ECO 602 Analysis of Environmental Data

FALL 2019 – UNIVERSITY OF MASSACHUSETTS

DR. MICHAEL NELSON

ECoGSS
Department of Environmental Conservation
Graduate Student Symposium
28 February 2020

Call for Abstracts!

What is ECoGSS?
Environmental Conservation (ECo) Grad Student Symposium, featuring grad student posters, talks, and panels. ECoGSS 2019 is Feb. 28, 2020.

Who should submit an abstract?
Any ECo-affiliated grad students in ECo, MS3, BCT, IMS, OEB, PB, and others.

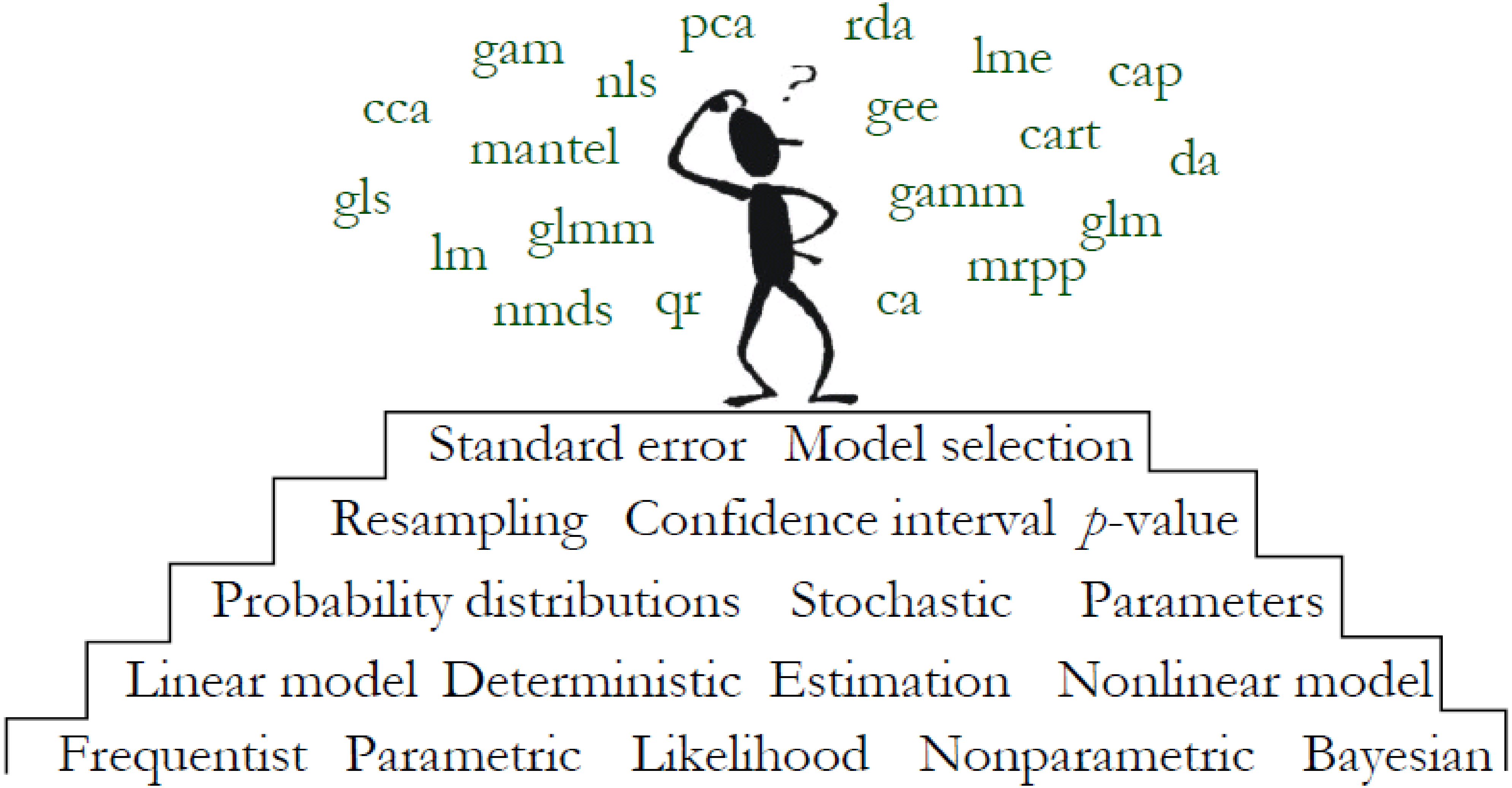
How do I submit an abstract?
Follow instructions here: https://blogs.umass.edu/ecogss/

When is my abstract due?
Abstracts are due Nov. 22, 2019!

