Tuesday Lecture:

Outline:

I. Adding deterministic components to our models ("regression")
   a. This is where the theoretician in me gets excited. These deterministic models are in many respects the bread-and-butter of theory. When we can link them up with stochastic processes and ultimately data, we're able to start testing theories.
   b. Continuous covariates (regression models):
      i. Linear regression
      ii. Quadratic regression
      iii. Nonlinear regression
      iv. EMD bestiary of functions
      v. Modeling variances
   c. Categorical/discrete covariates (ANOVA analogs)
      i. Model matrices
      ii. Contrasts
   d. Continuous + discrete covariates (ANCOVAs)
   e. Technological concerns
      i. Limits of numerical optimization
      ii. Finding good starting guesses:
         1. Plot things out
         2. Method of moments
         3. Simulate data
         4. Try a bunch and see what happens. (grid.mle2)

<break>

II. Model comparison/assessment
   a. AIC model comparison
      i. Akaike weights and weighted models.
      ii. Alternatives - AICc and BIC; emphasis on sample size and more conservative penalty for number of parameters.
   b. Likelihood ratio test and confidence intervals
      i. Chisq tests/distribution
      ii. Slices
      iii. Profiles
      iv. Fisher information CI's(?)
      v. Nested models and p-values/significance
      vi. R2 values? If time allows…
   c. Limitations of MLE approach
      i. Asymptotic normality
      ii. Potentially biased estimators
      iii. Bounded parameters
      iv. Numerical limitations

Tuesday lecture

Recall from yesterday that the poisson distribution

$$P(x|\lambda) = \frac{e^{-\lambda}\lambda^{x}}{x!}$$

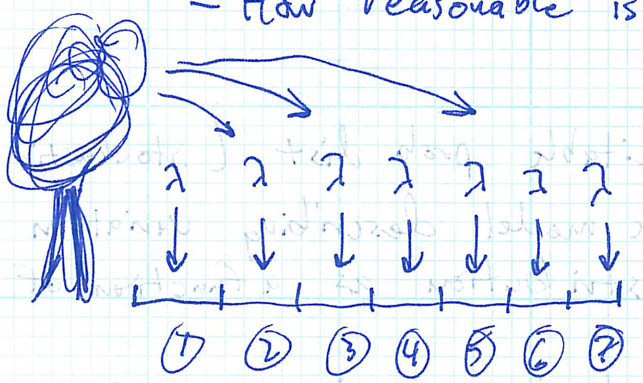
We used calculus + numerical approaches to come up with the maximum likelihood estimate of $\hat{\lambda}$, given a set of x data.

Remember that the poisson distribution models the distrib of number of individuals/arrivals/events/counts of things occuring in a unit of space/time/sampling effort, given a certain rate at which events occur.

Seeds falling in space (dispersal)

↳ we assumed that the rate of dispersal is uniform through space

– How reasonable is this?



It seems likely that the rate of seed fall through space is not constant.

Then, rather than thinking about modeling our X data as arising from a poisson distribution w/ constant lambda, we might want to allow λ to vary with distance from an adult tree. (assuming we also have data on these distances.

Instead of

$$X \sim Pois(\lambda)$$

We can consider $X \sim Pois(\lambda)$

where $\lambda \sim \exp(\lambda_0 + \lambda_1 \cdot distance)$

→ rate of dispersal decays exponentially with distance

Still just trying to model variation in X given now some extra info.

<u>This is a model that we can fit using MLE approach!</u>

- ~~additionally~~ We've just combined a <u>stochastic model</u> (pois. distrib) with a <u>deterministic model</u> / hypothesis describing how we believe a biological relationship works.
- additionally, we've generated now multiple models that we can compare (based on their likelihoods) as a way of drawing inferences about the process of seed dispersal.

This is where it gets exciting.

We can model data by combining <u>any</u> suitable prob dist (stochastic process) with a deterministic model describing variation in the parameters of the distribution as a function of some covariate.

- this includes all of the dist's we've covered + more
- pretty much any functional form you can think of (linear, nonlinear, even ODE's)
- many different kinds of covariates

<show Bolker fig 9.2, all_of_statistics.pdf> → Spans all of "generalized linear models", plus some, and builds the foundation of others beyond.

