Week 9: Modern regression models

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. What are modern regression models?
2. The models we'll go over:
   - Model II regression
   - Segmented regression
   - Non-linear least squares
   - Locally estimated scatterplot smoothing

2

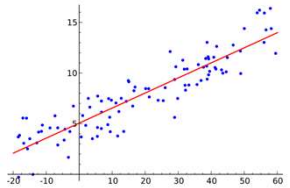
What are modern regression models?

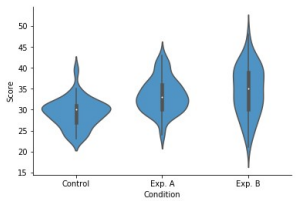
*Just a "wastebasket" phrase to describe newer regression methods

3

Quick review: general linear models (GLM)

- Models continuous DV as a *linear/additive* function of one or more IVs
- $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 \ldots + \beta_n X_n + \varepsilon$
- IVs can be continuous or categorical





4

Review: GLM assumptions

1. Relationship between DV and IVs is linear
2. IVs measured without error
3. Error terms have mean = 0 and are normally distributed
4. Error terms drawn from population with the same variance (homoskedasticity)
5. Error terms are independent

5

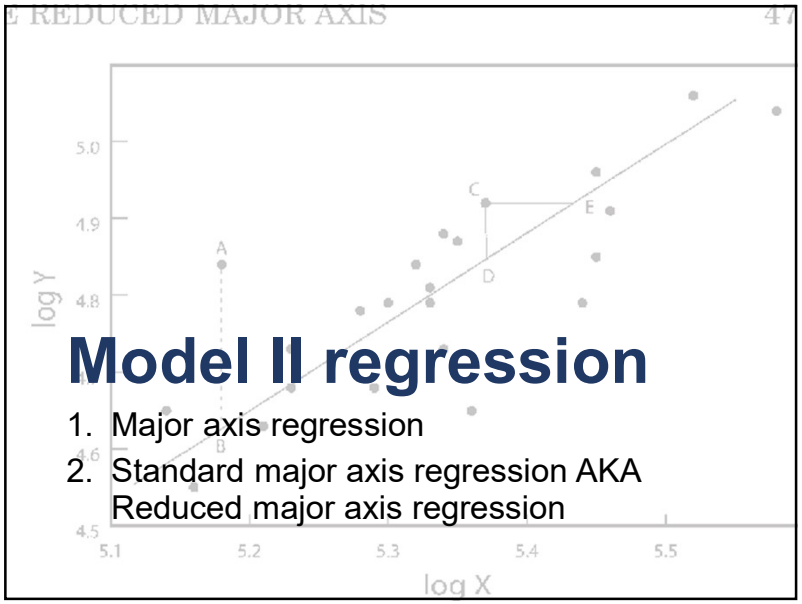
Relaxing GLM assumptions

1. Relationship between DV and IVs is linear
   - Segmented regression
   - Non-linear least squares
   - Locally estimated scatterplot smoothing
2. IVs measured without error
   - Model II regression

6

Questions?

7



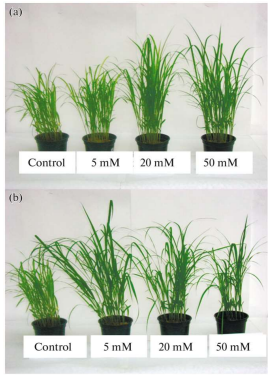
Model II regression

1. Major axis regression
2. Standard major axis regression AKA Reduced major axis regression

8

GLM: IV measured w/o error

- Makes most sense in experiments where IV is controlled/manipulated



El Omari et al., 2016

9

GLM: IV measured w/o error

- Makes most sense in experiments where IV is controlled/manipulated
- What if this is not the case & IV varies due to unmeasured factors (i.e., "has error")?
  - E.g., Measuring human weight ~ height
- Or unclear which variable is IV or DV
  - E.g., Petal length ~ width
- Need Model II regression methods

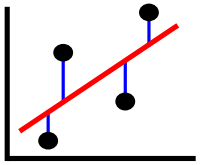
10

https://cran.r-project.org/web/packages/lmodel2/vignettes/mod2user.pdf

Model II regression

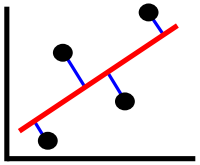
- Minimizes errors in BOTH variables to estimate regression intercept & slope
- Can hypothesis test (i.e., 95% CIs & P-values on coefficients), but cannot predict

Model I regression (least squares)

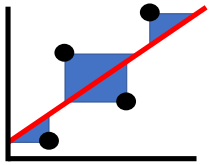


Model II regression

Major axis regression (MA)



