Modeling PalEON biomass

Wesley Brooks

UW-Madison

May 20, 2013

Outline

1. Data
   - Overview of the data
   - Models
2. Methodological details
   - Sources of randomness
   - Branching process details
   - Estimating gene copies from qPCR
   - Variance of the estimator
3. Analysis of experimental data
   - Luteinizing hormone
   - S. vulgaris

Goal

- Produce a model of per-species biomass at time of settlement

Table of Contents

1. Data
    - Overview of the data
    - Models

2. Methodological details
    - Sources of randomness
    - Branching process details
    - Estimating gene copies from qPCR
    - Variance of the estimator

3. Analysis of experimental data
    - Luteinizing hormone
    - S. vulgaris

Data

- Computed from settlement-era survey
- Working with composition, biomass, and stem density

qPCR as a branching process

- PCR is controlled so that each replication cycle k is discrete
- Each particle either doubles or does not during each trial
  - Probability of replication is typically high ($0.9 < p$)
- This defines a supercritical branching process that leads to exponential growth
- During early cycles ($k < 15$, say), the count is obscured by noise
- Availability of reaction chemicals attenuates the reaction after ~ 30 cycles
- The cycles between 15 and 30 are called the exponential phase.

Models

There are two divisions for modeling biomass data:

- One-stage vs. two-stage
- Smoothing splines vs. GMRF

Two-stage models

- First stage: zero/non-zero
  - Logistic regression
  - $Z \sim \text{Bernoulli}(\gamma)$
  - $\text{logit}(\gamma) = f(x, y, p_k)$
- Second stage: distribution of positive biomass
  - $Y|Z = 1 \sim \text{Gamma}(\alpha, \beta)$
  - $\text{E}(Y|Z = 1) = \mu = \alpha\beta = f(x, y, p_k)$

Tweedie model

The Tweedie model is a Gamma-Poisson mixture.
How to visualize a Tweedie random variable:

- Draw $N \sim \text{Poisson}(\lambda)$
- Now make N iid draws: $V_\ell \sim \text{Gamma}(\alpha, \beta)$
- $Y = \sum_{\ell=1}^{N} V_\ell$

Table of Contents

1. Data
   - Overview of the data
   - Models

2. Methodological details
   - Sources of randomness
   - Branching process details
   - Estimating gene copies from qPCR
   - Variance of the estimator

3. Analysis of experimental data
   - Luteinizing hormone
   - S. vulgaris

Sources of randomness

- N_0, the initial number of gene copies, is random
  - $E[N_0] = m_a$
  - $var(N_0) = \sigma_a^2$
- At each cycle of the reaction, each gene copy replicates randomly
  - $N_{n+1} = N_n + \text{Bin}(N_n, p)$

qPCR as a branching process

Note:

$$\begin{aligned} E[N_n] &= E[E(N_n|N_{n-1})] = E[(1+p)N_{n-1}] \\ &= \cdots = (1+p)^n \times E(N_0) \end{aligned}$$

- So $W_n = \frac{N_n}{(1+p)^n}$ is a positive martingale
- Thus, $W_n \to W$ almost surely for some W
- $E[W] = m_a$

qPCR as a branching process

Consider an idealized reaction experiment:

- If we knew p and could let the number of reaction cycles $n \to \infty$:
    - $W_i = \lim_{n\to\infty} \frac{N_{i,n}}{(1+p)^n}$
    - $W_1, W_2, \ldots, W_r \overset{\text{iid}}{\sim} W$
    - So $\frac{1}{r}\Sigma_{i=1}^{r} W_i \overset{\text{a.s.}}{\to} E[W] = m_a$
- But p is unknown and we can only observe ≈ 15 reaction cycles, so we need some other estimator.

Estimating p

- Since p is unknown, we estimate it with $\hat{p}$ via weighted least squares:

$$\begin{pmatrix} N_n \\ N_{n-1} \\ \vdots \\ N_1 \end{pmatrix} - \begin{pmatrix} N_{n-1} \\ N_{n-2} \\ \vdots \\ N_0 \end{pmatrix} = p \times \begin{pmatrix} N_{n-1} \\ N_{n-2} \\ \vdots \\ N_0 \end{pmatrix} + \epsilon$$

- Where $\epsilon_j \overset{\text{approx.}}{\sim} \text{Normal}(0, p(1-p)N_{j-1})$
- With weights $W_j = (N_{j-1})^{-1}$ the resulting estimator is:

$$\hat{p} = \frac{\Sigma_{i=1}^{n}(N_i - N_{i-1})}{\Sigma_{i=1}^{n} N_i}$$

Making the most of a finite sample

Reminder: our idealized estimator was $W(n) = \frac{N_n}{(1+p)^n}$

- W uses only the final observation (N_n)
- More efficient: use the sum $Y_n = \Sigma_{i=1}^{n} N_i$
- By the Toeplitz Lemma, $\frac{Y_n}{(1+p)^n} \overset{\text{a.s.}}{\rightarrow} \frac{1+p}{p} W \Rightarrow \frac{pY_n}{(1+p)^{n+1}} \overset{\text{a.s.}}{\rightarrow} W$
- Plug in $\hat{p}$ and the limit still holds.

