NRES 779: Bayesian Hierarchical Modeling in Natural Resources

Lab 03: Likelihood

Due: In classs 02/23

I Objective

To get the most out of this lab, you must think of the specific problems you confront here as examples from an infinite variety of analytical challenges that can be attacked using the same general approach:

- 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 of your process model to the data.
- Use maximum likelihood (in this exercise) or Bayesian methods (later) to learn about the parameters in your process model and associated uncertainties.
- In this exercise we will not attempt to distinguish among different sources of uncertainty, i.e., process variance, observation error, and random effects. These distinctions will be made soon enough, after we have developed a bit more statistical sophistication. Moreover, we are leaving the problem of model selection until later in the course.

II Problem

Coates and Burton (1999) studied the influence of light availability on growth increment of saplings of species of conifers in northwestern interior cedar-hemlock forests of British Columbia. They used the deterministic model,

$$\mu_i = \frac{\alpha(L_i - c)}{\frac{\alpha}{\gamma} + (L_i - c)},$$

where:

μ_i = 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)

L_i = measured index of light availability for the ith hemlock tree, i.e. the proportion of the hemisphere above canopy open to light × 100 (unitless).

We will return to this model several times during the course.

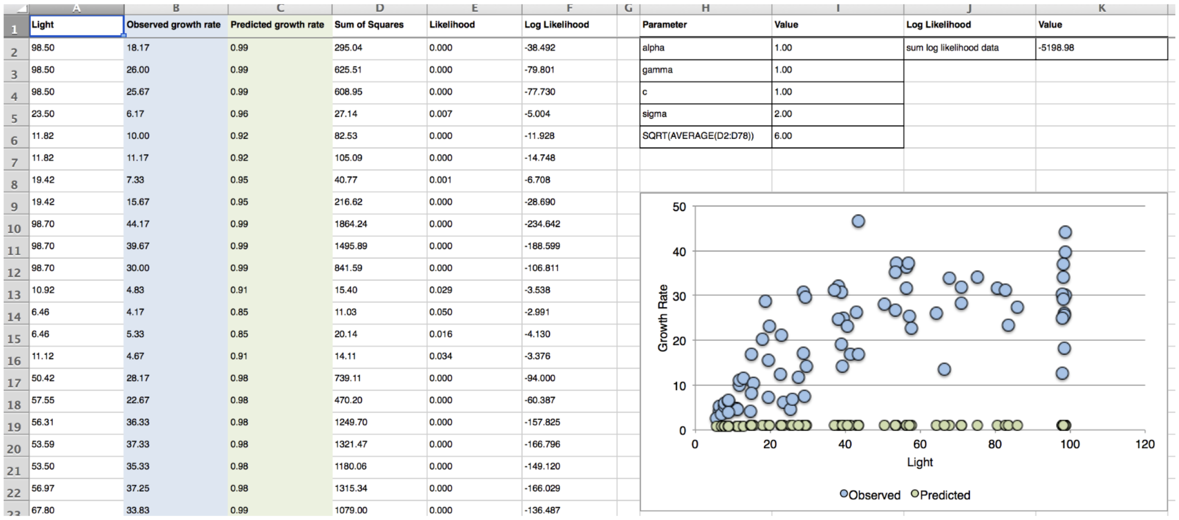
Assume that growth increment can be any real number. It can negative because moose can eat the tops of saplings.

1. Write a model for the data.

III Getting Started using Excel

Obtain maximum likelihood estimates (MLE's) of the model parameters using Solver in Excel (Excel? wtf!?). There is a good reason for using Excel here. When you write code in R, it is easy to fail to understand exactly what is happening "under the hood." The structure of a maximum likelihood analysis is much more transparent when you are forced to build a spreadsheet. You may be delighted to know that this is the last time you will do this in this course.

Open the spreadsheet containing the light limitation data (`HemlockData.csv`). In the next section you will add the proper formulas to columns and cells on this sheet to demonstrate that you know how likelihood works. Your spreadsheet will look something like this before answering questions 1 – 9:



