Week 11: Linear mixed models & autocorrelation

ANTH 674: Research Design & Analysis in Anthropology

Professor Andrew Du

Andrew.Du2@colostate.edu

Office Hours: Thursdays, 9:00am–12:00pm
In person: GSB 312
Virtual: https://tinyurl.com/F22ANTH674

1

Lecture outline

1. Linear mixed models
   1. What are they? What are they used for?
   2. Variance partitioning w/ crossed or nested designs
2. Measuring & dealing with autocorrelation
   1. Correlograms
   2. Association between two variables
      1. Cross-correlograms
      2. Generalized least squares

2

Linear mixed models

What are they? What are they used for?

3

What is a linear mixed model?

- A linear model with **fixed** AND **random** factors/effects
- I've always found the distinction between the two factors confusing, so let's go through them carefully


4

1

Fixed factors

- Measured levels are the only ones of interest
- E.g., want to know how intercept of growth rate ~ temperature changes w/ **different levels of nitrogen** (i.e., ANCOVA)
- Thus far, we have only looked at fixed factors
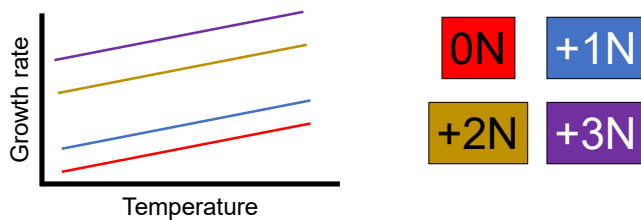

5

Random factors

- Measured levels are a random, representative sample of some population of levels
- E.g., want to infer something about population of intercepts (i.e., SD of intercepts across levels)
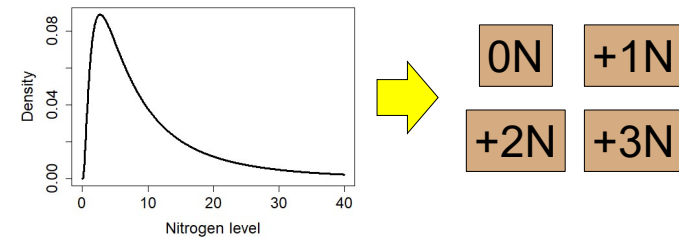
6

Ultimately, depends on research question!

- **Fixed factor**: estimate how intercept and/or slope changes ~ different N levels (i.e., ANCOVA)
- **Random factor**: estimate SD of population of intercepts and/or slopes across N levels
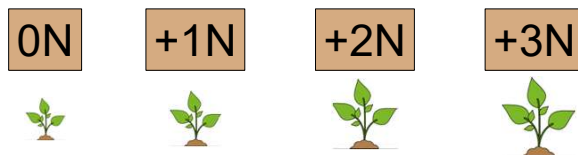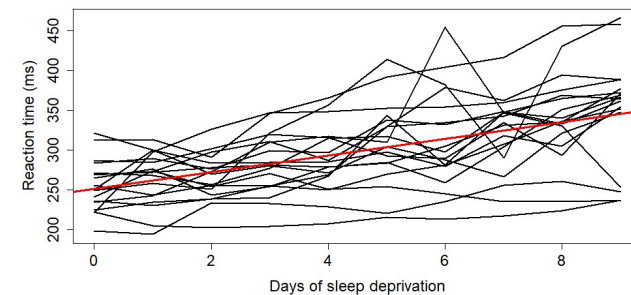

7

For example…

- Study measured reaction times of 18 subjects each day after full night's sleep & then 9 days of sleep deprivation (180 total observations)

8

Fixed factor approach

- Estimates intercept & slope of `reaction ~ days` for each subject (factor level)
- <u>ANCOVA</u>: `lm(reaction ~ days * subject)`

36 estimated coefficients for 180 data points!

(plot: Reaction time (ms) vs Days of sleep deprivation)
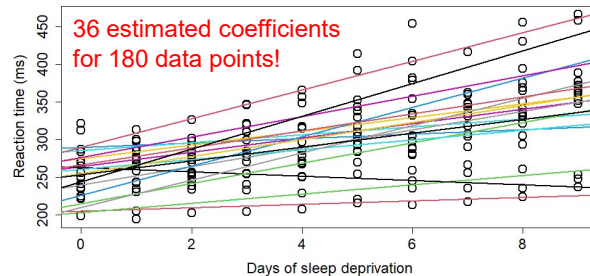
9

Random factor approach

- Estimates SD of intercepts & slopes across levels **ALONG WITH** intercept & slope fixed effects (via maximum likelihood)

```
Random effects:
             StdDev   Corr
(Intercept) 24.740241 (Intr)
Days         5.922103 0.066
Residual    25.591843
```

Coefficient estimates same as `lm(reaction ~ days)`!

```
Fixed effects: Reaction ~ Days
             Value  Std.Error  DF  t-value p-value
(Intercept) 251.40510 6.824516 161 36.83853     0
Days         10.46729 1.545783 161  6.77151     0
```

Can only calculate P-values in simple mixed models (e.g., balanced design, no crossed random effects)

