NRES779 Lab 03

Faith Machuca and Griffin Shelor

February 16, 2024

I Objective

Lab Objectives:

- Think about how the data arise.
- Develop a mathematical model of the process that produces the data
- Choose the appropriate likelihood function to tie the predictions in your process model to the data.
- Use maximum likelihood to learn about the parameters in your process model and associated uncertainties

II Introduction

Write a model for the data

$$y \sim Normal(\mu_i, \sigma^2)$$
$$\mu_i = \frac{\alpha(L_i - c)}{\frac{\alpha}{\gamma} + (L_i - c)}$$

where:

- μ_i = the prediction of growth increment of the i^th hemlock tree (cm/year);
- α = maximum growth rate (cm/year);
- γ = slope of curve at low light (cm/year);
- c = light index where growth = 0 (unitless); and
- L_i = measured index of light availability for the i^th hemlock tree, i.e. the proportion of the hemisphere above canopy open to light x 100 (unitless).

III Getting Started using Excel

See attached Excel document.

IV Setting up the Spreadsheet

Question 2

We can look at our observed data to help find good initial conditions for our model parameters. For example, if α is equivalent to the maximum growth rate, we can look at our maximum value of observed growth rate in column 2, which is approximately 46. Setting up the equation for predicted growth rate and graphing is also helpful, because you can see how changing your initial parameters adjusts the predicted growth rate to better match the observed data, playing around with γ and c to get the best looking fit.

Question 3

We adjusted our values to get the best predictions and came up with:

- $\alpha = 46$
- $\gamma = 1.25$
- $c = 3$

We can get a get a better initial value for σ by again looking at our observed data and implementing the following equation into Excel:

$$= STDEV(B2 : B78)$$

which yields a value of approximately 11.5.

Question 4

The full equation that is implemented in the formula in column E is:

$$[z|\mu, \sigma^2] = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(z-\mu)^2}{2\sigma^2}}$$

Question 5

The 'FALSE' argument calculates the probability density function (PDF) of the normal distribution, which provides the probability density (height of the probability curve) at a specific value of the observed growth rate.

Question 6

The 'TRUE' argument calculates the cumulative distribution function (CDF), which provides the probability that a random variable drawn from the distribution will be less than or equal to the value of the observed growth rate.

Question 7

We would use the product of the likelihoods instead of the sum.

Question 8

Some computational problems that can arise is underflow (when the product of many small numbers can become much smaller than what the computer can represent) and overflow (when the number exceeds than what the computer can represent). We also have issues with numerical instability and a loss of precision.

Question 9

It is violating the assumption of linearity because the reality is that the relationship between light availability and growth rate may not be linear at very low and very high levels, i.e., the growth rate could increase to a certain point and plateau, which we see by looking at our scatterplot. There also could be other covariates that are influencing the growth rate, such as water availability or influence from wildlife. This could potentially be addressed by performing a logarithmic transformation on the light availability variable to capture this nonlinear effect. Additionally, this model assumes that light availability is being measured perfectly, which is not likely to be the case. To fix this problem, we can implement a probability distribution to the y_i and μ_i parameters.

V Using the Excel Solver

When we use Excel Solver, we get the following parameters:

