ECO 602 Analysis of Environmental Data

FALL 2019 – UNIVERSITY OF MASSACHUSETTS

DR. MICHAEL NELSON

Today's Agenda

- Beyond general linear models: challenges
  - Nonlinearity
  - Non-Normal errors
  - Heterogeneity
  - Autocorrelation
- Finish Group 1 model interpretation activity
- Non-independent data

No question set 4

Instead, we're going to do a series of in-class model interpretation activities/quizzes.

Please upload a pdf of your critical paper review papers to Moodle.

Thursday's in-class activity will be customized by group.

Terminology clarification

Critical distance, critical difference, critical value

I was sloppy with terminology that I used in lecture and question set 3.

Proper terms are:

- 1-sample t-test: critical value
- 2-sample t-test: critical difference in means

Recall the key group 1 assumptions and limitations.

Linear relationships

Normal errors

Constant variance

Independent observations

Single response

Simple modifications of Group 1 models:

You can try:

1. Data transformation (usually the logarithm)
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

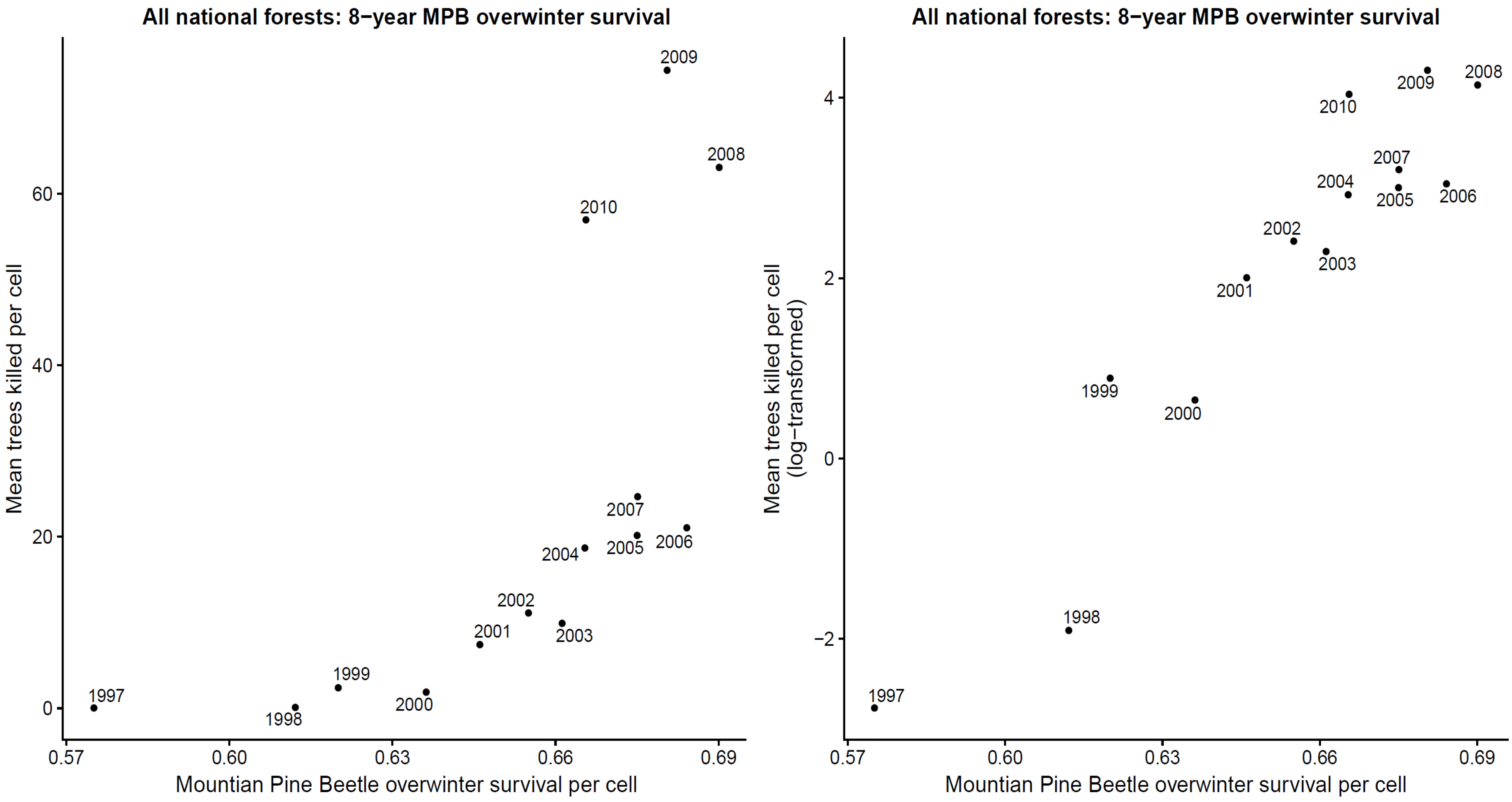
Log transformations: challenges

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.


All national forests: 8−year MPB overwinter survival
Mean trees killed per cell
Mountian Pine Beetle overwinter survival per cell
0
20
40
60
0.57
0.60
0.63
0.66
0.69
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
All national forests: 8−year MPB overwinter survival
Mean trees killed per cell
(log−transformed)
Mountian Pine Beetle overwinter survival per cell
−2
0
2
4
0.57
0.60
0.63
0.66
0.69
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010

Coefficient interpretations

Linear slope coefficient:

- "Every 1% increase in survival was associated with 2 additional killed trees per hectare per year."

Log-transformed coefficient:

- "Within a stand, a 1% increase in beetle survival was associated with a 6% proportional increase in tree mortality rate over the mortality rate of the previous year."

Additional model terms

Polynomial regression

1. Raise predictor variable to a power,
2. Nonlinear predictor/response relationship
3. Model parameters are still linear.

Interaction terms

- Example model: $Y_i = 1.3 + 2.0x_1 + 2.4x_2 + 2x_1x_2$
- 1-unit increase in predictor 1 associated with 2-unit increase in response.
- 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?

Interaction terms

- Example model: $Y_i = 1.3 + 2.0x_1 + 2.4x_2 + 2x_1x_2$
- What if we simultaneously increase predictor 1 and 2 by one unit?
- Without an interaction we would expect an increase of 4.4 (the **sum of beta1 and beta2**)
- With the interaction we get an increase of 6.4!

