Multiple regression

Multiple Regression

- $y=\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\ldots+\varepsilon$
- As our measuring gets better/cheaper, we have more variables
  - Earth sciences advance (climatology, remote sensing)
- Two key challenges
  - Too many variables to make sense of
  - Collinearity (correlation between variables)
    - Less noticed in ANOVA because usually factors are "orthogonal" – i.e. uncorrelated

"Too many" (aka lots of variables)

- Scaling
- Selection
  - Model selection
  - Stepwise selection
  - Thinking selection

Handle many variables 1 - scaling

- The β from traditional regression depends on units (measure in kg or g)
- This makes it impossible to compare coefficients
- To find the "important" independent variables
  - Look at |t| or p – reported in most packages
  - Standardize b
    - $b' = b s_x / s_y$ or convert x,y to z scores & do regression

Scaling in R

```
bd=read.table("/851/birds.csv",sep=
    ",",h=T)
names(bd)
bd2=bd[bd$Rangesize>0,]
summary(bd2)
summary(lm((log10(TotalAbund+.1))~(
    log10(Mass))+(log10(Rangesize))
    ,data=bd2))
summary(lm(scale(log10(TotalAbund+.
    1))~scale(log10(Mass))+scale(lo
    g10(Rangesize)),data=bd2))
```

Handling Many Variables II - Nested models

- Model A is nested in model B if all the terms in A are also in B
- Example
  - Y=a+bx is nested in y=a+bx+bx^2 and in y=a+bx+cz
- Einstein says "make a model as simple as possible but not simpler than needed"
  - You can compare the R^2 (adjusted?) of nested models
  - Can also do ANOVA tests (F-statistics) for statistical signifcance (degrees of freedom adjust for increased parameters)

Comparing nested models I - Adjusted R^2

- r^2
  - Goes up w/ increasing # of variables
    - Doesn't mean better
    - Have to get "enough" increase in r^2 to justify decrease in parsimony
  - Goes up w/ fewer rows
    - Aggregating data causes spurious increase in r^2
- $R^2=1-(n-1)/(n-p)(1-r^2)$

- Mallow's c better
- AIC covered in a few weeks is a much better way to do this

Comparing nested models II - Nested test

- In R: anova(m1,m2,m3,...)
- Reports progessive tests
- In general with nested models
  - Test most "complex" term in largest model
    - If not significant, throw it out and look at smaller model
  - EG x^2 or interaction term invalides interpretation of x or A+B
    - Don't use effect estimates, r^2 from model with higher order term in

Handling many variables III - Stepwise regression

- Have dozens of variables
- Let the computer pick the equation
- Forward and backward steps
  - Forward
    - start with intercept
    - do a regression with each variable added on
    - Pick one with greatest increase in adjusted r^2 or F
    - Repeat
  - Backward
    - Similar but start with all variables in
    - Calculate with each variable dropped, keep best
  - Stepwise = mix forwards and backwards

Stepwise Caveat

- Stepwise is OK for exploration
- Freedman's paradox
  - If you have enough variables, even random data will have some strong relations
  - Stepwise regression pulls these out
- Whittingham – bad! Murtaugh – fine!
  - On website
- Beware of significance tests on regressions developed stepwise
- If you want to do a significance test, have an a priori hypothesis!

Stepwise in R

```
em=read.table("/apache/serve/em
    lk.csv",sep=",",h=T)
names(em)
m=lm(log10(abund+.1)~.,data=em)
summary(m)
m.step=step(m)
```

Handling many variables IV - Really radical technique

- Think!
- Form hypotheses – use your knowledge to select variables a priori. Use biology.
- EG climate
  - What factors most important
    - Max Ann Temp, Growing season length correlated – which matters more? Use one!
  - Plant model – winter freezing, summer growth, summer drought
    - → Annual Min Temp, Growing Degree Days, Palmer Drought Severity Index

Multicollinearity

The correlation of x variables

Not an actual violation of OLS

But causes problems!