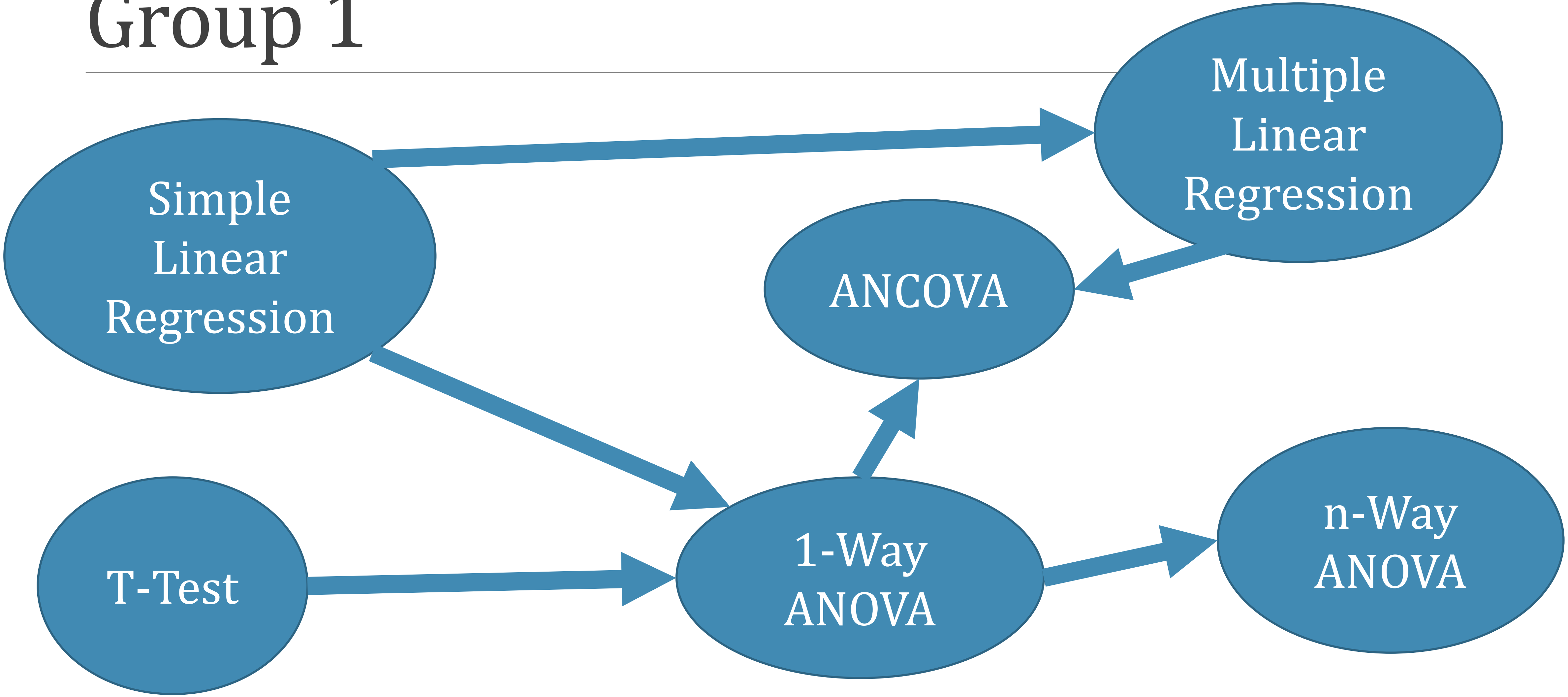
Today’s Agenda

- Key ideas from Group I analyses
- Recap/continuation of model output interpretation
- In-class data activity/quiz
- Beyond general linear models
- Final project/critical paper review in-class activity

Foundation for Understanding Statistical Methods…



Group 1



Simple Linear Regression
Multiple Linear Regression
ANCOVA
T-Test
1-Way ANOVA
n-Way ANOVA

Recall some of the key group 1 assumptions.