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

Finish interpretation activity from Tuesday

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:

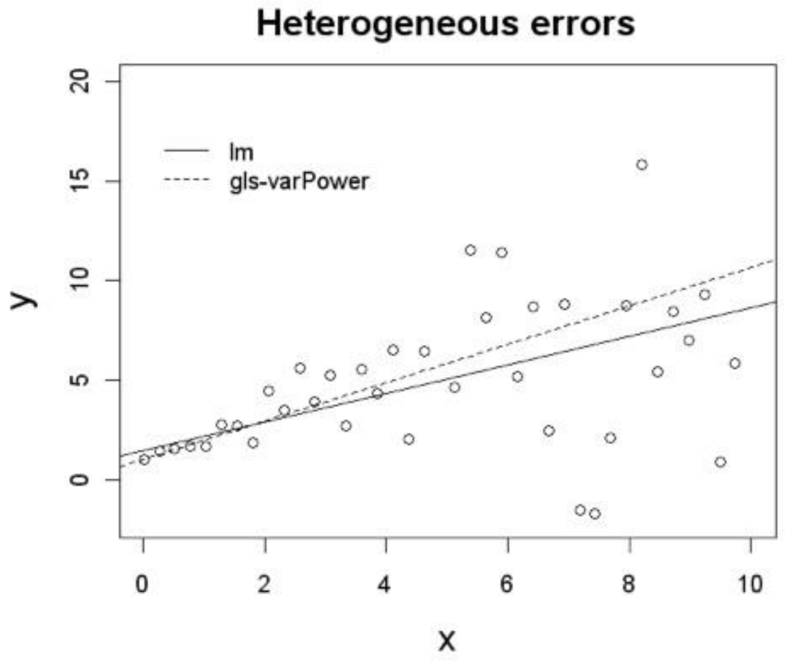
1. Constant variance: no heterogeneity
2. Normally-distributed errors
   - GLMs can accommodate certain types of non-normal errors
3. Independent observations

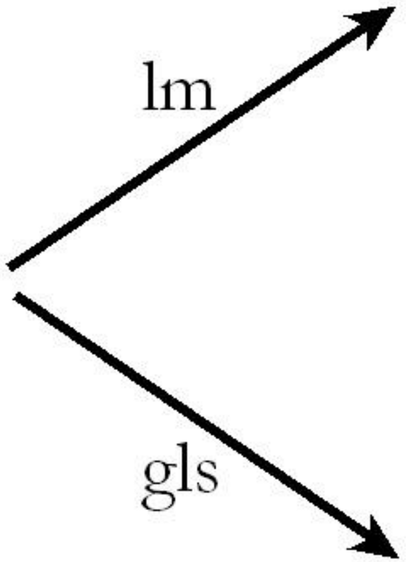
What is heterogeneity?

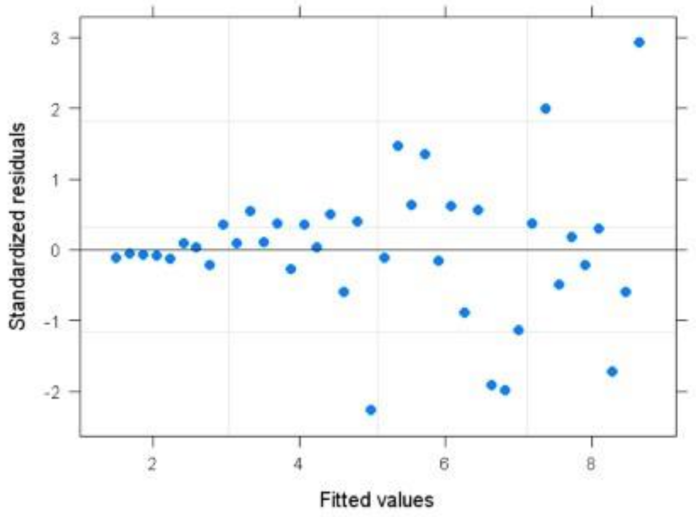
Landscape of Statistical Methods...

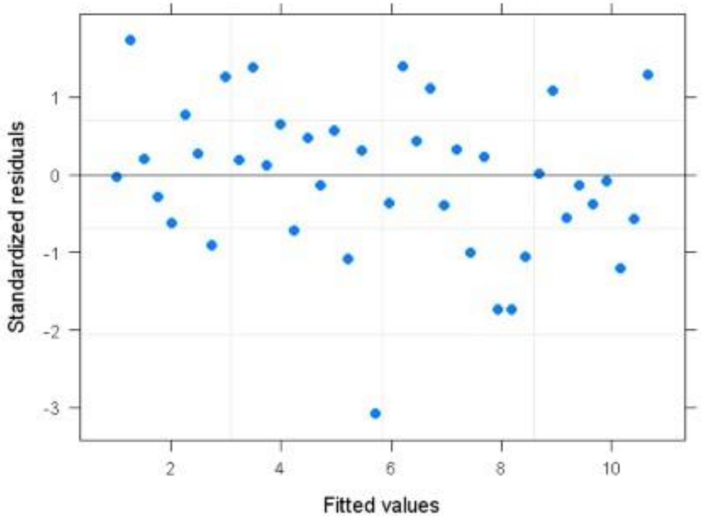
Dealing with heterogeneity

Linear model versus generalized least squares model:







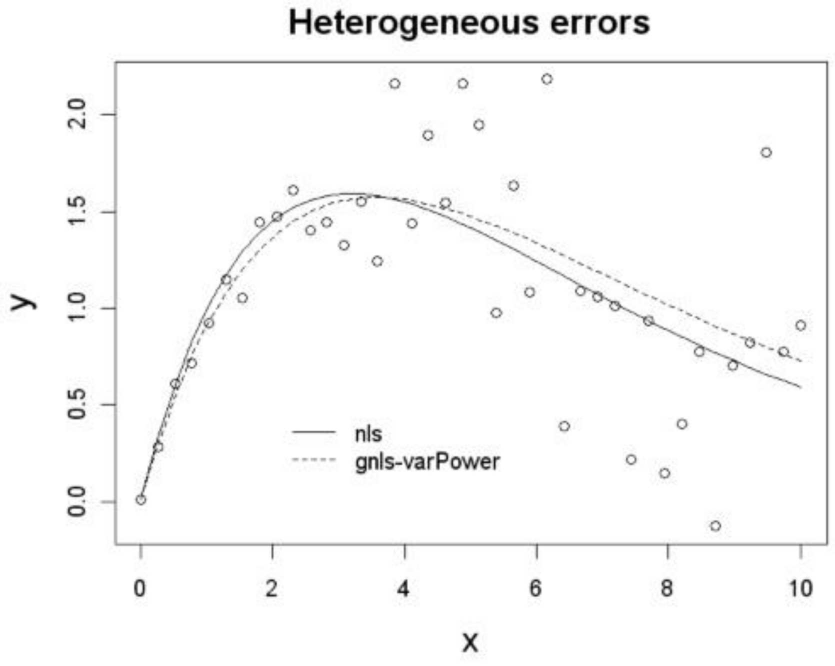



```
         Model df      AIC      BIC     logLik   Test  L.Ratio p-value
fit.lm       1  3 233.5692 238.4819 -113.78458
fit.gls2     2  4 195.3466 201.8969  -93.67328 1 vs 2 40.2226  <.0001
```

