Homework 8 - Maximum Likelihood

Due November 14 at 9:00am

Names Lindsay Platt & Ben Iuliano

Submit HW as a link to a GitHub repository containing your code and a pdf of this worksheet

Background
We're going to implement the three ways of fitting a linear regression model that were mentioned in lecture: (1) the analytical solution using the normal equation, (2) a numerical optimization approach based on minimizing the sum of squared errors, and (3) a numerical optimization approach based on maximizing the likelihood.

Q1: Simulate some data for a linear regression using the simple linear regression model: y_i = beta_0 + beta_1*x_i + error
Where error~N(0,sigma).

Make sigma large enough that it resembles "typical" ecological data when you plot it, but not so large that it completely obscures the relationship between x and y.

Record your true beta_0, beta_1, and sigma here:

Sigma: 12
Beta_0: 4
Beta_1: 1.3

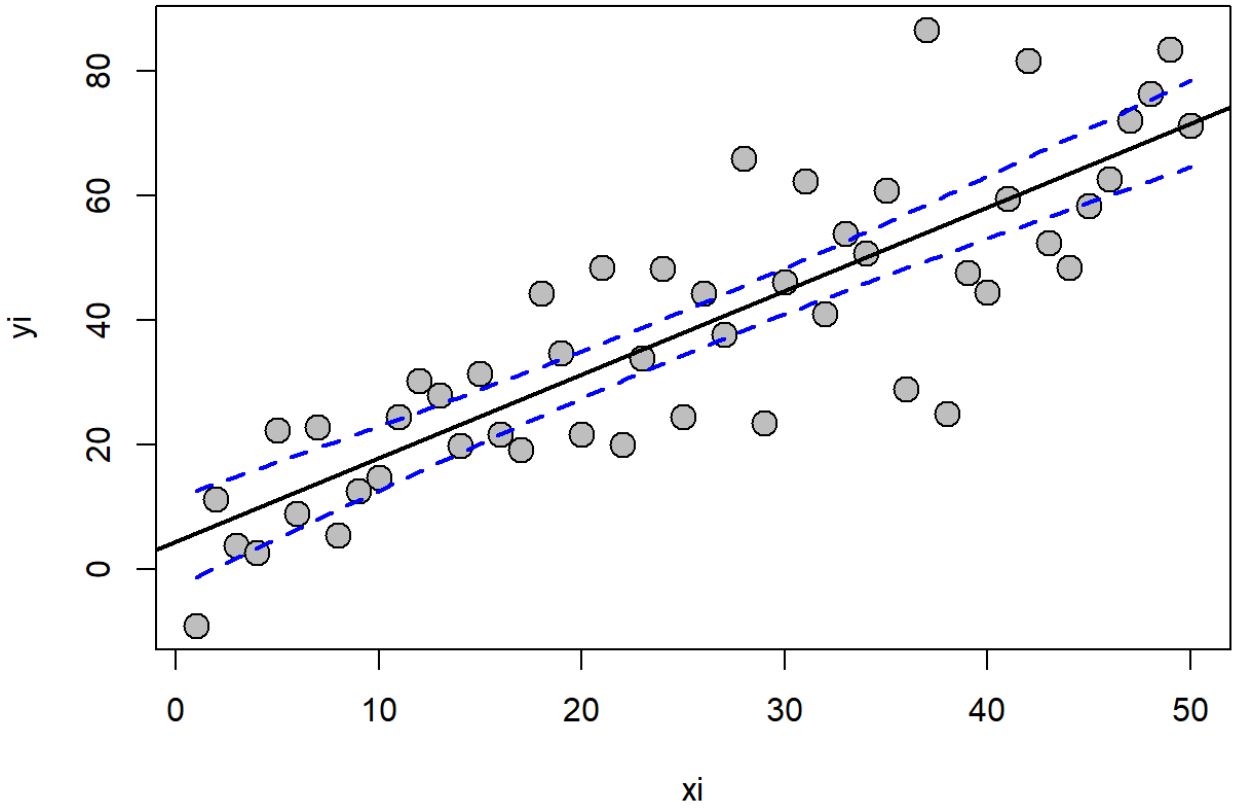
Fit a linear regression model using lm() and do your typical checks of model assumptions that we went over in class.

Paste your estimated model coefficients and their 95% confidence intervals below:

Intercept: 4.38, 95% confidence interval: -2.78 to 11.53
Slope: 1.34, 95% confidence interval: 1.10 to 1.59

Are the coefficient estimates close to the true values? Do the 95% confidence intervals cover the true values?