We can categorizes these options based on the kind of covariate that we want to include in our deterministic models of pdf parameters ↳ refered to generally here as Θ

① Continuous = Regression

- "Linear" regression $\Theta \sim b_0 + b_1 x$
- "Quadratic" regression $\Theta \sim b_0 + b_1 X + b_2 X^2$
- Multiple linear regression $\Theta \sim b_0 + b_1 X_1 + b_2 X_2$
  ↳ multiple covars

All "linear" from Stats perspective

- nonlinear regression $\Theta \sim f(X, b)$

  eg. $\Theta \sim \frac{aX}{1+abX}$ (type II functional response)

< EMD book, Table 3.1 >
↳ list of deterministic functions

② Discrete $\approx$ ANOVA (Categorical)

- Really essentially just linear models, but we make use of a coded form of the covariate.

Algae growth rate experiment
- 2 light levels (low, high)
- 3 temperatures (10°c, 20, 30)

① Typical data column, X, low, low, low, high, low, high, high
↳ can't do math on this

Recoded: $\bar{X} = 0, 0, 0, 1, 0, 1, 1$

Consider Model

$$\Theta \sim b_0 + b_1 \bar{X} \quad \begin{cases} b_0 & \text{when } \bar{X}=0 \text{ (low)} \\ b_0+b_1 & \text{when } \bar{X}=1 \text{ (high)} \end{cases}$$

$$\begin{cases} b_0 + b_1 \cdot 0 \\ b_0 + b_1 \cdot 1 \end{cases}$$

We can fit this (essentially linear) model to our data via mle as above, obtaining estimates for the value of

$$\theta_{low} = b_0$$

$$\theta_{high} = b_0 + b_1$$

If we treat algal growth as normally distributed and consider $\theta = \mu$, mean of normal distrib, We've essentially recreated a one-way ANOVA w/ 2 treatment levels

(2) Temperature only
- treat as categorical (but could treat as cont...)

Now instead of coded column, we need a matrix, called a model matrix

Temp	T1	T2
10	0	0
10	0	0
20	1	0
20	1	0
30	0	1
30	0	1

$$\theta \sim b_0 + b_1 T_1 + b_2 T_2$$

when $T_1 = 0$ + $T_2 = 0$ (Temp = 10)

estimates of treatment means.
$$\left\{\begin{array}{l} \theta \sim b_0 + b_1 \cdot 0 + b_2 \cdot 0 = b_0 \\ \theta \sim b_0 + b_1 \cdot 1 + b_2 \cdot 0 = b_0 + b_1 \\ \theta \sim b_0 + b_1 \cdot 0 + b_2 \cdot 1 = b_0 + b_2 \end{array}\right.$$

Note that we're estimating trt. means, not predicting quantitatively how much Temp should ↑ or ↓ growth rates
→ works even on non-ordinal data.

Alternately,

Temp	T_1	T_2	T_3
10	1	0	0
10	1	0	0
20	0	1	0
20	0	1	0
30	0	0	1
30	0	0	1

called a model matrix

$$\theta \sim b_1 \cdot T_1 + b_2 T_2 + b_3 T_3$$

individ. pars interpretable as means, but
↳ less efficient,
↳ get same answers *

③ Light + Temperature (2-way ANOVA)

$\boxed{A}$ - with out an interaction, (main effects only)

$$\theta \sim b_0 + b_1 \bar{X} + b_2 T_1 + b_3 T_2$$

$\theta \sim b_0$ = low light, Temp = 10

$\theta \sim b_0 + b_2$ = " " Temp = 20

$\theta \sim b_0 + b_3$ = " " Temp = 30

$\theta \sim b_0 + b_1$ = high light, Temp = 10

$\theta \sim b_0 + b_1 + b_2$ = " ", Temp = 20

$\theta \sim b_0 + b_1 + b_3$ = " ", Temp = 30

get Students to help fill in

$\boxed{B}$ w/ an interaction (ignore 1 level of temp)

Light	Temp	X_1	X_2	X_3
low	10	0	0	0
low	20	0	1	0
~~high~~				
high	10	1	0	0
high	20	1	1	1

X_1: main effect of light
X_2: main effect of Temp
X_3: extra effect of ↑ light + ↑ temp.