Landscape of Statistical Methods…

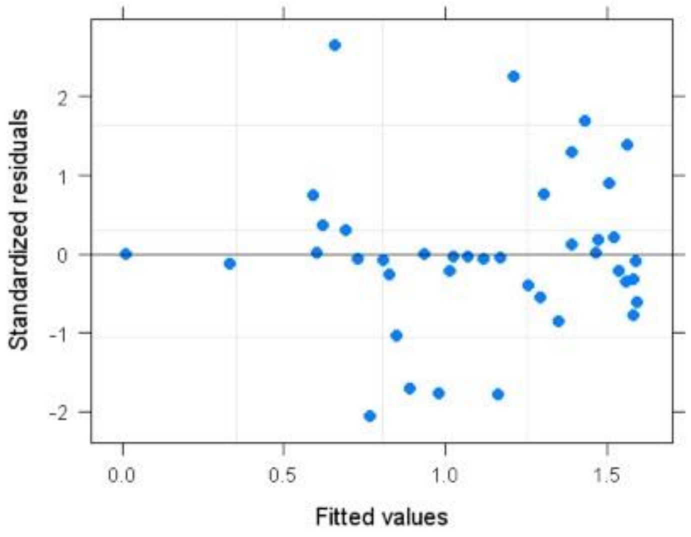
Dealing with heterogeneity

Nonlinear model with heterogeneity (generalized *nonlinear* least squares):

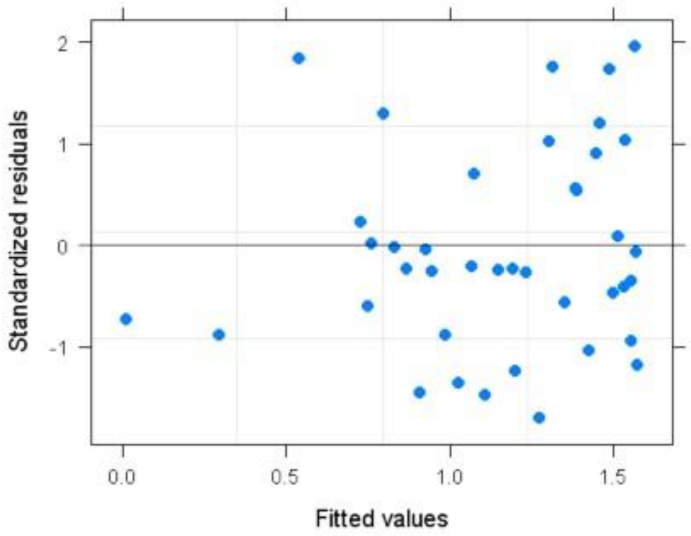


	df	AIC
fit.nls	3	50.52992
fit.gnls1	4	14.20585

nls



gnls



Challenge 2: Heterogeneity

GLS and GNLS: Generalized (Nonlinear) Least Squares

GLS/GNLS still require:

1. Independent observations
2. Normally-distributed errors

Generalized Linear Models (GLM) can accommodate some kinds of heterogeneity.

Challenge 3: Non-Normal errors

Generalized Linear Models can accommodate some types of non-normal errors.

Especially useful for binary or count data

Data transformations can sometimes fix non-normal errors.

- But data transformations cause interpretation difficulties.

Nonlinear Least Squares

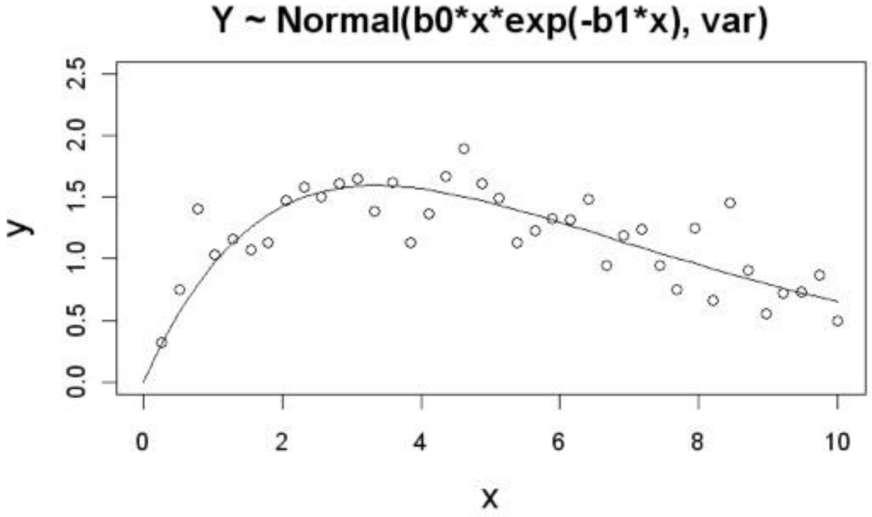
Useful with nonlinear functions such as Ricker, logistic, any other nonlinear mechanistic function we can propose!

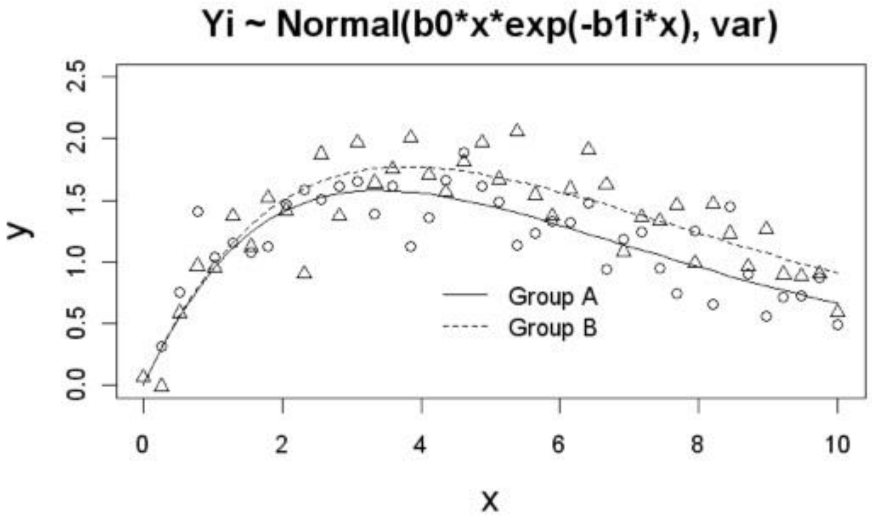
Least squares optimization criterion

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

Landscape of Statistical Methods…
Dealing with nonlinearity

4. Nonlinear least squares models (nls)
   - Relax the requirement of linearity (in the parameters) but keep the requirements of independence, normal errors and constant variance
   - Method: *numerical least squares*





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.