10

Random factor approach
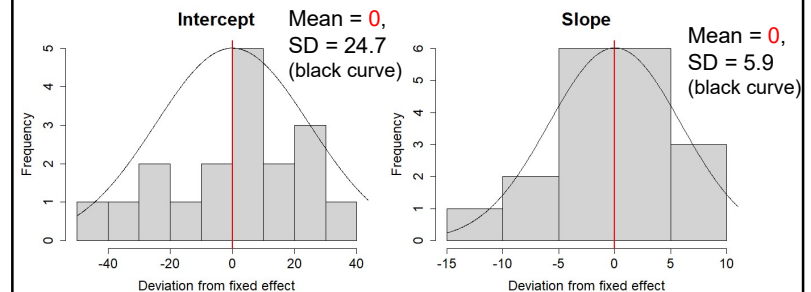
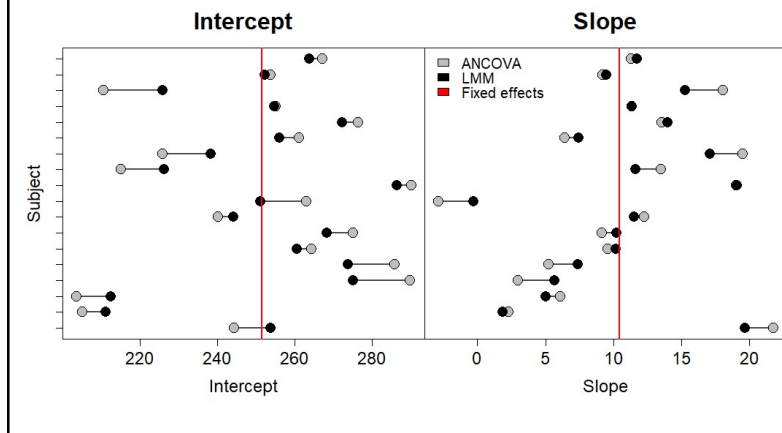- Can get intercept & slope for each subject (**NOT** estimated coefficients!)

Intercept   **Slope**

Drawn from pop. w/ mean = 0, SD = 5.9

Drawn from pop. w/ mean = 0, SD = 24.7

(Deviation from fixed effect)

11

Random factor approach

- Can get intercept & slope for each subject (**NOT** estimated coefficients!)

Intercept   Mean = 0, SD = 24.7 (black curve)

Slope   Mean = 0, SD = 5.9 (black curve)

(histograms: Frequency vs Deviation from fixed effect)

12

3

Compared to ANCOVA

Intercept   **Slope**

13

Complete pooling
One intercept & slope for all data

No pooling
ANCOVA w/ diff. intercepts & slopes

Partial pooling
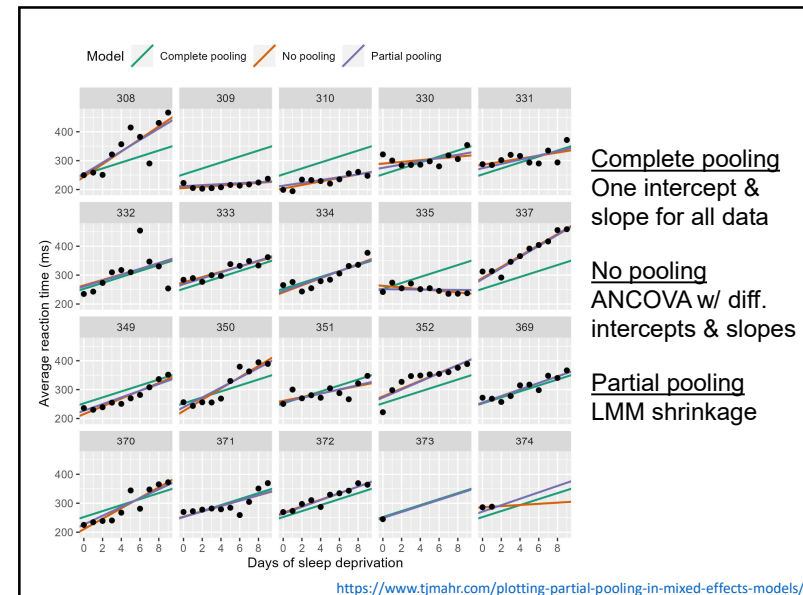LMM shrinkage

https://www.tjmahr.com/plotting-partial-pooling-in-mixed-effects-models/
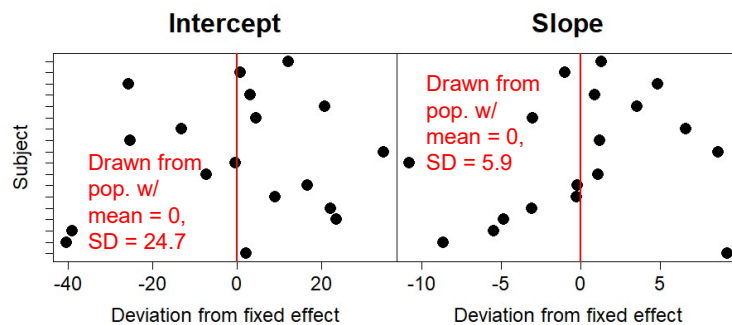
14

