.

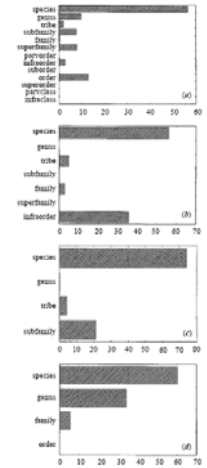
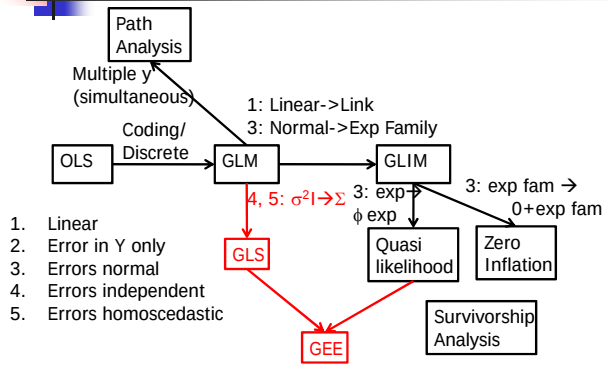


Figure 4. The proportion of variation in the geographic range sizes (logarithmically transformed) of species that is explained at different taxonomic levels for: (a) New World birds; (b) wildfowl; (c) woodpeckers; and (d) regular sea urchins. From: Gaston & Blackburn (1995b), Blackburn *et al.* (1998), and data in Easter (1995) and Kier & Lawton (1978).

- McGill 2008, Gaston 1998

Road map



29

In R

```
library(AED)
data(Squid)
names(Squid)
Squid$MONTH<-factor(Squid$MONTH)
plot(Testisweight~DML,data=Squid)
plot(Testisweight~MONTH,data=Squid)
#heteroscedastic against DML & Month
library(nlme) #contains gls command
m_nc<-gls(Testisweight~DML*MONTH,data=Squid)
plot(m_nc) #still heteroscedastic
m_lm<-lm(Testisweight~DML*MONTH,data=Squid)
m_lm
m_nc  #same as GLS with no variance/correlatin
```

In R II

```
vfix<-varFixed(~DML) #variance object prop to DML
vfix<-Initialize(vfix,Squid) #initialize with specific
   data
varWeights(vfix)
#varying s go in via weights option
m_fix<-
   gls(Testisweight~DML*MONTH,data=Squid,weights=vfix)
m_fix #big improvement in likelihood!
#what about MONTH
vmo<-varIdent(form=~1|MONTH)
vmo<-Initialize(vmo,Squid)
varWeights(vmo)
m_mo<-
   gls(Testisweight~DML*MONTH,data=Squid,weights=vmo)
anova(m_nc,m_fix,m_mo)
```

In R III

```
#var input directly
m_exp<-
   gls(Testisweight~DML*MONTH,data=Squid,weights=v
   arExp(form=~DML))
#combine
m_com<-
   gls(Testisweight~DML*MONTH,data=Squid,weights=v
   arExp(form=~DML|MONTH))
m_nc #estimates are different
anova(m_nc,m_fix,m_mo,m_exp,m_com) #which model
```

In R IV - Block

```
library(AED)
data(RiceFieldBirds)
d<-RiceFieldBirds
d$Rich=rowSums(d[,8:56]>0)
d$Field=factor(d$FIELD)
library(lattice)
xyplot(Rich~Time|Field,data=d,type=c("p",
  "smooth","grid"))
summary(m_ub<-lm(Rich~Time,data=d))
summary(m_bl<-lm(Rich~Time+Field,data=d))
```

In R V - GEE

```
library(AED)
data(RiceFieldBirds)
d<-RiceFieldBirds
d$Rich=rowSums(d[,8:56]>0)
d$Field=factor(d$FIELD)
library(lattice)
xyplot(Rich~Time|Field,data=d,type=c("p","smooth","grid
  "))
summary(m_ub<-lm(Rich~Time,data=d))
summary(m_bl<-lm(Rich~Time+Field,data=d))
library(geepack)
m_gee<-
  geeglm(Rich~Time,data=d,family=poisson,id=Field,corst
  r="ar1")
summary(m_gee)
```

In R – phylogenies (from Gene Hunt – see web link)

```
library(ape)
tree <- rtree(5)
tree$edge
tree$edge.length
tree$tip.label
plot(tree)
nodelabels(-1:-4) #internal nodes
size<-c(1,1,2,3,5) #continuous trait
larv<-c(0,0,1,1,1) #discrete trait
plot(tree,show.tip.label=F)
tiplabels(text=size) #put trait on tips
plot(tree, show.tip.label=F) #fancier
tiplabels(pch=19, cex=size)
plot(tree, show.tip.label=F)# two traits
tiplabels(pch=larv+21, bg=larv+1)
```

Phylogeny in R II

```
aa <- ace(size, tree) #calculate
  ancestral values using MLE
aa
plot(tree, show.tip.label=F)
tiplabels(pch=21, cex=size)
nodelabels(pch=21, cex=aa$ace)
lm(size~larv)
m=gls(size~larv,cor=corBrownian(1,tree))
```