	A	B	C	D	E	F
1	Light	Observed growth rate	Predicted growth rate	Sum of Squares	Likelihood	Log Likelihood
2	98.50	18.17	0.99	295.04	0.000	-38.492
3	98.50	26.00	0.99	625.51	0.000	-79.801
4	98.50	25.67	0.99	608.95	0.000	-77.730
5	23.50	6.17	0.96	27.14	0.007	-5.004
6	11.82	10.00	0.92	82.53	0.000	-11.928
7	11.82	11.17	0.92	105.09	0.000	-14.748
8	19.42	7.33	0.95	40.77	0.001	-6.708
9	19.42	15.67	0.95	216.62	0.000	-28.690
10	98.70	44.17	0.99	1864.24	0.000	-234.642
11	98.70	39.67	0.99	1495.89	0.000	-188.599
12	98.70	30.00	0.99	841.59	0.000	-106.811
13	10.92	4.83	0.91	15.40	0.029	-3.538
14	6.46	4.17	0.85	11.03	0.050	-2.991
15	6.46	5.33	0.85	20.14	0.016	-4.130
16	11.12	4.67	0.91	14.11	0.034	-3.376
17	50.42	28.17	0.98	739.11	0.000	-94.000
18	57.55	22.67	0.98	470.20	0.000	-60.387
19	56.31	36.33	0.98	1249.70	0.000	-157.825
20	53.59	37.33	0.98	1321.47	0.000	-166.796
21	53.50	35.33	0.98	1180.06	0.000	-149.120
22	56.97	37.25	0.98	1315.34	0.000	-166.029
23	67.80	33.83	0.99	1079.00	0.000	-136.487

H	I	J	K
Parameter	Value	Log Likelihood	Value
alpha	1.00	sum log likelihood data	-5198.98
gamma	1.00		
c	1.00		
sigma	2.00		
SQRT(AVERAGE(D2:D78))	6.00		

IV Setting up the Spreadsheet

Let's think about the columns and the rows. This is the benefit of this exercise, so please linger on this, discussing the layout of this spreadsheet with your partner.

- Columns A and B should be easy, these are the data.
- Column C contains the prediction of your model for each level of light. These predictions depend on the values for α, γ, and c contained in column I.
- Column D contain the squared difference between observations and predictions.

- Column E contains the the probability density of the data conditional on μ_i, and σ, one value for each data point. The Excel formula for this is `=NORMDIST(B2,C2,I$5,FALSE)`.
- Cells I2 – I4 contain the values for α, γ, and c that are used to form the predicted growth rate as a function of light level (column C). Cell I5 contains the value for σ. Right now these cells are set to either 1 or 2. You will replace these with better initial values before using the solver to find the maximum likelihood estimates (MLE) for each of these parameters.
- Make a scatterplot of the data and the model predictions, where the x-axis is light level and the y-axis is growth increment. Plot the observed growth increments in blue and the predicted growth increments in green. For the moment, the predicted growth increments should form a line of points along the x-axis.

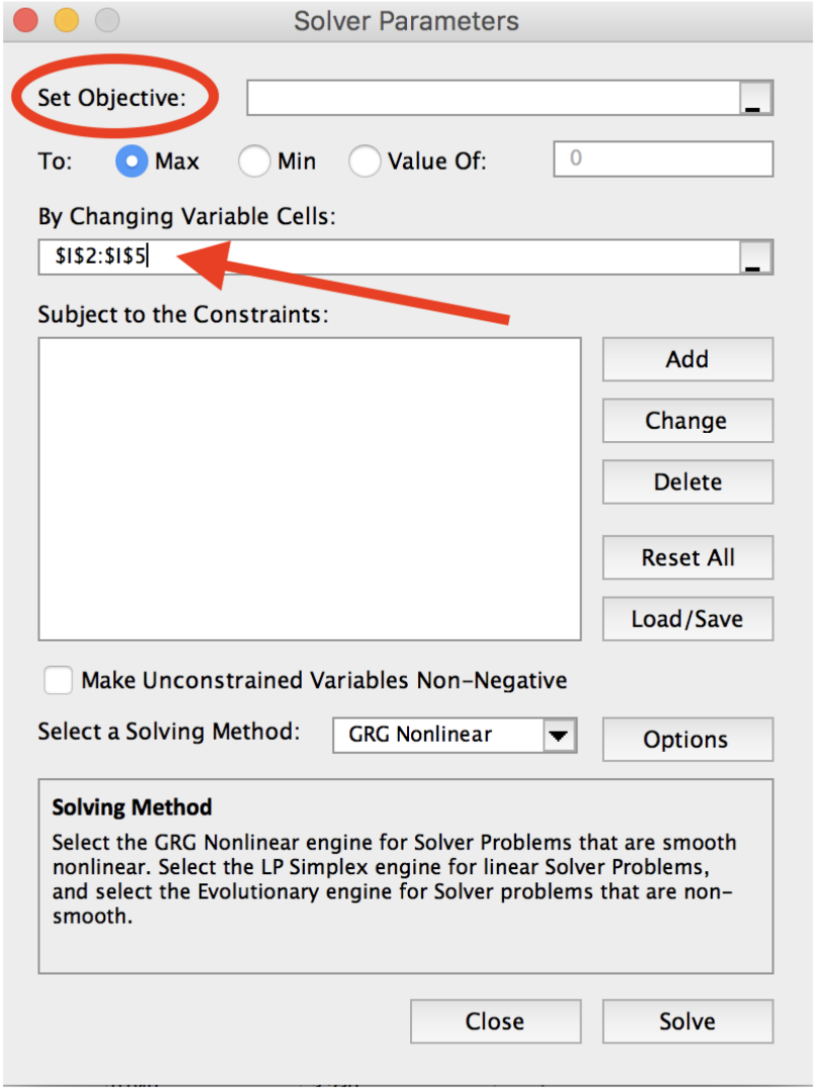
Answer the following questions before proceeding.