Shrinkage or regularization

- In ANCOVA, each line is fit to data for one factor level **only**
- In LMM, levels are drawn from a population, so they are expected to share characteristics (w/ a distribution mean = fixed effect)
  - This is how you can predict for new factor levels w/ LMM (can't do this w/ ANCOVA)
- If a level has few observations, it borrows more information from other levels and is pulled more towards the average (fixed effect)

15

Fixed vs. random factors

- **Fixed factors**
  - Interested in how intercept and slope changes *independently for each subject*
  - Potentially need to estimate *A LOT* of coefficients
- **Random factors**
  - Interested in the SD of intercepts and slopes *across all **possible** subjects*
  - Estimates fixed effects AND SDs
  - Shrinks intercepts & slopes towards means

https://dynamicecology.wordpress.com/2015/11/04/is-it-a-fixed-or-random-effect/
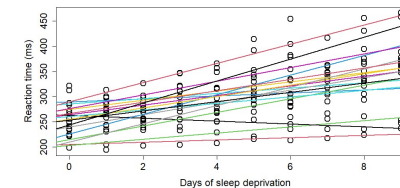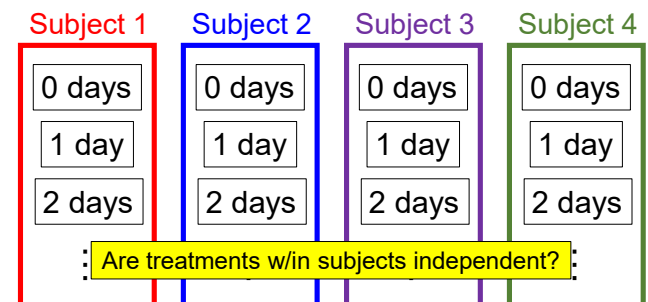
16

Questions?

17

Reasons for doing LMM

1. Factor levels are random sample from larger population → interested in estimating SD of population of intercepts/slopes
2. Ratio of data points to # coefficients in ANCOVA too small → overfitting (only need to estimate SDs in LMM)

36 estimated coefficients for 180 data points!

18

Reasons for doing LMM

1. Factor levels are random sample from larger population → interested in estimating SD of population of intercepts/slopes
2. Ratio of data points to # coefficients in ANCOVA too small → overfitting (only need to estimate SDs in LMM)
3. Shrink intercepts and slopes towards mean
4. Want to account for non-independence in data using a blocking factor (very popular reason)

19

E.g., sleep study

- Each subject is treated as a block, each w/ treatments of different days of sleep deprivation (AKA **repeated measures design**)

Subject 1	Subject 2	Subject 3	Subject 4
0 days	0 days	0 days	0 days
1 day	1 day	1 day	1 day
2 days	2 days	2 days	2 days

Are treatments w/in subjects independent?

20

Analyzing block design

- Can do ANCOVA, but problematic if lots of levels & not enough data
- Also, many times blocks are not of interest (i.e., don't care how intercept/slope change w/ block)
- **Can use LMM & include block as a random factor to account for it**
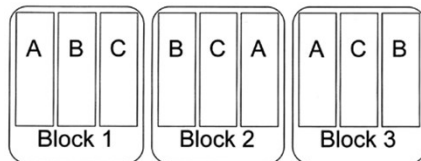
21

Random factor rules

- Random factor **MUST** be categorical (i.e., it must be a factor w/ levels)
  - A continuous IV can be entered into model as covariate (only one extra estimated coefficient)
- Need at least five levels (would you calculate SD of a sample w/ < 5 data points?)
  - Otherwise, just make factor a fixed effect

22

Generalized linear mixed models (GLMMs)

- Mixed model w/ non-normally distributed errors (cf. generalized linear models)
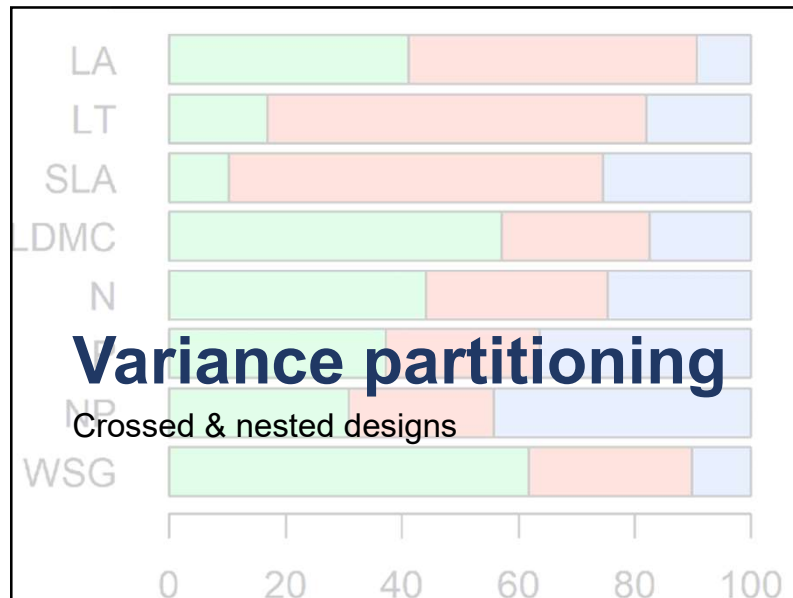- E.g., presence/absence of lions ~ wildebeest population density w/ plot as blocking factor

23

Questions?