$$\theta \sim b_0 + b_1 X_1 + b_2 X_2 + b X_3$$

$\theta \sim b_0$ = low light, Temp = 10

$\theta \sim b_0 + b_1$ = ~~high~~ high light, Temp = 10

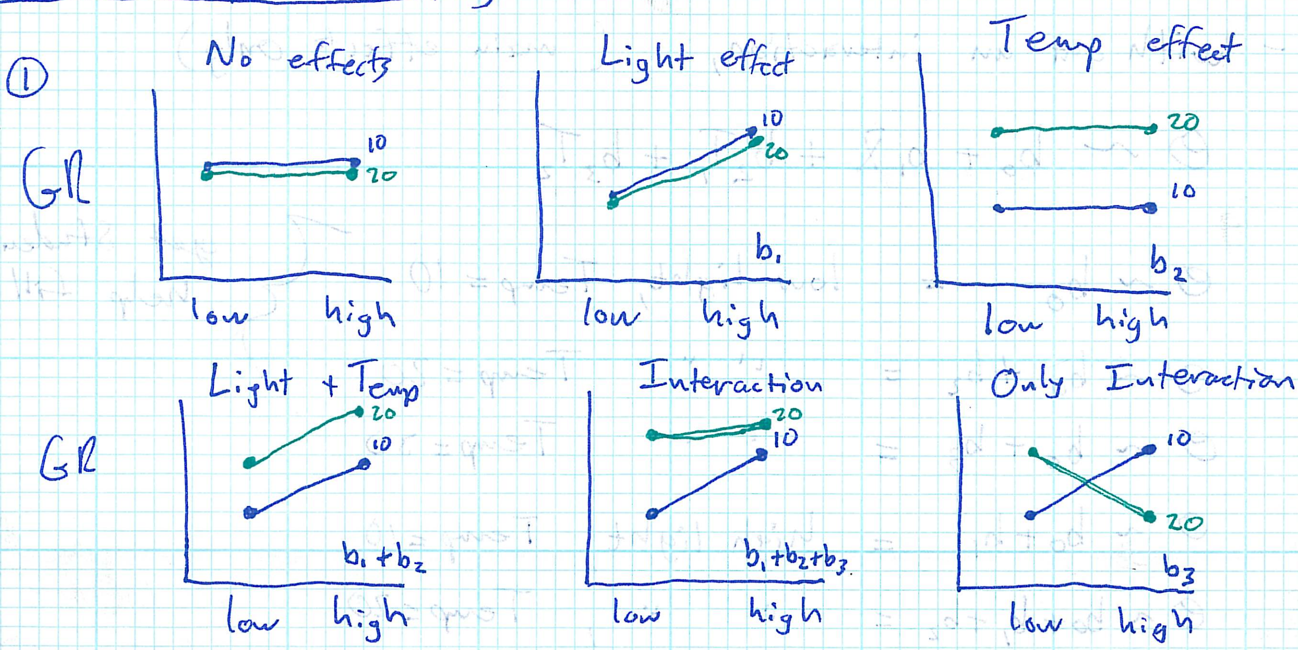
$\theta \sim b_0 + b_2$ = low light, Temp = 20

~~$\theta \sim b_0 + b_3$ =~~

$\theta \sim b_0 + b_1 + b_2 + b_3$ = high light, Temp = 20

If $b_3 \neq 0$, ~~then~~ we get a non additive treatment effect, an interaction, where individual effects of $b_1 + b_2$ aren't sufficient for explaining ↑ light, ↑ Temp

Last example Visually

① 

Cases all determined by whether or ~~not~~ the b's are different from 0.

$$\begin{cases} b_1 \neq 0 \Rightarrow \text{light effect} \\ b_2 \neq 0 \Rightarrow \text{Temp. effect} \\ b_3 \neq 0 \Rightarrow \text{interaction} \end{cases}$$

Can be extended to account for more covariates and more interactions, although # parameters start to ↑

Ends our survey of modeling discrete data

→ Note Normality assumption isn't required for MLE, just for consistency w/ ANOVA tradition.

---

③ Discrete + Continuous can be mixed together (ANCOVA)

What if we thought growth rate increased w/ temp. linearly?

$$\theta \sim b_0 + b_1 X + b_2 \cdot \text{Temperature}$$

where X is (0, 0, 1, 1, etc) coding low vs. high light.

Ⓐ What might the advantage of this be?

Discuss