- $\alpha =$ 38.5
- $\gamma =$ 1.73
- $c =$ 4.72
- $\sigma =$ 7.26

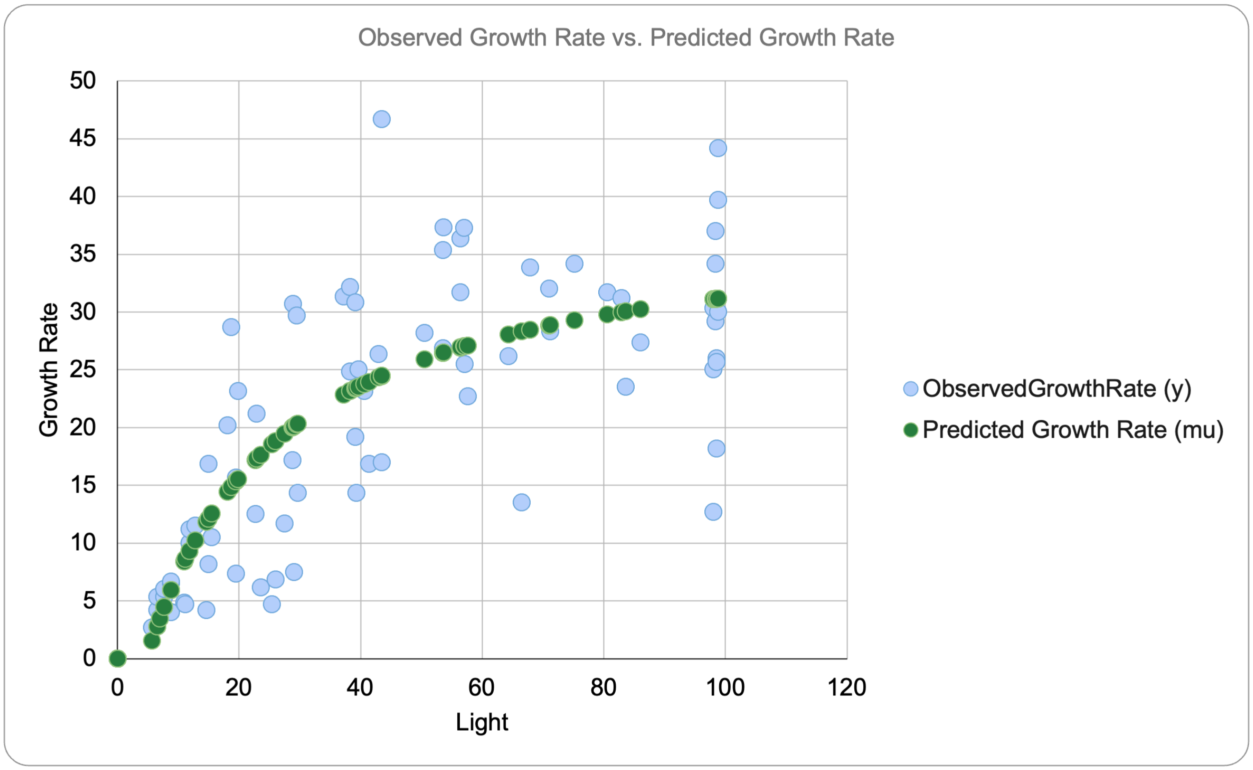

Figure 1: Observed growth rate versus predicted growth rate.

Question 10

You can take the square root from the average of the column D values, which gives us a nearly identical sigma (7.261996414) compared to the Excel Solver output (7.261984676).