24

Variance partitioning

Variance partitioning

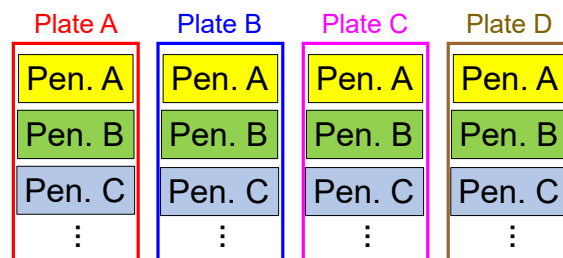Crossed & nested designs

25

Variance partitioning

- Can partition DV variance **not accounted for by fixed effects** among multiple random factors
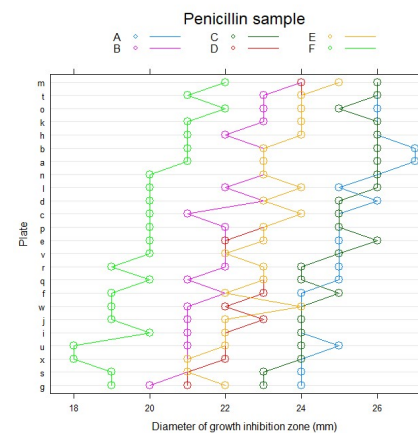- Or, can fit an intercept-only model to partition DV variance among random factors

26

Multiple random factors

1. **Crossed random factors**: each level from a factor is w/ each level from the other
- E.g., how penicillin inhibiting growth of *Bacillus* varies across plates & penicillin samples

Plate A	Plate B	Plate C	Plate D
Pen. A	Pen. A	Pen. A	Pen. A
Pen. B	Pen. B	Pen. B	Pen. B
Pen. C	Pen. C	Pen. C	Pen. C
⋮	⋮	⋮	⋮

27

Crossed random factors

Penicillin sample

- Fully crossed! Every combination of `sample` and `plate` is represented
- Unreplicated and completely balanced (only one observation per `sample`/`plate` combo)
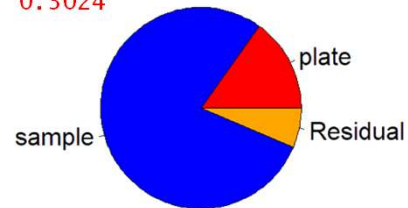
Diameter of growth inhibition zone (mm)

28

7

Crossed random factors

- Intercept-only LMM w/ `plate` & `sample` as crossed random effects

- ```
  Random effects:
  Groups    Name          Variance
  plate     (Intercept)   0.7169
  sample    (Intercept)   3.7311
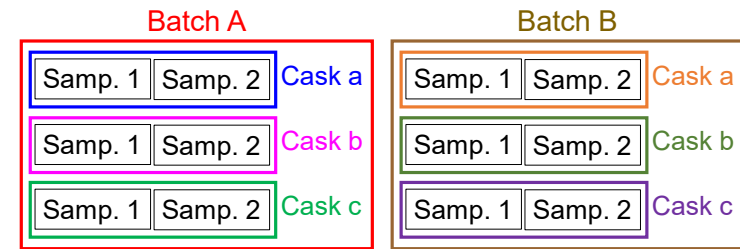 Residual   0.3024
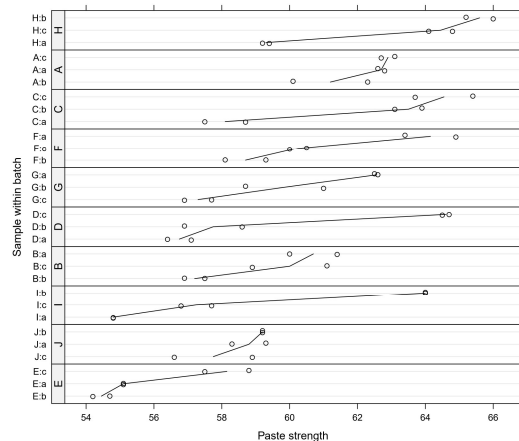 ```

Multiple random factors

2. **Nested random factors**: factors are nested within each other

- E.g., strength of chemical paste, where two samples were taken from each cask from each batch

29

30

Nested random factors

- 10 batches (capital letters)
- 3 casks per batch (lower case)
- Two samples per cask

Nested random factors

- Intercept-only LMM w/ `cask` nested within `batch`