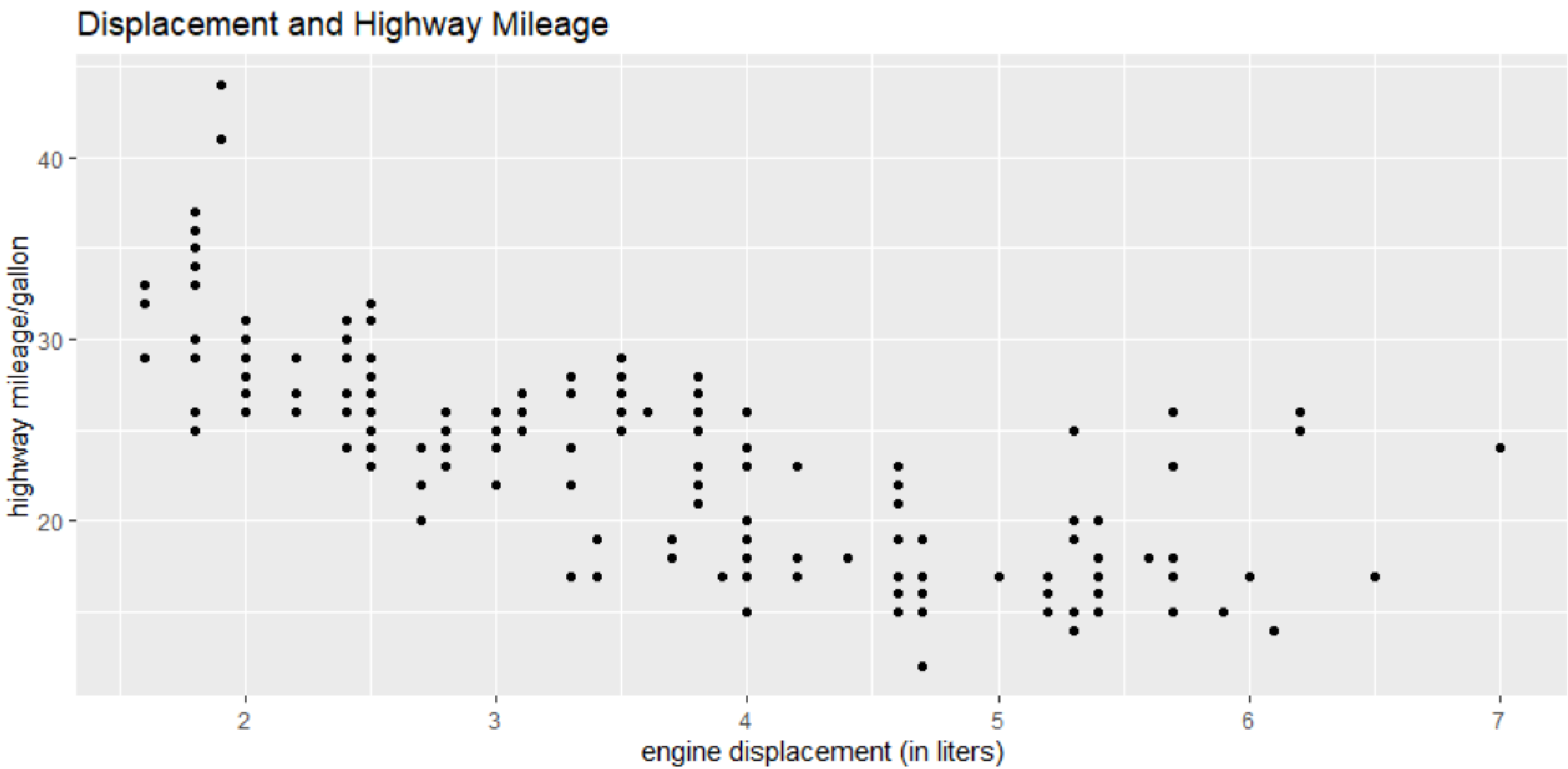
Additive Models: GAM

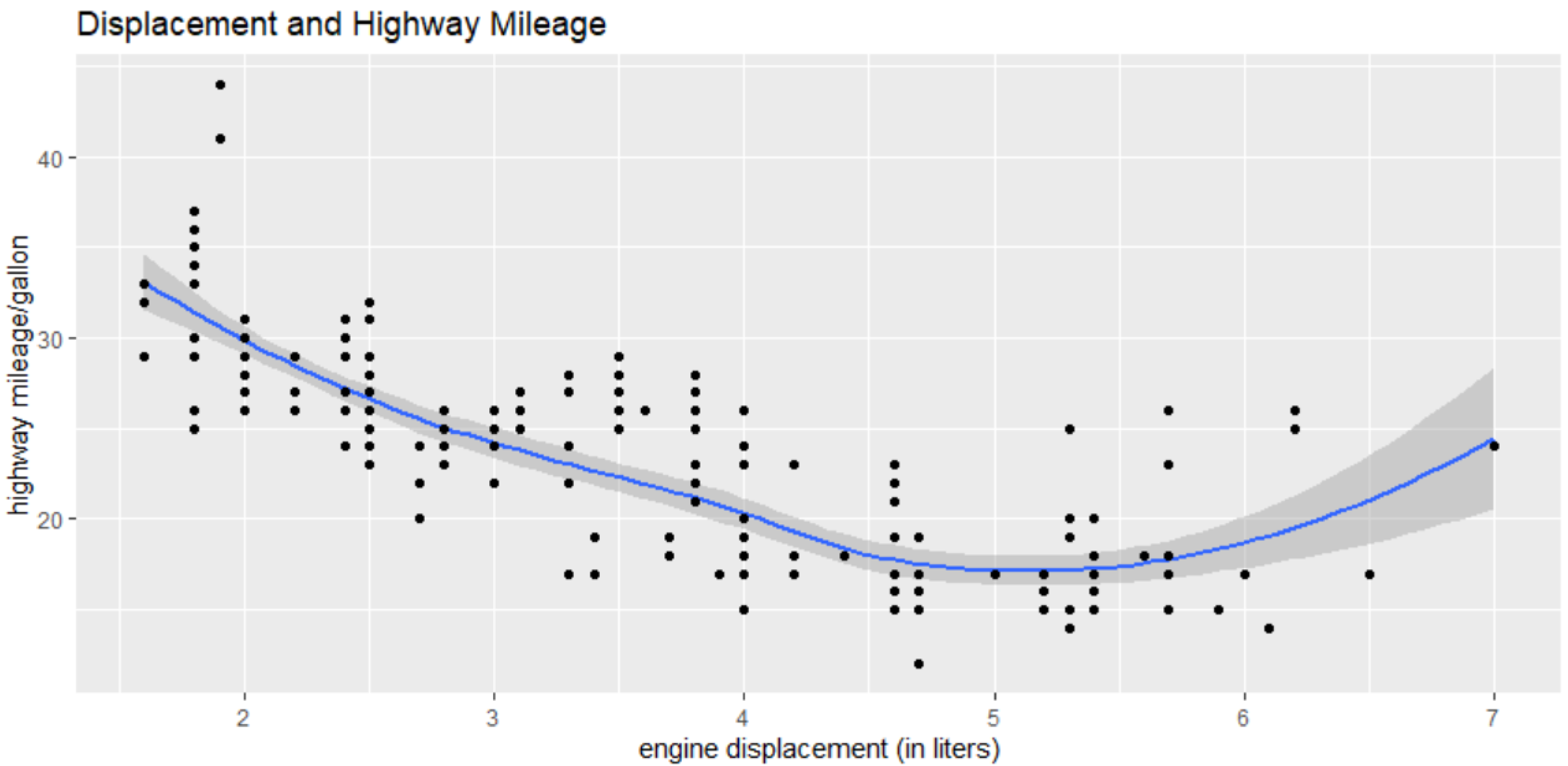
Local regression: general idea

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
- Genera**lized** Linear Models