Standard major axis regression (SMA)

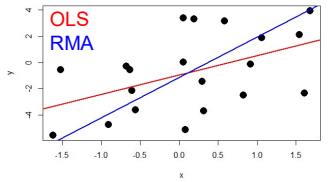


11

Standard major axis regression

- MA assumes X & Y in same units
- SMA is MA done on standardized X & Y
- SMA called reduced major axis regression (RMA) in bio. anth. (and other fields)
- $\text{Slope} = \pm \text{SD}_y / \text{SD}_x$ (sign determined by corr., r)
- $\text{Intercept} = \bar{y} - \bar{x} \times \text{slope}$ (as in OLS)

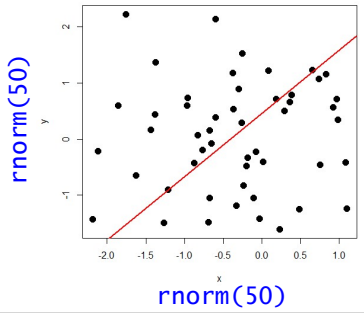
- RMA slope also = OLS slope / $|r|$
- So, RMA slope magnitude ≥ OLS slope magnitude



12

Be careful!

- Even if no relationship between two variables, RMA will plot a positive or negative fitted line (b/c slope $= \pm \mathrm{SD}_y/\mathrm{SD}_x$)



"If r is not significantly different from zero, [RMA slope] should not be computed."

-Legendre & Legendre 2012:550

13

Final Model II thoughts

- I find RMA slope difficult to interpret: what does $\mathrm{SD}_y/\mathrm{SD}_x$ mean?
- Perhaps Model II best for comparing coefficient estimate(s) to value(s) expected from theory
- If error in DV is much greater than that in IV, okay to use OLS (Model I)
- I never use Model II regressions

14

Questions?

15

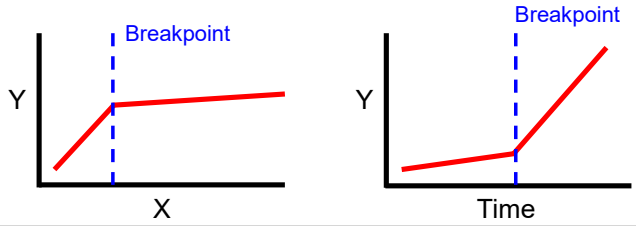
Segmented regression

AKA piecewise regression
AKA breakpoint regression
AKA change point regression

16

When it's used

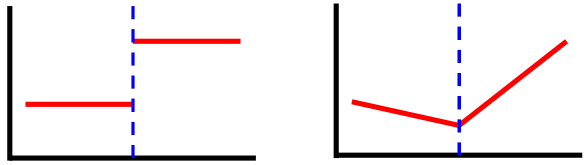
- When relationship between DV & IV can be described w/ ≥2 linear segments
  - E.g., threshold effect as IV increases
  - E.g., change in relationship w/ time
- Regression segments separated by breakpoints



17

How it works

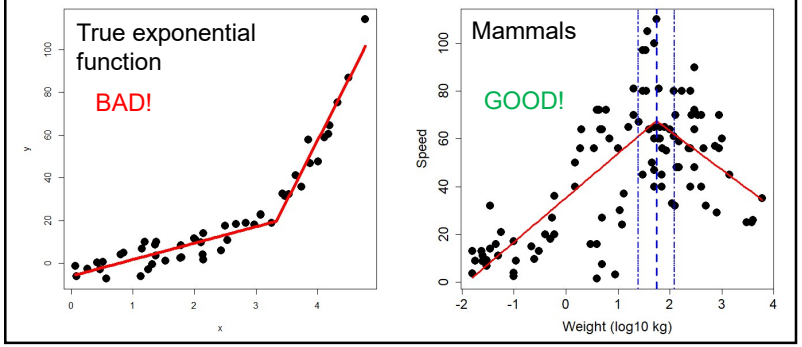
- Need to specify number of breakpoints (*n*)
  - Changing intercepts only → 2*n* + 1 coefficients
  - Changing intercepts & slopes → 2*n* + 2 coeff.
- Coefficients estimated algorithmically using maximum likelihood estimation
- Can hypothesis test all estimated coefficients, including breakpoints (95% CI & P-values)