- ```
  Random effects:
  Groups    Name          Variance
  cask      (Intercept)   8.434
  batch     (Intercept)   1.657
  Residual                0.678
  ```
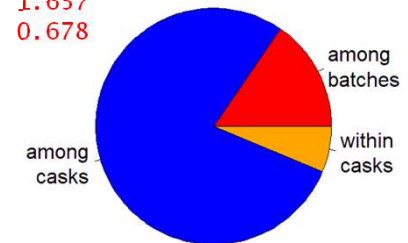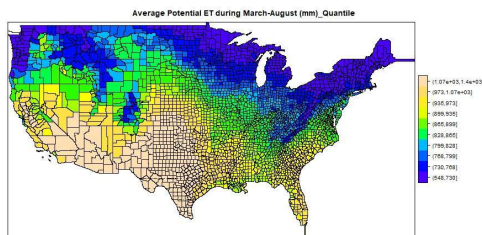

31

32

8

Questions?

Average Potential ET during March-August (mm)_Quantile

Measuring & dealing w/ autocorrelation

Non-independent data

- Data across space, time, or a phylogeny will almost always be non-independent
- Data closer in space/time/relatedness will be more similar

---

Three components

- Patterns in space & time (& phylogeny) can be decomposed into three components
- $z(d) = \mu(d) + \eta(d) + \varepsilon$, where (d) means component is a function of distance
  - i.e., pattern = trend + autocorrelated (red) noise + independent (white) noise
  - Temporal data can also include a cyclical component too (e.g., monthly, seasonality, orbital cycles)
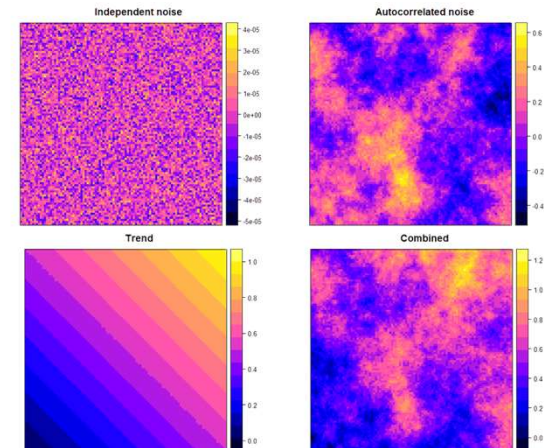
Simulated example for <u>time</u>
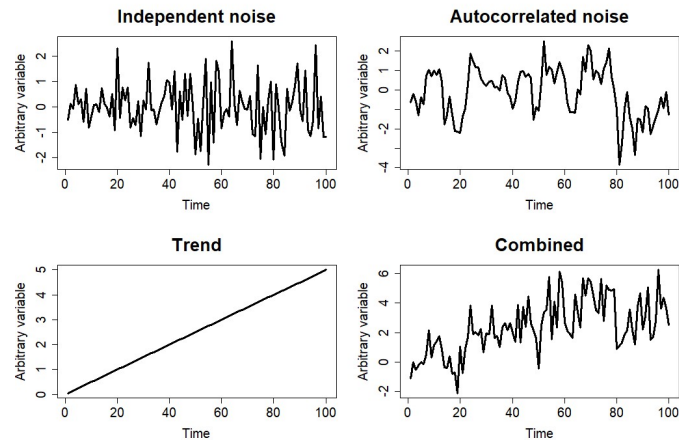
37

Simulated example for <u>space</u>


38

Autocorrelated noise: nuisance or interesting?

- **<u>Nuisance:</u>** non-independence violates one of the assumptions of linear models (increases Type I error rate)
  - Must correct for this (e.g., w/ blocking)!
- **<u>Interesting:</u>** How does strength of autocorrelation vary as a function of distance in space/time/phylogeny?
  - May tell us something about the underlying process generating autocorrelation!
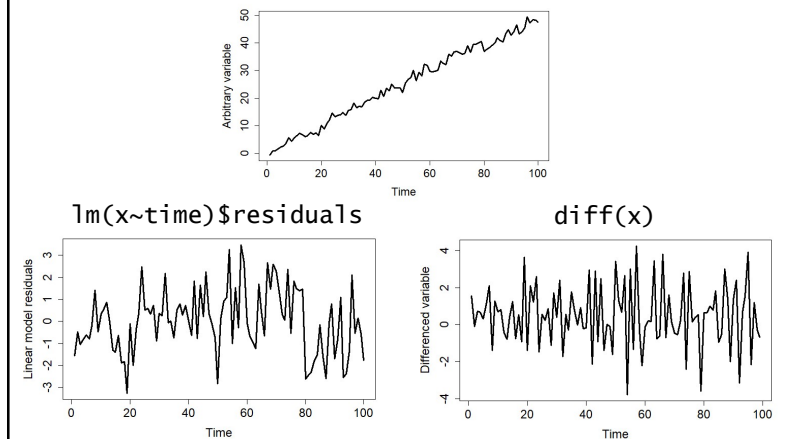
39

Questions?

40

10

Measuring autocorrelated noise

Correlograms
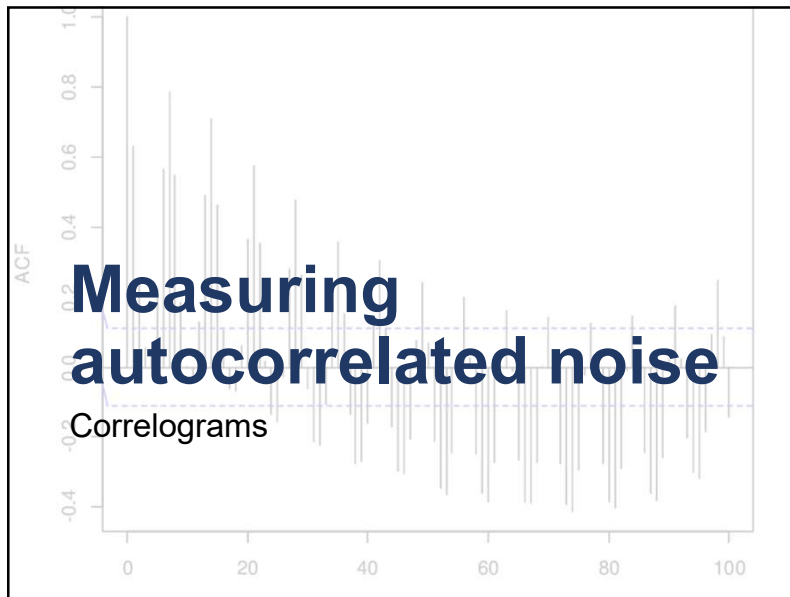