2. How could you use the data to help you find good initial conditions for model parameters?
3. Adjust the values for α, γ, and c until you get predictions that look reasonable in your plot. How could we get a better initial value for σ?
4. Write the mathematics (the full equation) that is implemented in the formula in column E.
5. What is the reason for the argument "FALSE" in the Excel formula in column E?
6. What does the function return when that argument is "TRUE"?
7. In column F we take the logs of the likelihoods, which are summed in cell K2. If we had not taken the logs and instead, worked directly with the likelihoods, what formula would we use in K2?
8. What are some potential computational problems with using the individual likelihoods rather than the log likelihoods to estimate the total likelihood?
9. This model violates a fundamental assumption of traditional regression analysis. What is that assumption? How might you fix the problem? (Hint: think about what we are assuming about the covariate, light availability.)

V Using the Excel Solver

When you have your spreadsheet constructed, use Solver to find the maximum likelihood estimates of the parameters. Solver is a very sophisticated non-linear numerical optimizer that uses Newton's method to find values of parameters of a function that maximize or minimize the output of the function.

If you have never used Solver, the main dialog box looks something like this:



Set Objective:

To: ● Max ○ Min ○ Value Of: 0

By Changing Variable Cells:

$I$2:$I$5

Subject to the Constraints:

□ Make Unconstrained Variables Non-Negative

Select a Solving Method: GRG Nonlinear

Solving Method

Select the GRG Nonlinear engine for Solver Problems that are smooth nonlinear. Select the LP Simplex engine for linear Solver Problems, and select the Evolutionary engine for Solver problems that are non-smooth.

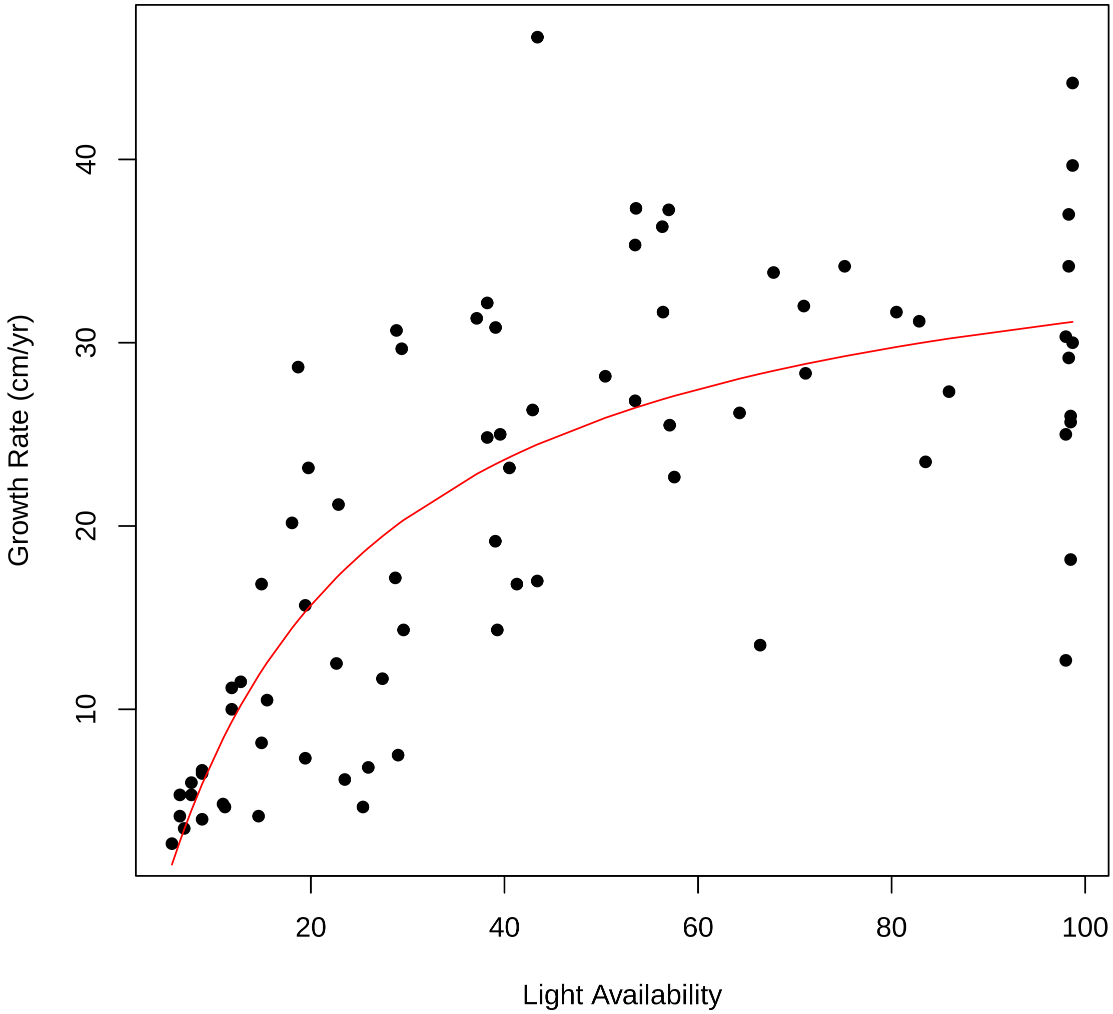
Put the cell containing the sum of the log likelihoods in the Set Objective field. The cells containing the parameter values will go into the By Changing Variable Cells field. Most likely, you will be able to do the exercises without putting constraints on the parameter values if you give them reasonable starting values. However, constraining parameters to reasonable values (e.g., α must be positive and can't be too large) will prevent numerical errors and speed execution time.