18

Recommendations

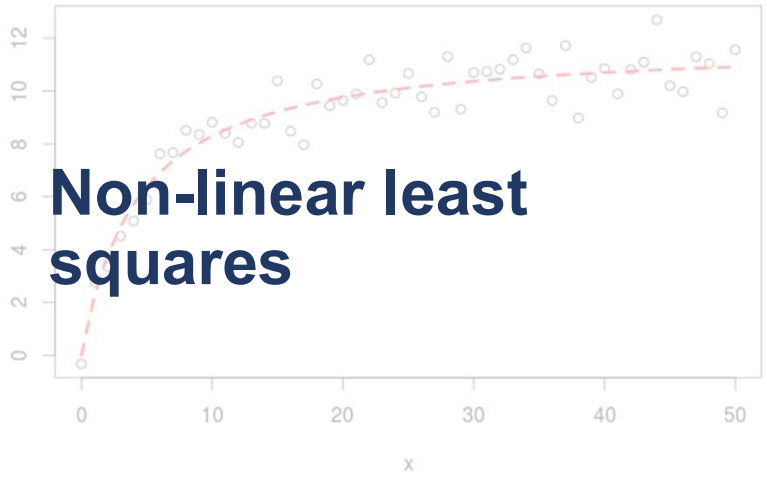
- Don't use to approximate non-linear functions
- Best to use when breakpoint is of main interest



19

Questions?

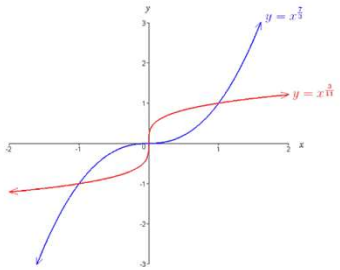
20



21

GLM: coefficients linearly enter model

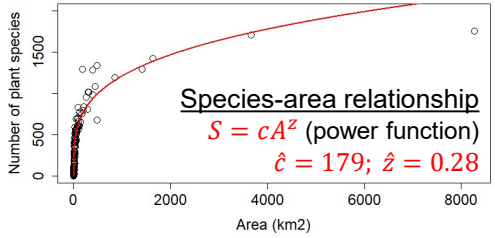
- $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 \ldots + \beta_n X_n + \varepsilon$
- May want to fit function where this isn't the case
  - E.g., power function: $Y = \beta_0 X^{\beta_1}$



22

Non-linear least squares

- Coefficients estimated algorithmically → computer tries different values to minimize residuals (i.e., least squares criterion)
- Need prespecified function (e.g., based on theory)
- Can do hypothesis testing & prediction



23

Questions?

24

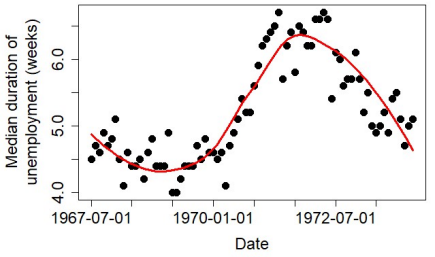
Locally estimated scatterplot smoothing

LOESS regression

25

When it's used

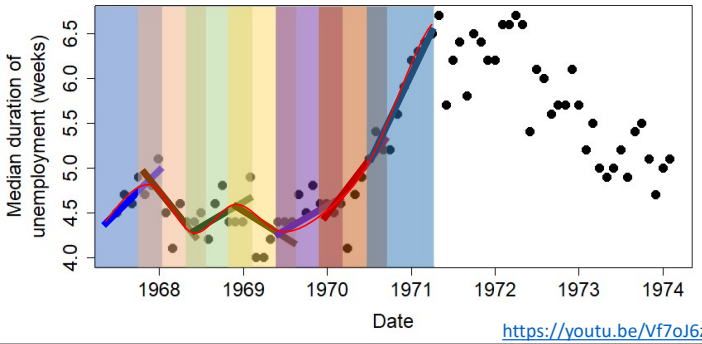
- No *a priori* function to explore or test (e.g., linear, segmented, exponential)
- Simply want to draw audience's eye to average trend, using a smoothed curve



26

How it works in principle

- Data are smoothed by iteratively moving a window & fitting a linear model each time

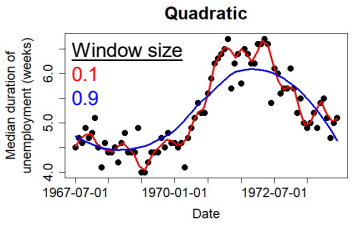


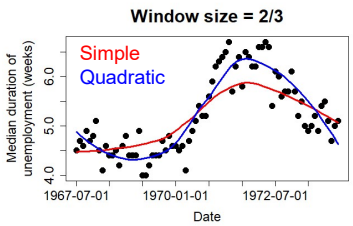
https://youtu.be/Vf7oJ6z2LCc

27

Adjusting the smoothing

- Window size set as proportion of total data pts.
  - Smoother as window size increases
- Window linear model can be simple linear regression or quadratic regression
  - Smoother if simple linear regression





28

Questions?