VI Using R to do the same thing

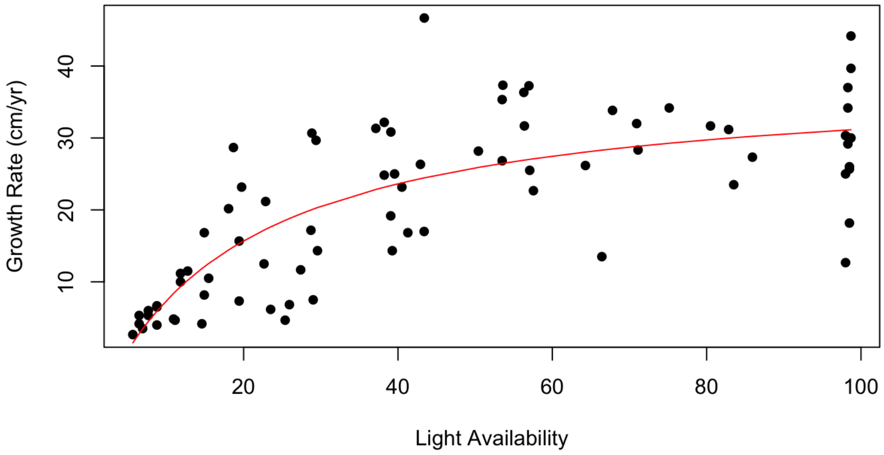

Figure 2: A recreation of the plot from the lab instructions

alpha	gamma	c
38.5	1.732	4.723

Table 1: Values Produced by nls model in R

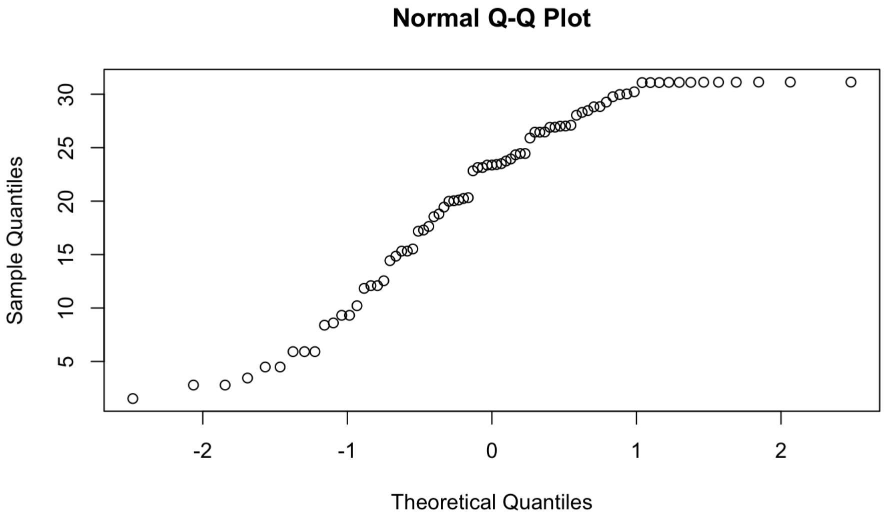

Figure 3: QQnorm plot using the predicted values from the nls model

The qqnorm plot for the predicted values do not indicate that the model residuals are normally distributed. There is a long tail at the right end of the plot and the line is not very straight, so we should not assume that the residuals of the predicted data from the model are normally distributed.

VII Incorporating Prior Information in an MLE

Question 11

$$y_i \sim Normal(\mu_i, \sigma^2)$$
$$\mu_i = \frac{\alpha(L_i - c)}{\frac{\alpha}{\gamma} + (L_i - c)}$$
$$\alpha \sim Normal(35, 4.25^2)$$

The mathematics for incorporating prior information would be:

$$[x|y] \propto \prod_{i=1}^{n} \left(\frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(y_i - \mu_i)^2}{2\sigma^2}}\right) \frac{1}{\sqrt{2\pi}4.25} e^{-\frac{(\alpha-35)^2}{2*4.25^2}}$$

Question 12

You combine likelihoods to obtain the total likelihood by taking the product of all individual likelihoods or the sum of all individual log likelihoods:

$$L(\theta|y) = \prod_{i=1}^{n} L[y_i|\theta]$$
$$log(L(\theta|y) = \sum_{i=1}^{n} log[y_i|\theta]$$

With priors we would add the additional information after the product of likelihoods or sum of log likelihoods (multiply by the prior with the normal distribution, or add it respectively).

Question 13

The α value decreases when using the Excel Solver with the sum of log likelihood that includes prior information. The prior information gave us a mean of 35, but the original model based solely on the data gave us a value of approximately 38. Therefore, including prior information of α would influence that parameter a bit more towards the prior mean.

Question 14

Increasing the standard deviation would make the prior weaker and would make the estimation of α based more off of the data, whereas shrinking the standard deviation would make the prior stronger, making the prior estimation of α have more influence from the data.

Question 15

When we have a lot of data (observations), we can drown out the prior since there is only one prior distribution being added to many individual log likelihoods to create the total likelihood. If we do not have a lot of data, the prior distribution would instead highly impact our outputs if it's a strong prior (i.e., a small standard deviation) because the data would not have enough observations to counteract the influence of the prior distribution.

VIII R code Appendix

```
library(pacman)
p_load(here, tidyverse, gtExtras, RColorBrewer)
hemlock <- read_csv(here(”Labs”, ”Lab3”, ”HemlockData.csv”))
x <- hemlock$Light
y <- hemlock$ObservedGrowthRate
qqnorm(y)
plot(x,y,ylab=”Growth␣Rate␣(cm/yr)”, xlab = (”Light␣Availability”), pch=16)
```

```
set.seed(802)
model <- nls(y ~ a*(x-c)/((a/s)+x-c), trace = TRUE,start=c(a=46,s=1.25,c=3))
summary(model)
p <- coef(model)
a.hat <- p[1]
s.hat <- p[2]
c.hat <- p[3]
yhat <- predict(model)
lines(x,yhat,col="red")
qqnorm(yhat)
```