Can handle heterogeneity in the errors

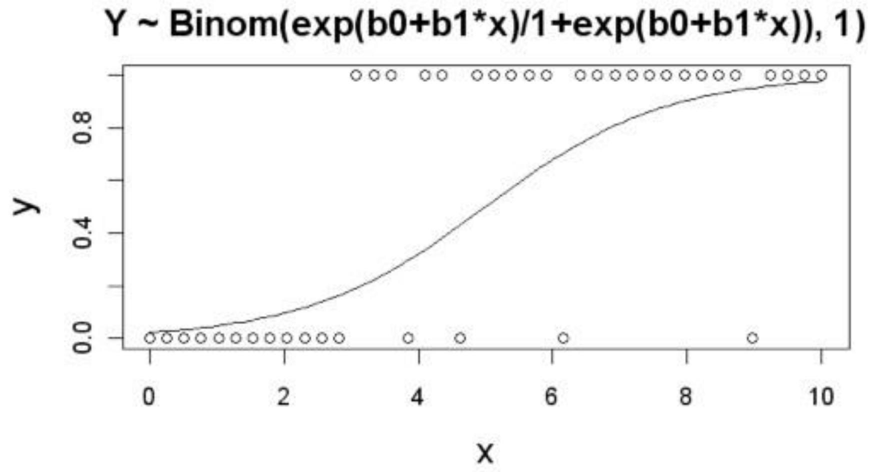
Extremely useful for binary and count data: logistic and Poisson regression

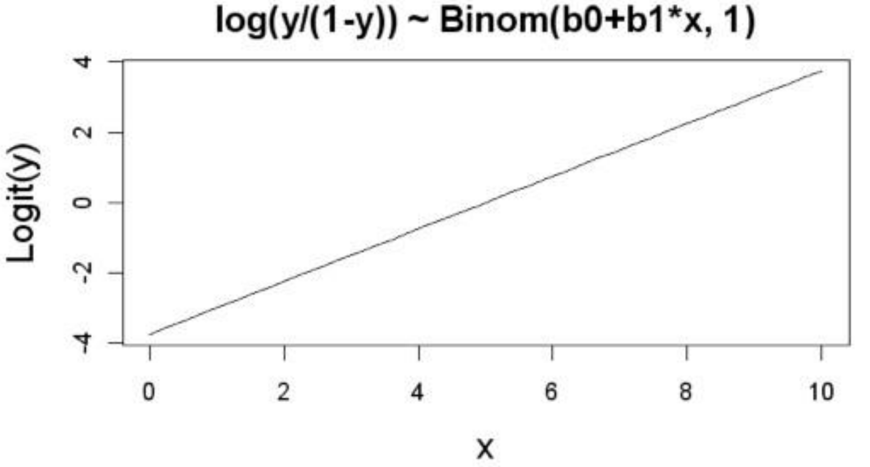
Useful with certain kinds of nonlinearity and non-normal errors

Landscape of Statistical Methods…

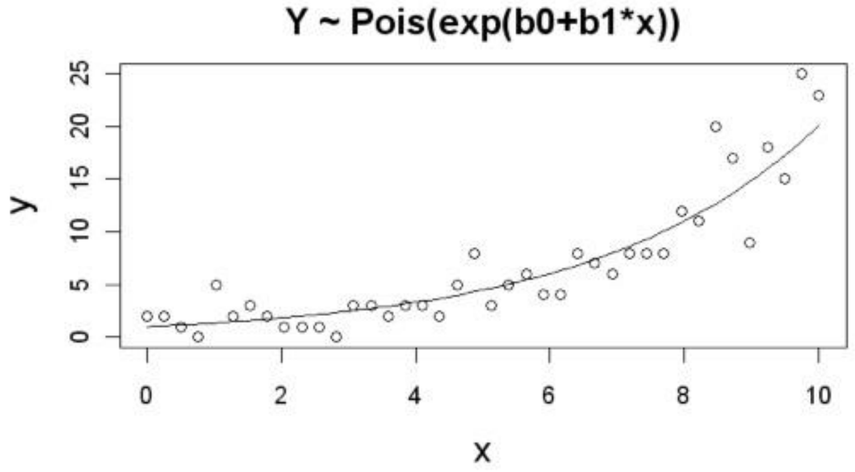
Generalized linear models (GLMs)

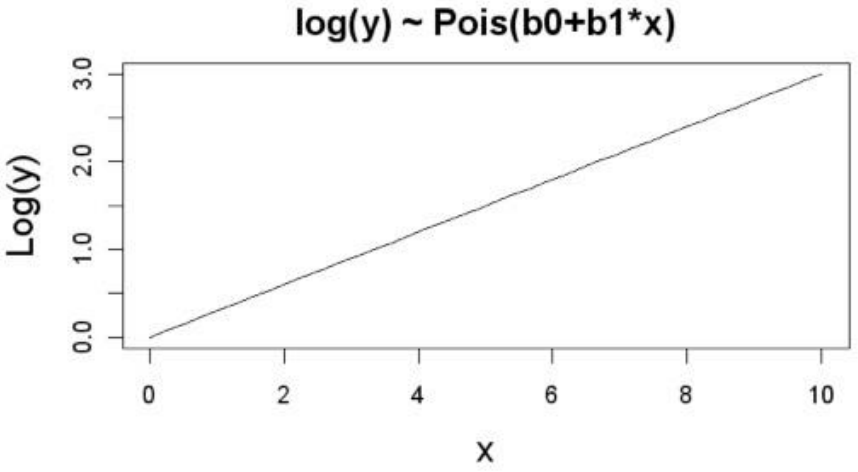
Logistic regression:





Poisson regression:





Challenge 3: Non-independent observations/errors

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?

Autocorrelation

Does the value of your current observation help you guess what you will observe next?

Observations nearby in space or time might be more similar than expected due to chance alone.

Walter Tobler's 1st law of Geography:

"Everything is related to everything else, but near things are more related than distant things."

Temporal dependence

Are observations close in time?

Can we guess the high temperature on July 28th, 2000 if we know the high temperature on July 27th, 2000?