Detrending

- If temporal or spatial data are strongly trended, need to first remove trend (i.e., **detrending**) to focus on autocorrelated noise
1. Can fit a trend to data and take the residuals
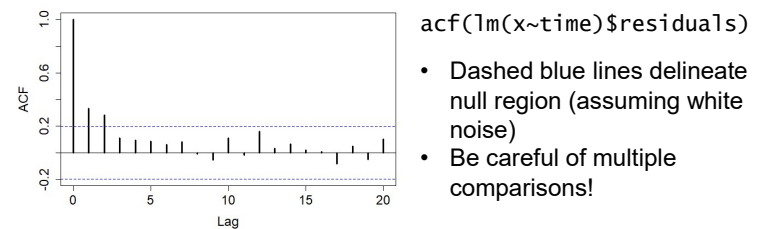2. Calculate first differences: $x_i' = x_{i+1} - x_i$

Example w/ temporal data

`lm(x~time)$residuals`   `diff(x)`


Autocorrelation function

- Quantifies how time series is correlated with itself at different lags
- E.g., `c(1, 2, 3, 4, 5)`
  - <u>time lag = 1</u>, `c(1, 2, 3, 4)` vs. `c(2, 3, 4, 5)`
  - <u>time lag = 2</u>, `c(1, 2, 3)` vs. `c(3, 4, 5)`

`acf(lm(x~time)$residuals)`

- Dashed blue lines delineate null region (assuming white noise)
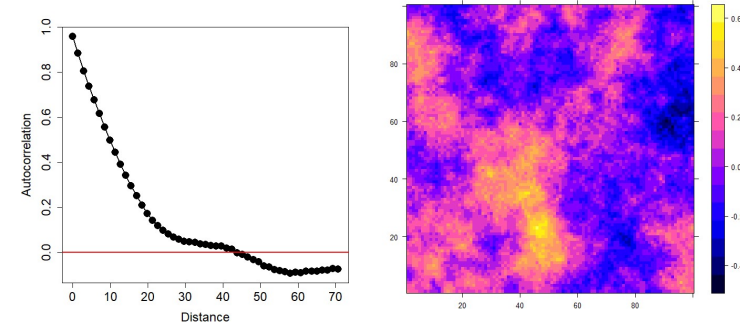- Be careful of multiple comparisons!

A real example

- Quarterly approval rating of US president from 1945 to 1974
- Not strongly trended, so no need to detrend
- Can begin to hypothesize what's generating AC!
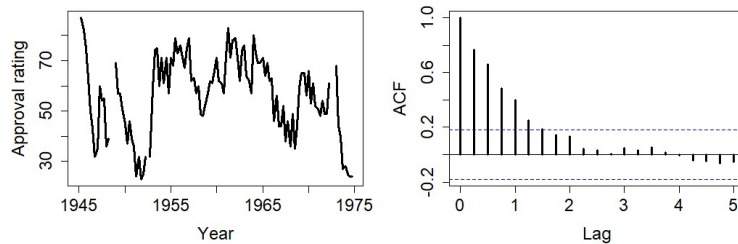
45

Spatial correlogram

- Quantifies correlation between pairs of points as you increase the distance (lag) between them
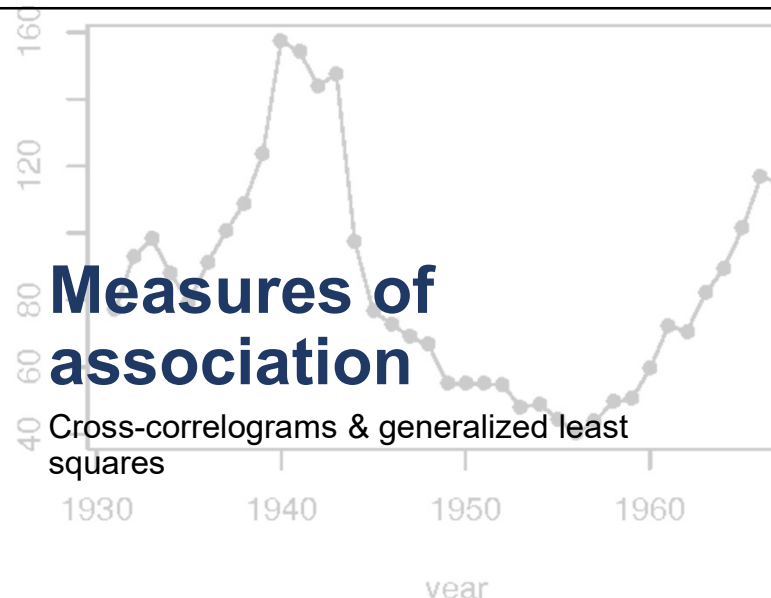

46

Questions?

47

Measures of association

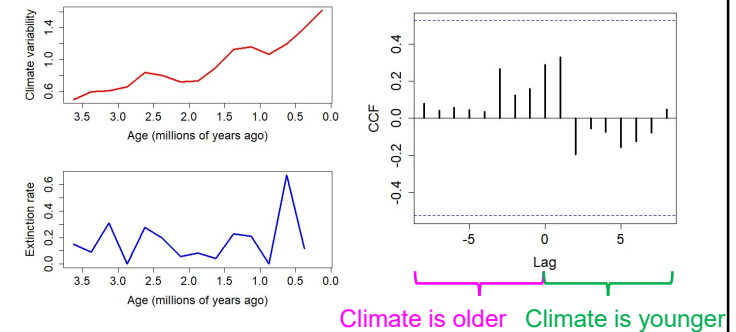Cross-correlograms & generalized least squares

48

12

Comparing two variables

1. **Cross-correlograms**: looks at how two variables are correlated at different lags
   - Perhaps one variable responds to another only after a certain amount of time/distance
   - As w/ correlation, symmetry between variables
2. **Generalized least squares**: models a DV as a function of one or more IVs, while explicitly modeling errors as non-independent