The problem: Bouncing β

- Exteme case – two identical columns
  - Sometimes shows up – mass, volume*density
- There is no correct answer for β_1 vs β_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
  - Body size, speed
  - Abundance of maple & pine covary negatively
  - Etc, etc…

Dealing with multicollinearity

- Measure & assess
  - Regression symptoms
  - Explore the correlation matrix
  - VIF
- Proceeding ahead
  - Principle component regression
  - Residual regression
  - Borcard Partition
  - Path analysis

Assesment I - Common symptoms in regression

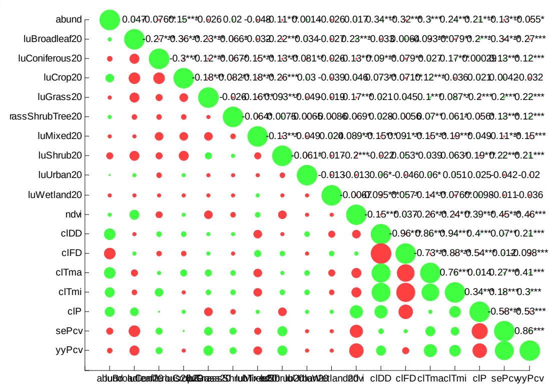
- From Belsley 1991
  - Coefficients have the wrong sign
  - Important predictors have high p-values
  - Deletion of one column or row causes the betas to change drastically

Variance-covariance matrix

- A square pxp (p=#variables) matrix containing covariances
  - $P=c*\exp(-(x-\mu)\Sigma^{-1}(x-\mu)/2)$
- Standardized matrix
  - each column converted to a z-score (subtract mean, divide by std dev)
  - Avoid units making some variables appear more important
- Standardized covariance matrix=correlation matrix

V_1	C_{12}	C_{13}
C_{12}	V_2	C_{23}
C_{13}	C_{23}	V_3

Assessment II - Correlation matrix example

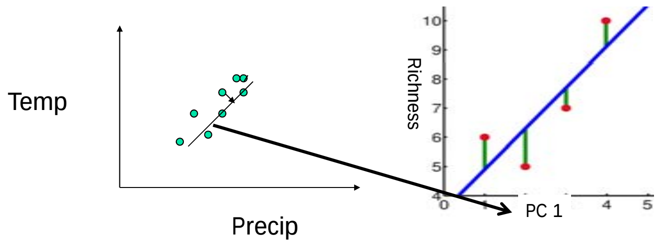


Assessment III - VIF

- Run regression with each variable as dependent
  - $VIF_i=1/(1-R^2)$
    - high if other variables explain this variable well
    - 0 if orthogonal
  - VIF>5-10 says highly collinear
  - Eliminate high VIF variables?
- library(car) has function vif(m)

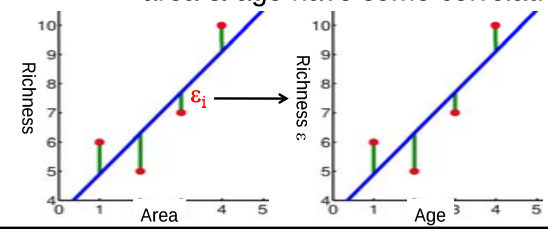
Proceeding I – Principle Components

- EG richness vs. climate
- But climate many variables – all correlated



Proceeding II – residuals regression

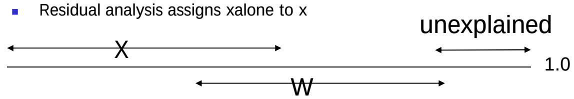
- Hypothesis: Island richness depends on age of isolation (relaxation effect)
- But island richness=f(area,age)
  - area & age have some correlation



Wilcox 1978 on website

Proceeding III – Borcard partition

- Partition r^2 among interesting/uninteresting variables
  - Y=f([X W])
  - X "interesting", W "uninteresting"
  - E.g. X=environment, W=space
- Calculate 3 variables
  - xr2=r^2 from y~x
  - wr2=r^2 from y~w
  - totr2=r^2 from y~x+w
- Then
  - Unexplained=1-tot
  - Explained=tot
  - x alone=totr2-wr2
  - w alone=totr2-xr2
  - xw combined=totr2-xalone-walone
- Residual analysis assigns xalone to x



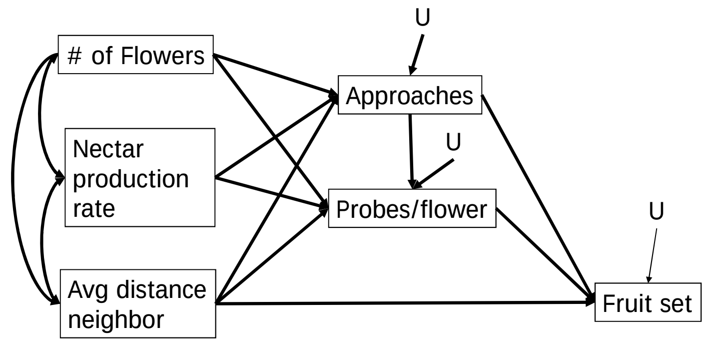
Proceeding IV - Path analysis

- Start with a priori hypothesis of causality
- Does two things
  - Somewhat stronger test of causality
  - Lessens problem of multicollinearity by a priori specification of variables, structure