Independent observations

Linear relationships

Normal errors

Constant variance

Interpreting model output

Two important tools:

1. ANOVA table
2. Model coefficient table

SLR: Iris petal width predicted by petal length

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  -0.363076   0.039762  -9.131  4.7e-16 ***
Petal.Length  0.415755   0.009582  43.387  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2065 on 148 degrees of freedom
Multiple R-squared:  0.9271, Adjusted R-squared:  0.9266 
F-statistic:  1882 on 1 and 148 DF,  p-value: < 2.2e-16
```

Multiple regression: petal width predicted by petal length, sepal length

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.008996   0.182097  -0.049   0.9607
Petal.Length  0.449376   0.019365  23.205   <2e-16 ***
Sepal.Length -0.082218   0.041283  -1.992   0.0483 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2044 on 147 degrees of freedom
Multiple R-squared:  0.929, Adjusted R-squared:  0.9281
F-statistic: 962.1 on 2 and 147 DF,  p-value: < 2.2e-16
```

ANOVA: Species and petal width

```
Coefficients:
                     Estimate Std. Error t value Pr(>|t|)    
(Intercept)           0.24600    0.02894    8.50 1.96e-14 ***
I(Species)versicolor  1.08000    0.04093   26.39  < 2e-16 ***
I(Species)virginica   1.78000    0.04093   43.49  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2047 on 147 degrees of freedom
Multiple R-squared:  0.9289, Adjusted R-squared:  0.9279 
F-statistic:   960 on 2 and 147 DF,  p-value: < 2.2e-16
```

ANCOVA: species and petal length predictors

```
Coefficients:
                     Estimate Std. Error t value Pr(>|t|)    
(Intercept)          -0.09083    0.05639  -1.611    0.109    
Petal.Length          0.23039    0.03443   6.691 4.41e-10 ***
I(Species)versicolor  0.43537    0.10282   4.234 4.04e-05 ***
I(Species)virginica   0.83771    0.14533   5.764 4.71e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1796 on 146 degrees of freedom
Multiple R-squared:  0.9456, Adjusted R-squared:  0.9445 
F-statistic: 845.5 on 3 and 146 DF,  p-value: < 2.2e-16
```

In-class model interpretation

Beyond the Group 1 methods

Real data is complicated, poorly-behaved, and noisy!

Simple modifications of Group 1 models:

You can try:

1. Data transformation
2. Adding polynomial or power terms
3. Adding interaction terms

Each option has pros and cons

Data transformations

Can help with:

1. Stabilizing the variance: log transformations
2. Linearizing the relationship

Data transformations: challenges

Transformations affect both the deterministic and stochastic model components

Transformed model coefficients can be difficult to interpret or explain to others.

Coefficients are now in terms of proportional increases/decreases not constant amounts.

It's not straightforward to 'back-transform' coefficients.

Log transformed response

Example interpretation of linear slope coefficient:

- “A 1 degree increase in temperature was associated with a beetle population density increase of approximately 130 beetles per hectare.”

Example interpretation of a log-transformed response:

- “A 1 degree increase in temperature is associated by a 7.4% increase in bark beetle population”

Additional model terms

Polynomial regression: raise predictor variable to a power, but parameters are still linear.

Interacting predictors:

- β1 = 2: 1-unit increase in predictor 1 associated with 2-unit increase in response.
- β2 = 2.4: 1-unit increase in predictor 2 associated with 2.4-unit increase in response.
- What if we simultaneously increase predictor 1 and 2 by one unit?

Beyond Simple Linear Models

More sophisticated models are needed when simple adjustments cannot address:

- Nonlinear relationships
- Heterogeneity: nonconstant variance
- Non-normal errors
- Non-independent observations

Challenge 1: non-linear relationship

NLS: Nonlinear Least Squares

GLM: Generalized Linear Models

GAM: General Additive Models, i.e. smoothing models

Challenge 1: non-linear relationship

NLS, GLM, GAM still require:

1. Constant variance: no heterogeneity
2. Normally-distributed errors
    - GLMs can accommodate certain types of non-normal errors
3. Independent observations

Challenge 2: Heterogeneity

GLS and GNLS: Generalized (Nonlinear) Least Squares

GLS/GNLS still require:

1. Independent observations
2. Normally-distributed errors

Custom models with custom variance/covariance structures

- A difficult (but not impossible) field!
- Zuur 2009 has some good descriptions and examples.

Challenge 3: Non-independent observations

Violates the assumption of independent, randomized sampling.

Results in data with **lower information content**.

- This seems really strange.
- Can we reason out why this might be?

Nonlinear Least Squares

Useful with nonlinear functions such as Ricker, logistic, any other nonlinear mechanistic function we can propose!

Least squares optimization criterion

- Find model parameter values tht minimize the sum of squared residuals

Nonlinear Least Squares: challenges

Needs numerical methods to estimate parameters

1. Relies on initial guesses for parameter values
2. Poor guesses can converge to local maxima

Very sensitive to outliers

1. Uses squared errors (like Group 1 methods)

Additive Models: GAM

What if we have no theoretical or mechanistic model for our system?