   - Good if variables are obviously DV or IV, or if you want to do prediction (as w/ linear models)
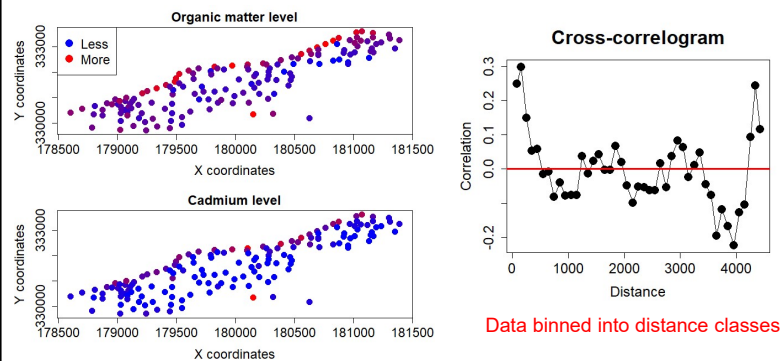
49

Cross-correlation function

- E.g., climate variability vs. mammal extinction rates in Plio-Pleistocene eastern Africa

Climate is older   Climate is younger

50

Spatial cross-correlogram

- E.g., soil organic matter & cadmium level in floodplain of Meuse River, Netherlands


Data binned into distance classes

51

Generalized least squares

- Relaxes GLM assumption of independent errors
- GLS estimates extra parameters to model non-independent errors
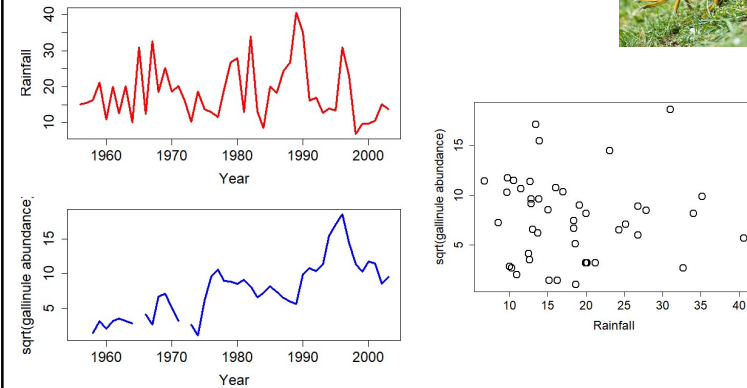- Parameters estimated using maximum likelihood

52

13

Temporal data

- AR1 autocorrelation error structure is popular
- Correlation between errors one time step apart estimated as ρ (usually positive), two steps apart is ρ^2, three steps is ρ^3, etc.
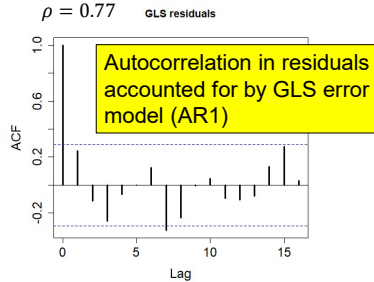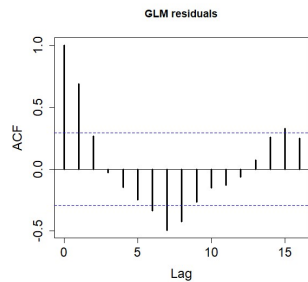

$\rho = 0.8$
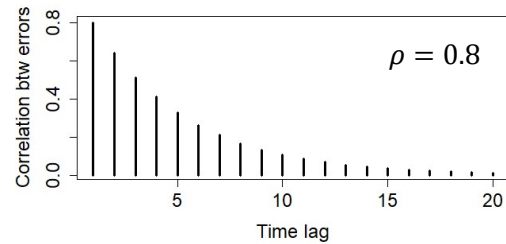
53

E.g., Hawaiian gallinule

54

sqrt(gallinule)~Rainfall+Year
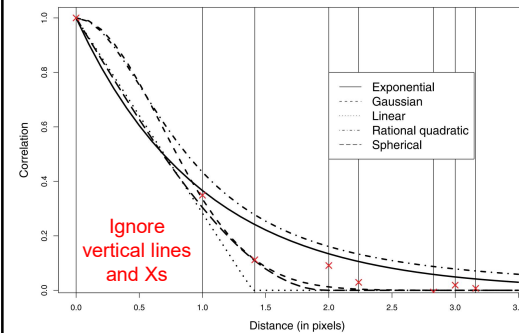
Coefficients are very similar!

lm coefficients:

	Estimate	Pr(>|t|)
(Intercept)	-4.777e+02	1.26e-10
Rainfall	8.604e-04	0.986
Year	2.450e-01	8.15e-11

GLS coefficients:

	Value	p-value
(Intercept)	-436.4326	0.0030
Rainfall	-0.0098	0.7649
Year	0.2241	0.0026

$\rho = 0.77$

Autocorrelation in residuals accounted for by GLS error model (AR1)

55

Spatial data

- Non-independence between errors modeled as a function of distance between them