10. How might you use the squared error column D to compute σ? Make this computation and compare with your maximum likelihood estimate of σ obtained using Solver.

VI Using R to do the Same Thing

Check your results using the `nls` function in R, which does non-linear estimation for normally distributed data. Examine the assumption that the model residuals are normally distributed using `qqnorm`. To speed things along, I have given you the syntax, but for it to be useful to you, you must study it and experiment a bit. In particular, you must do a help on `nls` and look at its methods—summary, predict, coef, and residuals. The hemlock light and growth increment data are included on WebCampus.

```
d=read.csv("HemlockData.csv")
x=d$Light
y=d$ObservedGrowthRate

plot(x,y,
     ylab="Growth Rate (cm/yr)",
     xlab = ("Light Availability"),
     pch=16)


model=nls(y~a*(x-c)/(a/s+x-c), trace = TRUE,
          start=c(a=50,s=2,c=8))

## 6.3e+03 (7.38e-01): par = (50 2 8)
## 4.1e+03 (8.03e-02): par = (39 1.7 5.5)
## 4.1e+03 (1.98e-03): par = (39 1.7 4.7)
## 4.1e+03 (3.10e-04): par = (38 1.7 4.7)
## 4.1e+03 (7.81e-05): par = (39 1.7 4.7)
## 4.1e+03 (1.97e-05): par = (38 1.7 4.7)
## 4.1e+03 (4.94e-06): par = (39 1.7 4.7)

summary(model)

##
## Formula: y ~ a * (x - c)/(a/s + x - c)
##
## Parameters:
##   Estimate Std. Error t value Pr(>|t|)
## a   38.500      4.370    8.81  3.8e-13 ***
## s    1.732      0.497    3.49  0.00083 ***
## c    4.722      2.069    2.28  0.02533 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.4 on 74 degrees of freedom
##
## Number of iterations to convergence: 6
## Achieved convergence tolerance: 4.94e-06

p=(coef(model))
a.hat=p[1]
s.hat=p[2]
c.hat=p[3]
yhat=predict(model)

lines(x,yhat,col="red")
```



VII Incorporating Prior Information in an MLE

Suppose that a previous study reported a mean value of $\alpha = 35$ with a standard deviation of the mean = 4.25.You may use a normal distribution to represent the prior information on α.

11. Write a model for the data that includes the prior information.

    Incorporate these prior data in your new MLE estimate of α. Hint: create a likelihood function for the probability of the new value of α conditional on the previous value and its standard deviation.

12. How do you combine likelihoods (or log likelihoods) to obtain a total likelihood?

13. Describe what happens to the estimate of α relative to the one you obtained earlier. What is going on?

14. What is the effect of increasing the prior standard deviation on the new estimate? What happens when it shrinks?

15. There is a single log likelihood for the prior distribution but the sum of many for the data. This seems “unfair.” Explain how the prior distribtion can overwhelm the data and vice versa.

References

Coates, K. D., and P. J. Burton. 1999. Growth of planted tree seedlings in response to ambient light levels in northwestern interior cedar-hemlock forests of British Columbia. Canadian Journal of Forest Research **29**:1374–1382.