Smoothers can fit a phenomenological model to any dat.

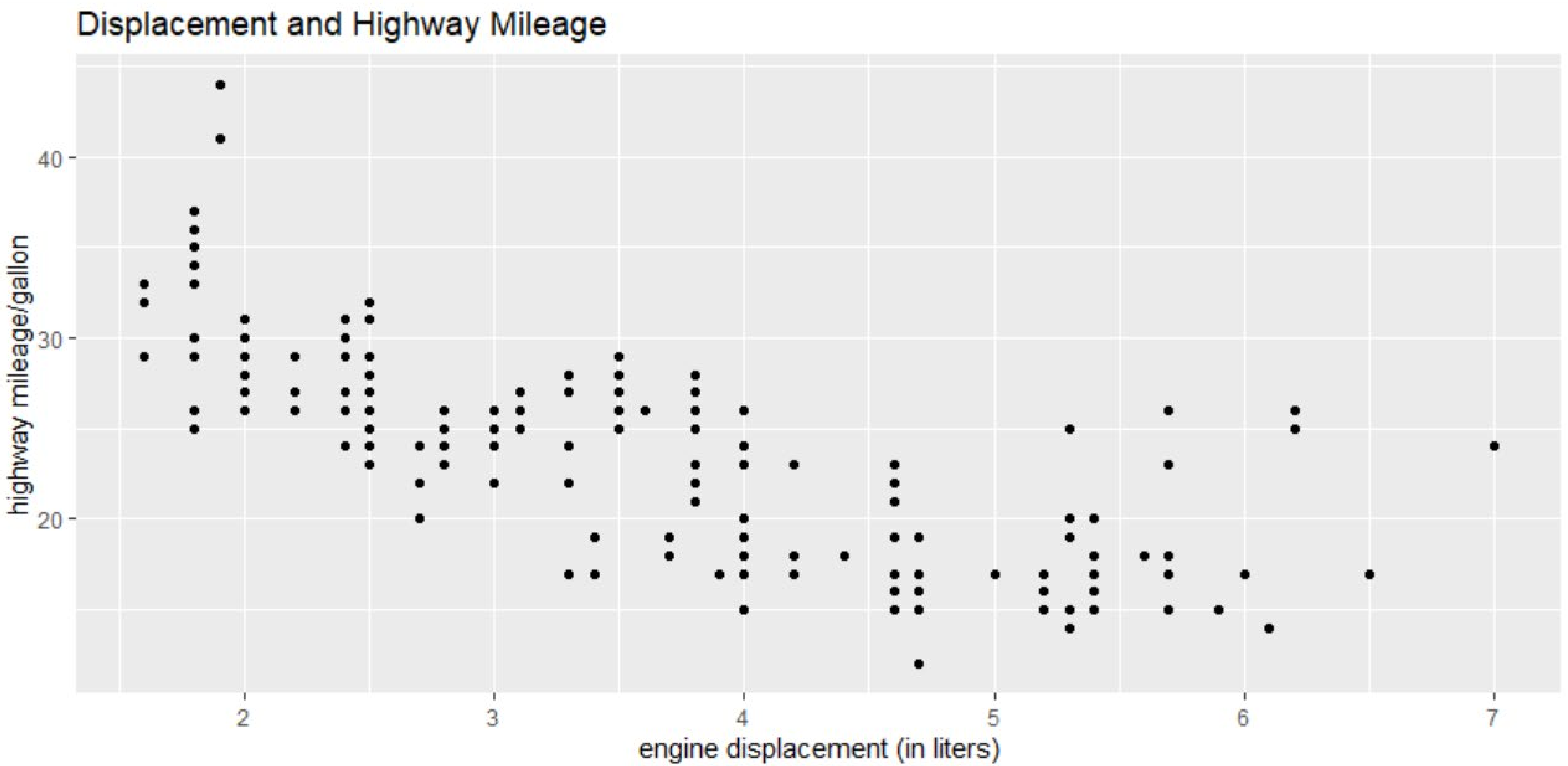
Additive Models: GAM

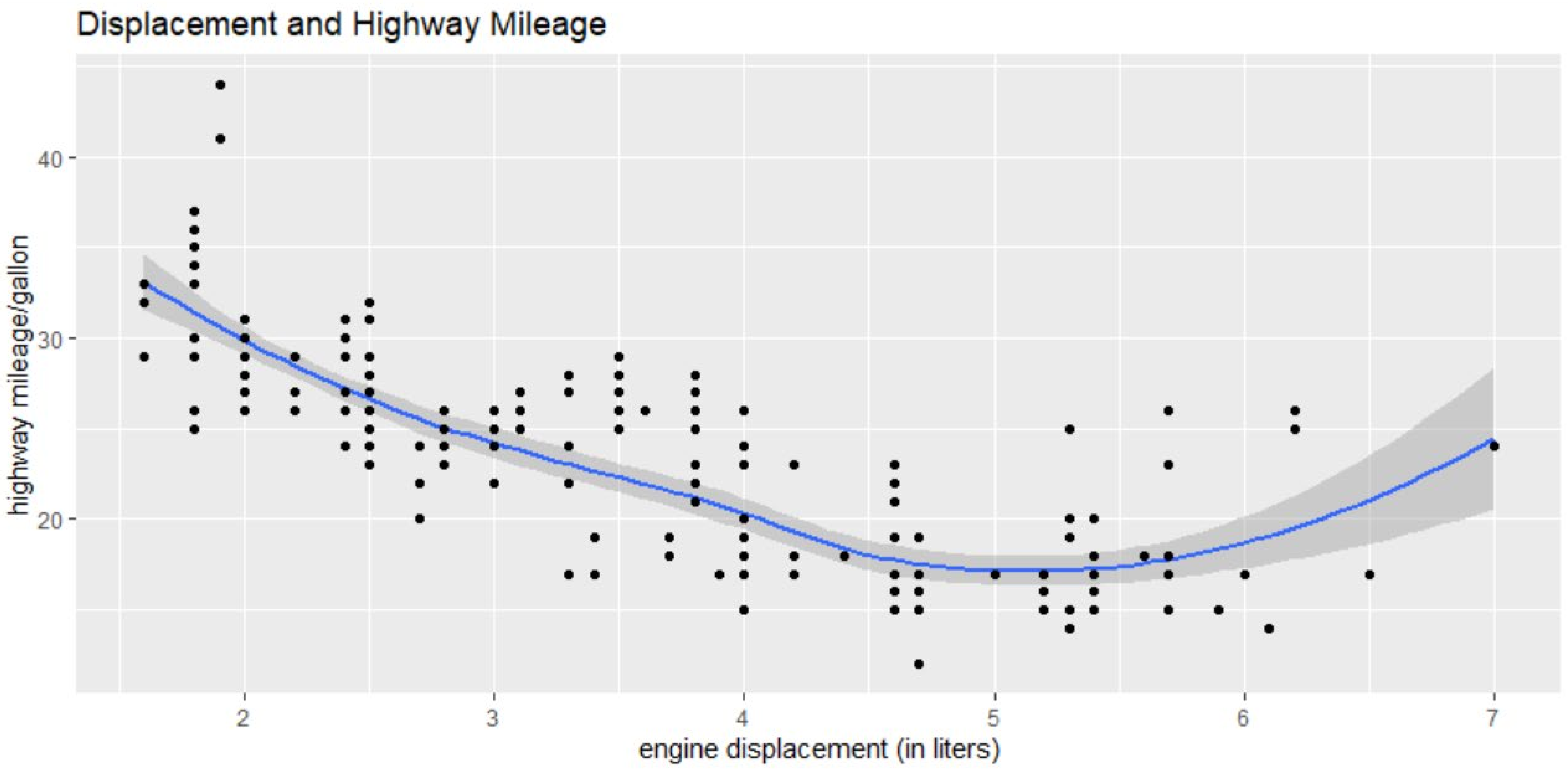
Local regression: general idea

1. for each point on the parameter space, calculate a new regression using a subset of points.
2. Give greater importance to nearby observations

Locally Weighted Regression – LOESS/LOWESS

Splines


Displacement and Highway Mileage
highway mileage/gallon
40
30
20
2
3
4
5
6
7
engine displacement (in liters)


Displacement and Highway Mileage
highway mileage/gallon
40
30
20
2
3
4
5
6
7
engine displacement (in liters)

Generalized Linear Models GLM

Unfortunate terminology similarity:

- General Linear Models
- Generalized Linear Models

Useful for binary and count response data

Can handle heterogeneity in the errors

Logistic and Poisson regression

Generalized Linear Models GLM

Allow non-constant variance

Allow certain kinds of nonlinear functions

- Exponential family of functions

Allow certain kinds of non-normality in the errors

Critical paper review: group activity

Choose group paper to use

For next time:

Continuation of methods to deal with more complicated data: McGarigal chapters 11a and 11b

Introduction to multivariate analyses: McGarigal chapter 11c