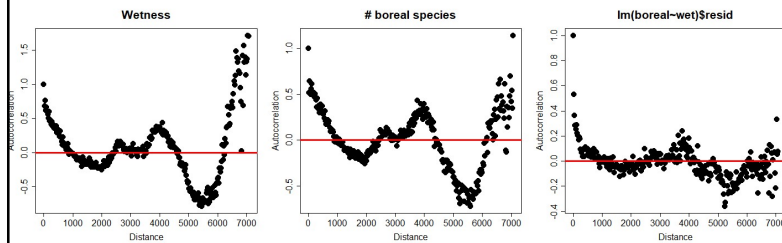Ignore vertical lines and Xs

- Five common models
- Different parameterizations will slightly change shape of each model

56

14

One last note

- If an IV captures autocorrelation in DV, then there will be no autocorrelation remaining in residuals (GLM assumptions not violated!)

Questions?

Summary

- Use LMMs (random effects) if:
  1. Factor levels are random, representative samples from some larger population and are interested in estimating SD of population of intercepts/slopes
  2. Have too few data (prevent overfitting w/ ANCOVA)
  3. Want to account for non-independence w/ blocking (and have too few data)
  4. Interested in shrinkage of estimates
  5. You want to partition variance in DV
- Correlograms quantify autocorrelation structure, which is interesting in itself!
- Cross-correlograms & GLS quantify association btw variables in time/space/phylogeny (but remember, OLS estimates are unbiased!)

Statistics vignette

Let's play a game…

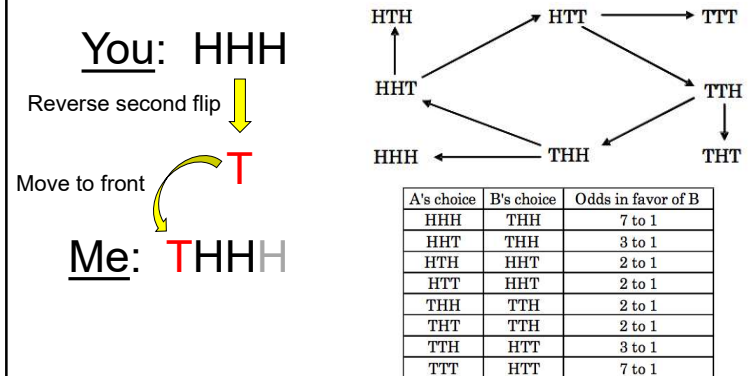<u>Rules</u>

1. Pick an outcome of three coin flips, e.g., HHH
2. I pick my own outcome
3. If your outcome appears first in *n* flips, you win

61

What's the secret?

• This is a game called "Penney's game"

<u>You</u>: HHH

Reverse second flip

Move to front   T

<u>Me</u>: THHH

A's choice	B's choice	Odds in favor of B
HHH	THH	7 to 1
HHT	THH	3 to 1
HTH	HHT	2 to 1
HTT	HHT	2 to 1
THH	TTH	2 to 1
THT	TTH	2 to 1
TTH	HTT	3 to 1
TTT	HTT	7 to 1

62

Why does it work? (an example)

<u>You</u>: HHH
<u>Me</u>: THH

You win only if HHH comes up in first three flips:
HHH…

Otherwise, T *must* precede any HHH, and I win:
…<u>T</u>HHH…

63