The coefficients are pretty close to the true values and the 95% confidence intervals cover the true values for the coefficients, but do not include all of the data (see below plot)



Relevant functions: rnorm()

Q2: Analyze the data generated in Q1 using the normal equation:

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

Paste your estimated model coefficients below.

Intercept: 4.38
Slope: 1.34

Are the coefficient estimates close to the true values?

Yes, they are pretty close!

Bonus Q1: Can you find an analytical solution for the se and 95% CI for the model coefficients? (note: this is not in the lectures or reading, you'll have to do some

searching). Write the equations and resulting answers below. Do the 95% confidence intervals cover the true values?

Did not have time to do the bonus questions.

Relevant functions: solve(), t()

Q3: Analyze the data generated in Q1 using a grid search to ***minimize the sum of squared errors*** (no need to iterate more than twice):

Paste your estimated model coefficients below.

Intercept: 4.2
Slope: 1.35

Q4: Analyze the data generated in Q1 using a grid search to ***minimize the negative log likelihood*** (no need to iterate more than twice). Note, there is a third parameter that you will need to estimate here: sigma

Paste your estimated model coefficients below.

Intercept: 4
Slope: 1.35
Sigma: 12

Q5: Analyze the data generated in Q1 ***using optim()*** to minimize the negative log likelihood. Note, there is a third parameter that you will need to estimate here: sigma

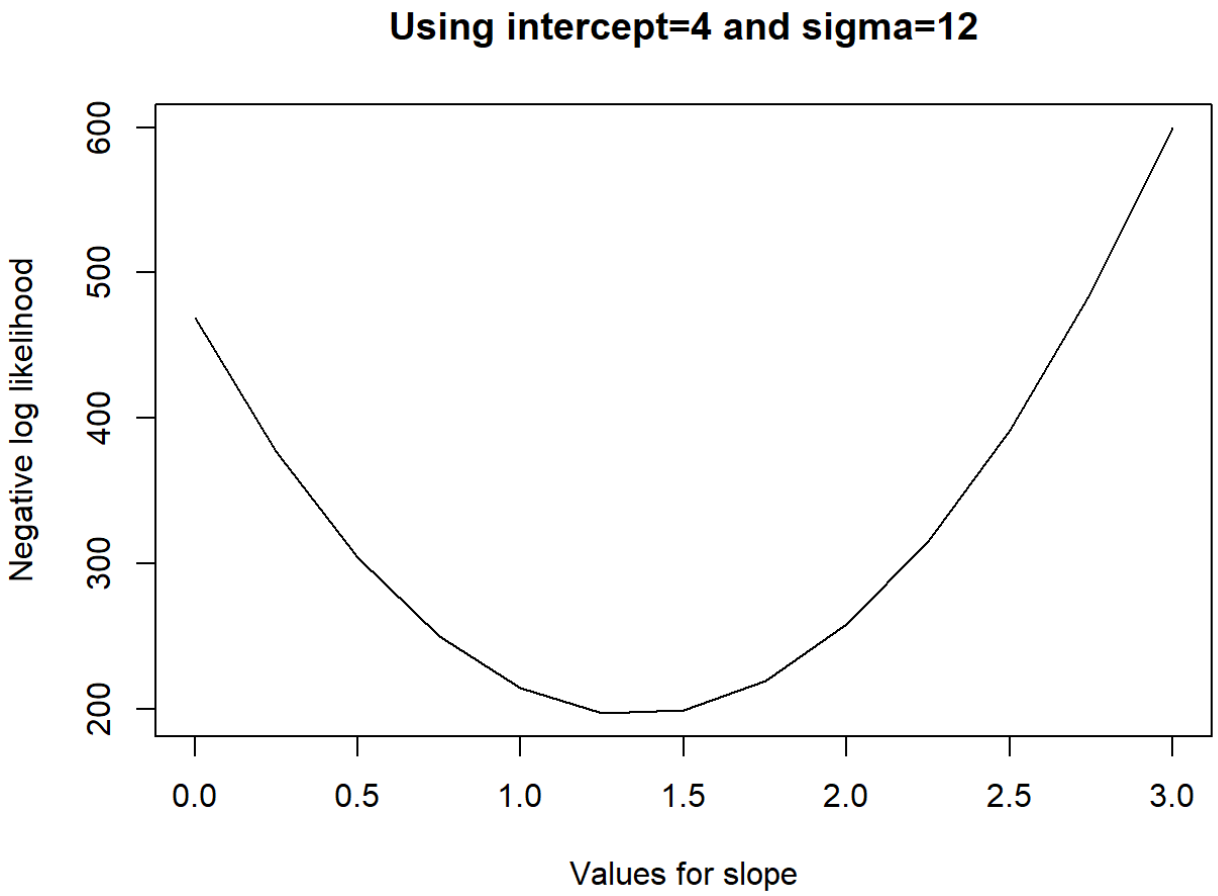
Paste your estimated model coefficients below.