Path analysis

- Two inputs
  - Correlation matrix
  - Path “diagram”

Mitchell’s Pollination Path

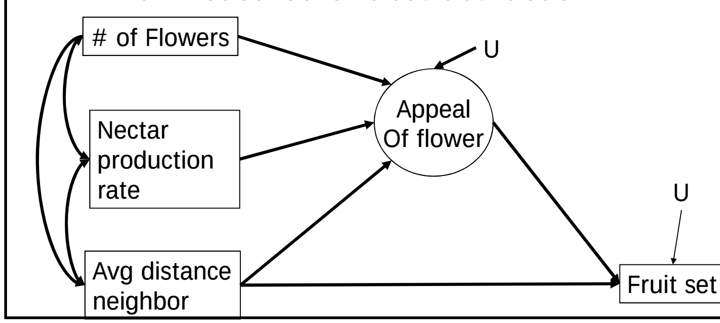


Many factors contribute

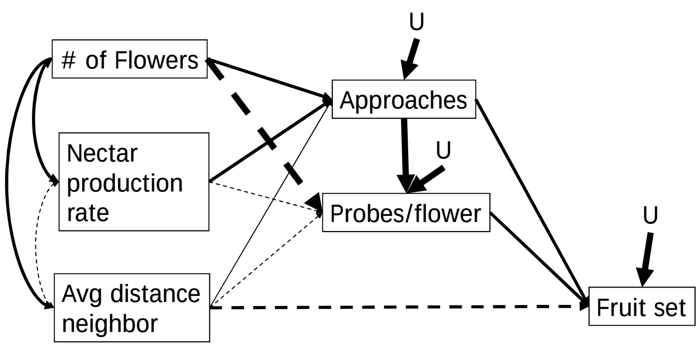
- Probes/flower driven by:
  - # approaches
  - Direct effect of # flowers
  - Indirect effect of # flowers on # approaches
  - Etc

Latent variables

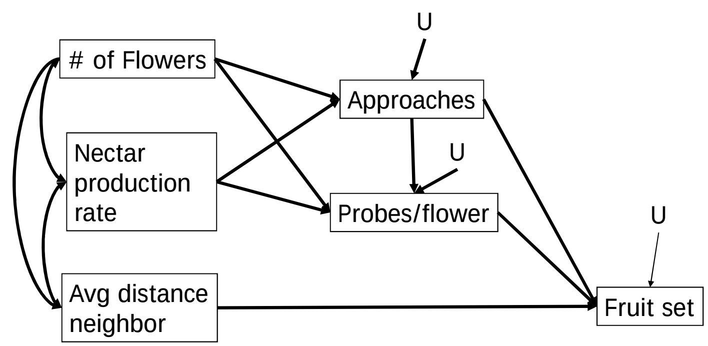
- Can introduce latent variables as unmeasured or abstract ideas



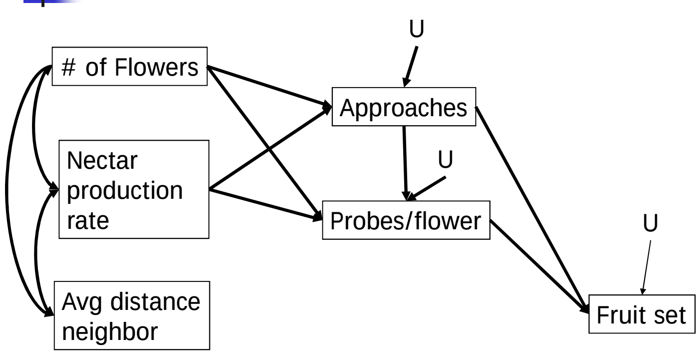
Results



Alternative 1 – Nested submodel



Alternative 2



Path analysis in R

```
library(sem)   #download from CRAN
cov<-matrix(c(
              3,   0,  0,
          -2.4,   2,   0,
              4, 1.8,  6), 3, 3, byrow=T)
cov<-(cov+t(cov))/2  #supposed to work w/ LT form but doesn’t
rownames(cov)<-colnames(cov)<-c(‘Label1’,’Label2’,’Label3’)
mod<-specify.model(file=“c:/…/mod.txt”)
```

or

```
Mod<-specify.model()
1:label1->label2,gamma12,NA
2:label1->label3,gamma13,NA
3:label2->label3,gamma23,NA
4:

sem<-sem(mod,cov,N=50)
summary(sem)
```