Can we guess the high temperature on July 28th 2012?

Autoregressive order 1: AR1

Model assumes that the current observation is related to the immediately previous observation.

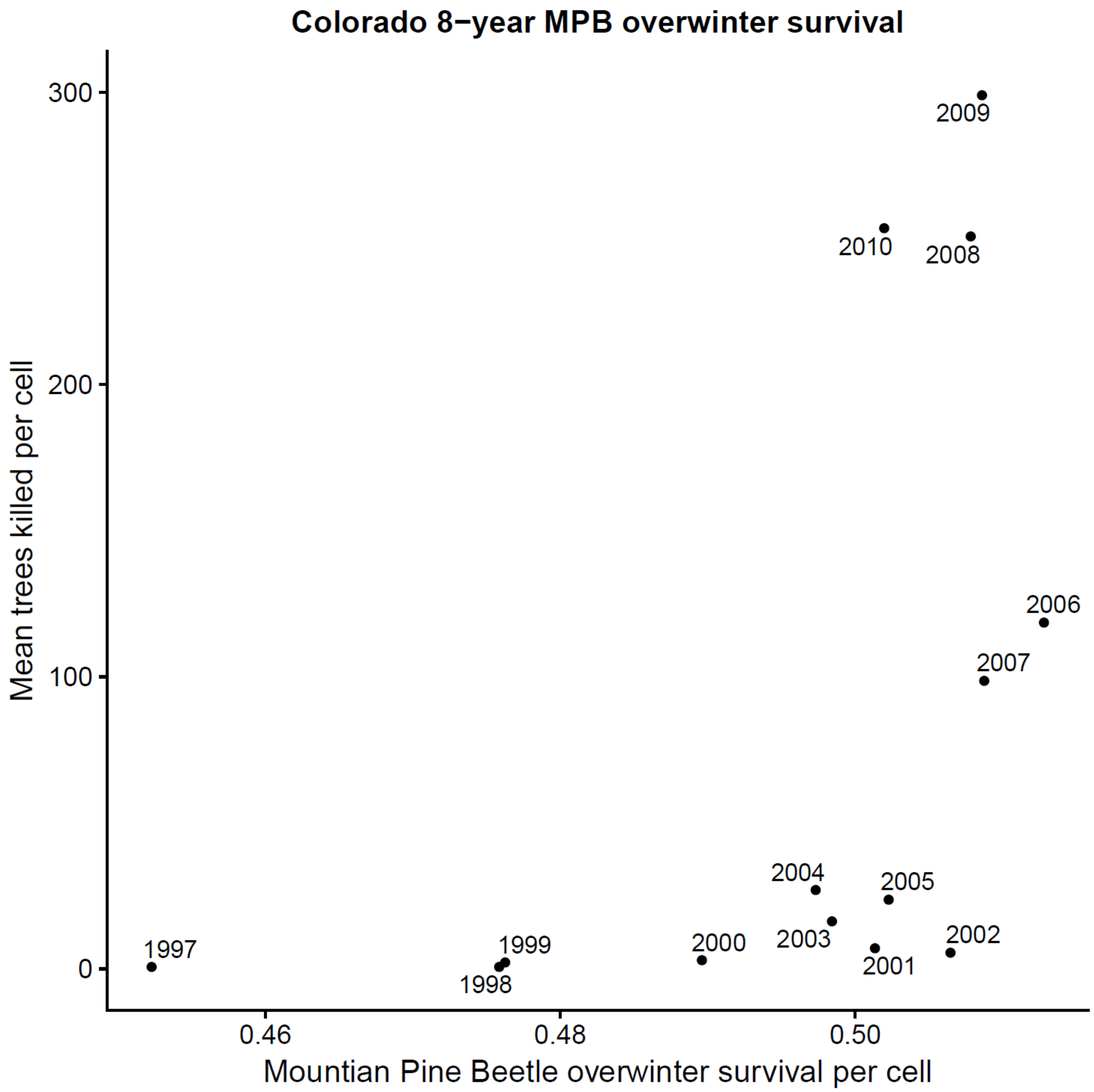
- Not correlated with observations more than 1 time-lag behind.
- Includes a model prediction term for the t-1 observation

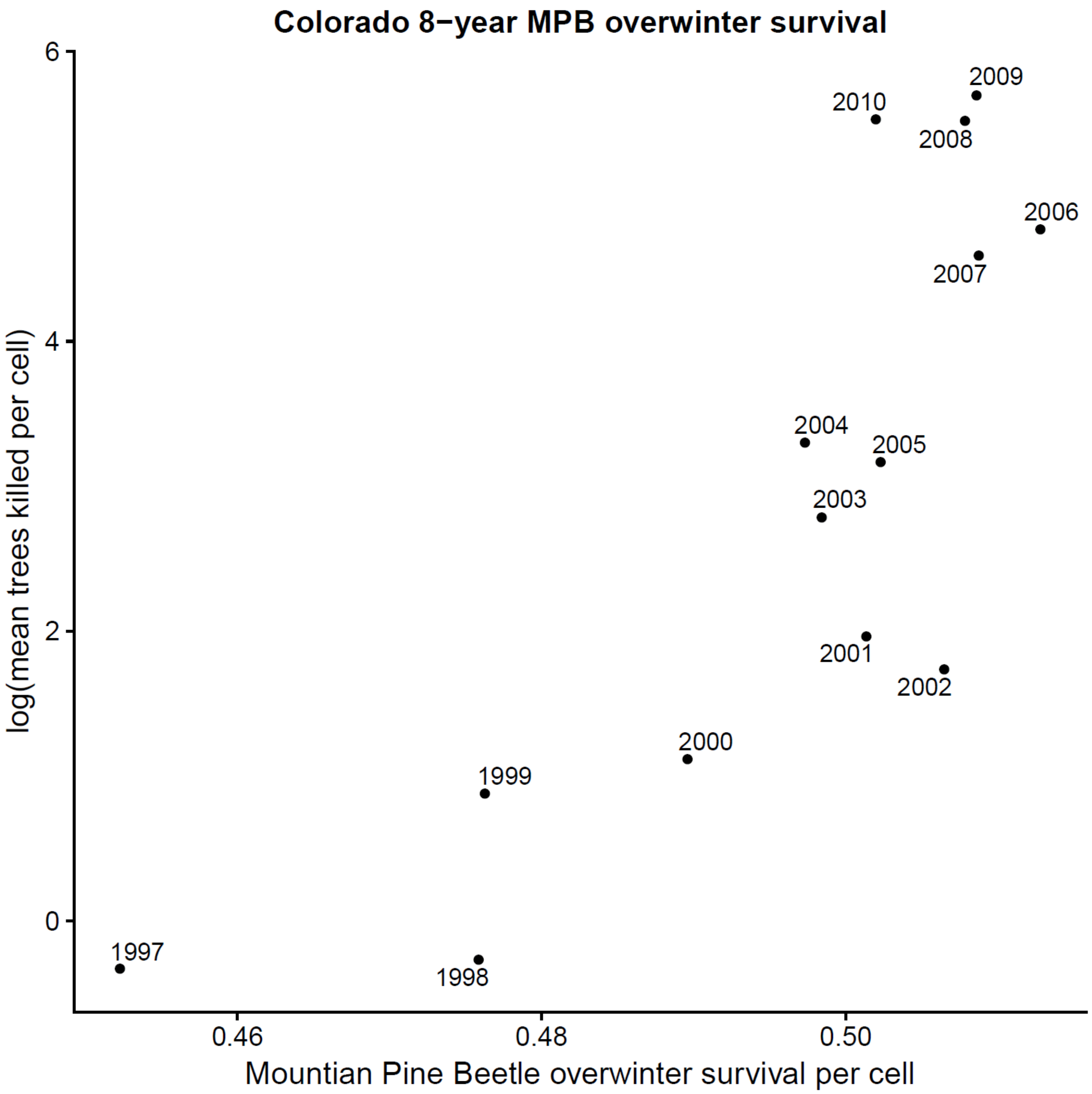
Temporal autocorrelation: Mountain Pine Beetle example