Strategy for quantitation

- Collect data on r independent reactions
- For reaction i ($i = 1, 2, \ldots, r$), compute the statistic $M_i = \frac{\hat{p}_i Y_{n_i}}{(1+\hat{p}_i)^{n_i+1}}$
- Average $M_1, M_2, \ldots, M_r$ to get $\bar{M}$
- $\sqrt{r}(\bar{M} - m_a) \xrightarrow{d} \text{Normal}(0, \sigma_L^2)$
- Where $\sigma_L^2 = \sigma_a^2 + m_a E[\frac{1-p}{1+p}]$

Variance of the estimator

$$
\begin{aligned}
\sigma_L^2 = \text{var}[\frac{N_n}{(1+p)^n}] &= E(\text{var}[\frac{N_n}{(1+p)^n}|p]) + \text{var}(E[\frac{N_n}{(1+p)^n}|p]) \\
&= E(\text{var}[\frac{N_n}{(1+p)^n}|p]) + \text{var}(m_a) \\
&= E(\text{var}[\frac{N_n}{(1+p)^n}|p])
\end{aligned}
$$

Variance of the estimator

$$\begin{aligned}
\mathrm{var}[\frac{N_n}{(1+p)^n}|p] &= \frac{1}{(1+p)^{2n}}\mathrm{var}[N_n|p] \\
&= \frac{1}{(1+p)^{2n}}(E(\mathrm{var}[N_n|N_{n-1},p]|p) + \mathrm{var}(E[N_n|N_{n-1},p]|p)) \\
&= \frac{1}{(1+p)^{2n}}(E[N_{n-1}p(1-p)|p] + \mathrm{var}((1+p)N_{n-1}|p)) \\
&= \frac{1}{(1+p)^{2n}}(m_a(1+p)^{n-1}p(1-p) + (1+p)^2\mathrm{var}[N_{n-1}|p]) \\
&= \frac{m_a p(1-p)}{(1+p)^{n+1}} + \frac{\mathrm{var}[N_{n-1}|p]}{(1+p)^{2n-2}} \\
&= \dots \\
&= \frac{m_a p(1-p)}{(1+p)^{n+1}} + \frac{m_a p(1-p)}{(1+p)^{n}} + \dots + \frac{m_a p(1-p)}{(1+p)^{2}} \\
&+ \frac{\mathrm{var}[N_0|p]}{(1+p)^{2n-2n}}
\end{aligned}$$

Variance of the estimator

$$\begin{aligned}\mathrm{var}[\frac{N_n}{(1+p)^n}|p] &= \frac{m_a p(1-p)}{(1+p)^2}\Sigma_{k=0}^{n-1}\frac{1}{(1+p)^k}+\sigma_a^2\\ &\to m_a\frac{1-p}{1+p}+\sigma_a^2\end{aligned}$$

So:

$$\begin{aligned}\mathrm{var}[\frac{N_j}{(1+p)^n}] &= E(\mathrm{var}[\frac{N_n}{(1+p)^n}|p])\\ &\to E(m_a\frac{1-p}{1+p}+\sigma_a^2)\\ &= m_a E(\frac{1-p}{1+p})+\sigma_a^2=\sigma_L^2\end{aligned}$$

Table of Contents

1. Data
    - Overview of the data
    - Models
2. Methodological details
    - Sources of randomness
    - Branching process details
    - Estimating gene copies from qPCR
    - Variance of the estimator
3. Analysis of experimental data
    - Luteinizing hormone
    - S. vulgaris

Experimental data - luteinizing hormone

- The goal with the experimental data was *relative* quantitation
  - Estimate ratio of gene expression between conditions C and T
- The sample was divided into two parts
- One part was diluted to one-third the original concentration
- Sixteen reactions were run under each condition (diluted, normal)

Experimental data - luteinizing hormone

Experimental data - Strongylus vulgaris

- Again, the goal of the was relative quantitation
- One part diluted to one-tenth the original concentration
- Ten reactions run under each condition (diluted, normal)

Experimental data - Strongylus vulgaris