Question: Can we use the winter beetle survival rate to predict the number of trees killed in the following summer?

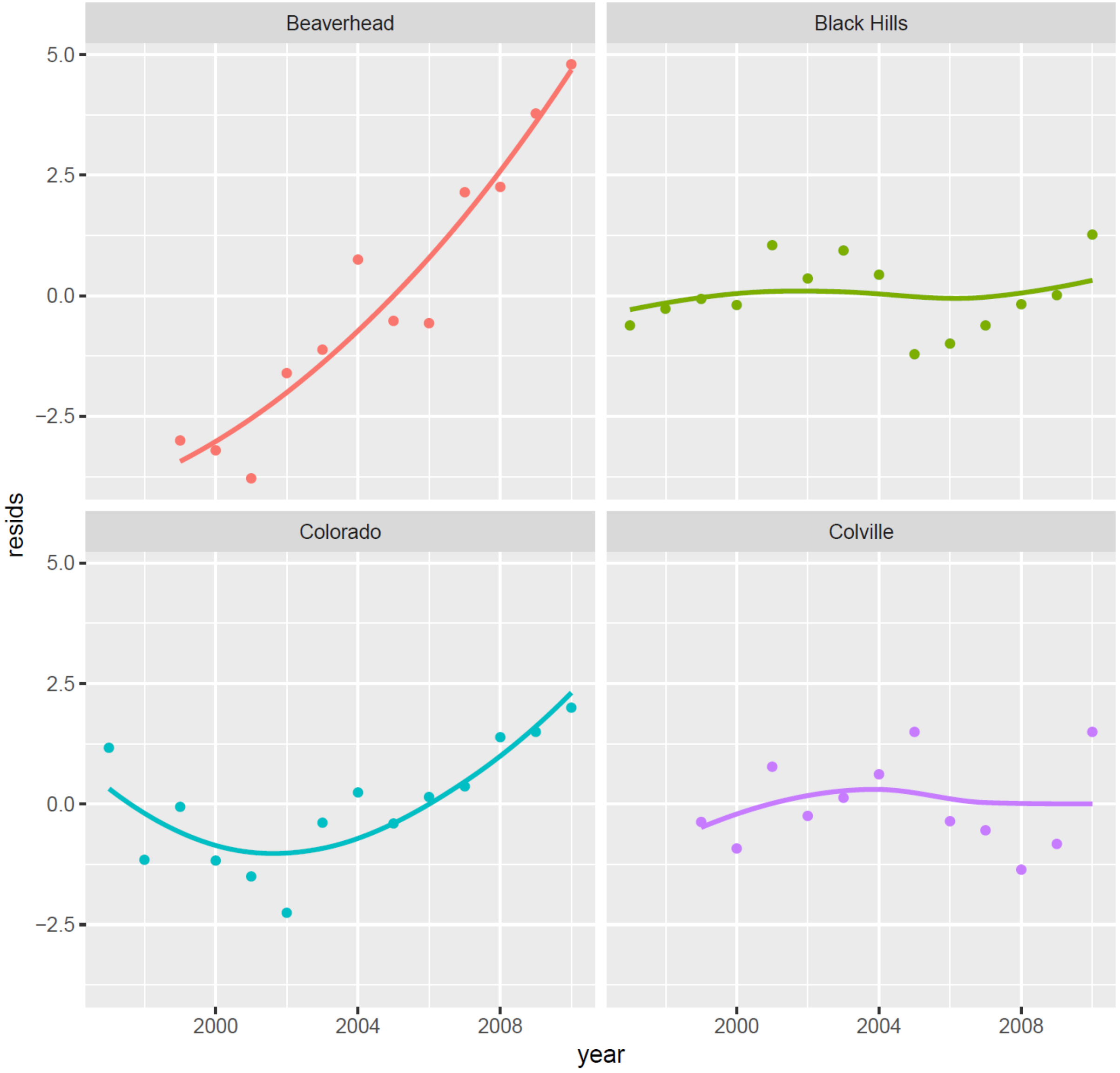
Two models:

- Simple linear model
- Simple linear model with AR(1) structure

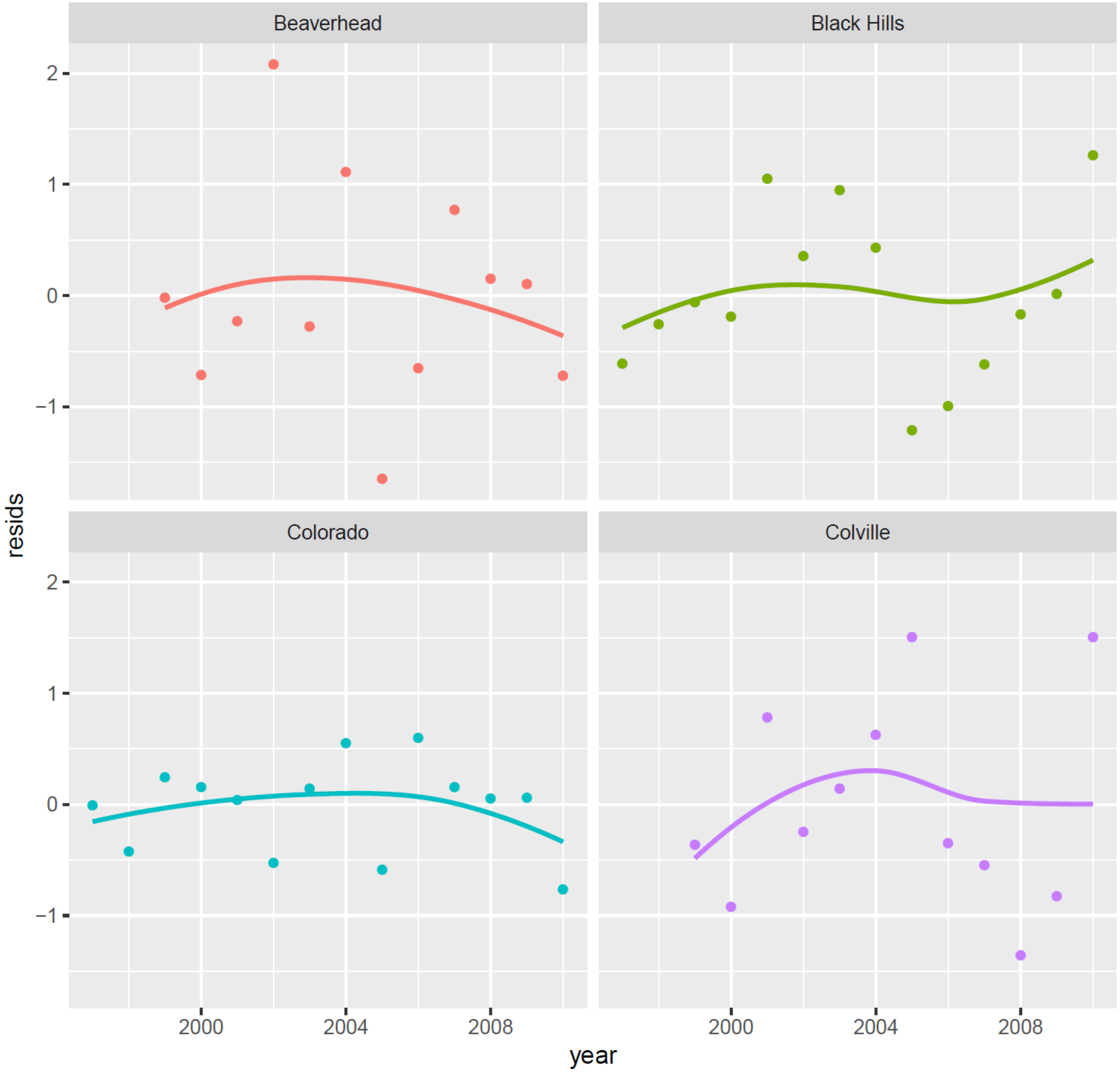

Colorado 8−year MPB overwinter survival
Mean trees killed per cell
Mountian Pine Beetle overwinter survival per cell
0
100
200
300
0.46
0.48
0.50
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010


Colorado 8−year MPB overwinter survival
log(mean trees killed per cell)
Mountian Pine Beetle overwinter survival per cell
0
2
4
6
0.46
0.48
0.50
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010

OLS residuals, mean mpb survival previous 8 years.



OLS with AR residuals, mean mpb survival previous 8 years.



Autoregressive order n: AR(n) AR(n) with moving average

Includes terms for n time lags

Moving average: adds a moving average

Spatial Dependence

Correlation among observations might decrease with increasing distance:

- Nearby observations are more similar than observations separated by large distances.
- Points separated by a **critical distance*** are not correlated.

*This is not an official term

We can use a **variogram** to quantify the spatial dependence

Landscape of Statistical Methods…

Dealing with (auto-)correlated errors

1. Spatial correlation – (semi-)variogram

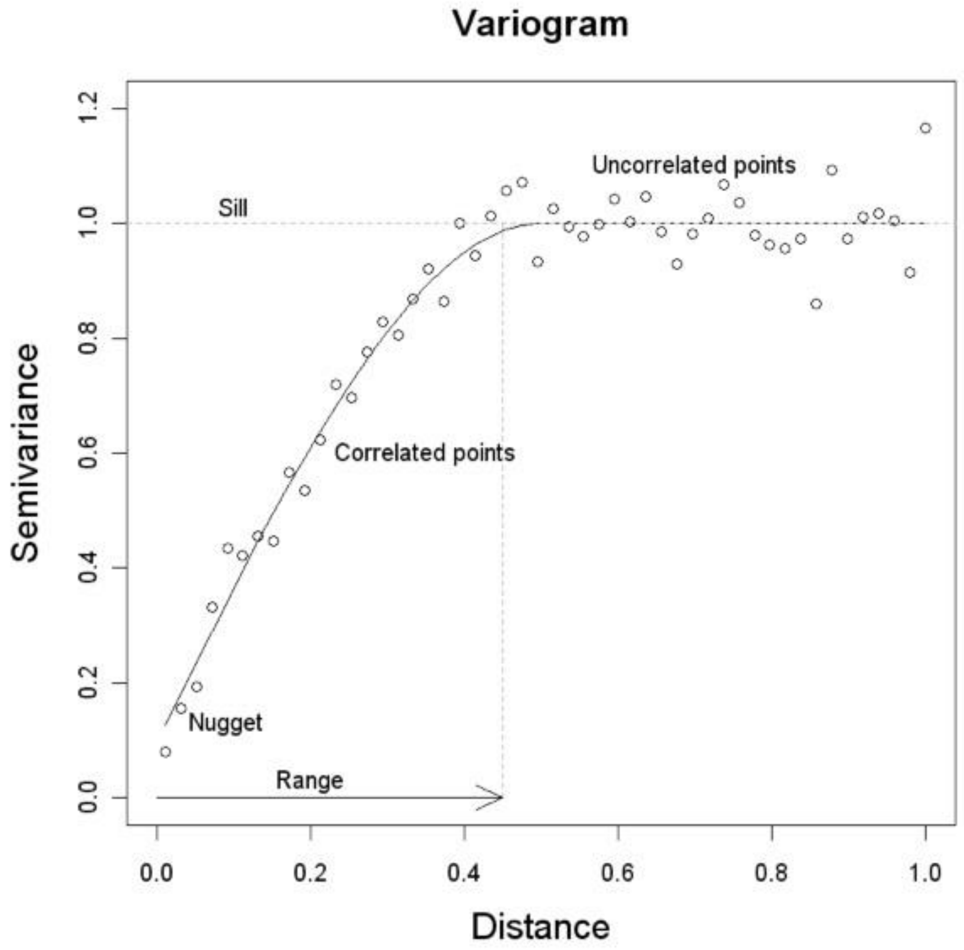
Variogram:

Semivariance:

$$\gamma(x_i, x_j) = \frac{1}{2} E\left[\left(Z(x_i) - Z(x_j)\right)^2\right]$$

Experimental variogram:

$$\hat{\gamma}(h) = \frac{1}{2N(h)} \sum_{i=1}^{N(h)} \left[z(x_i + h) - z(x_i)\right]^2$$



Regression with autocorrelated errors

Custom models with custom variance/covariance structures for heterogeneity,

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

For next time:

Make sure you have read your critical paper review/final project papers.

There will be an in-class activity/quiz on your group's paper.

We're going to talk about hierarchical models and an overview of multivariate methods

Read McGarigal Ch 12b and 12c