Intercept: 4.38
Slope: 1.34
Sigma: 12.14

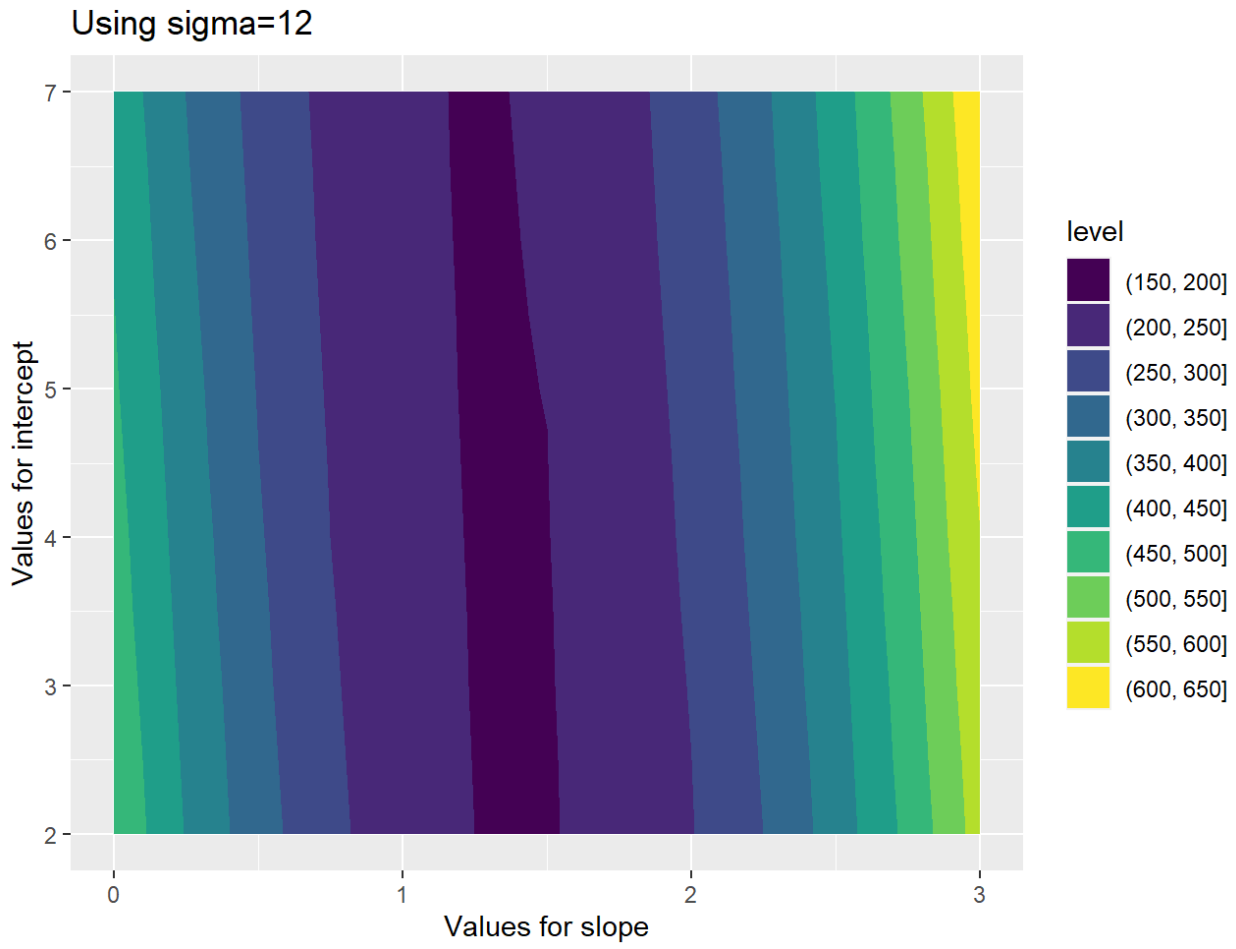
Did the numerical optimization algorithm converge? How do you know? Yes, it converged. We know because the value under the `convergence` item in the output of # `optim()` was 0, which indicates a successful completion.

Did the numerical optimization algorithm find a global solution? How do you know? Yes because all of the values are above our initial parameter inputs and the optimization converged, so it successfully tried parameters above those initializations, too.

Q6: Plot a likelihood profile for the slope parameter while estimating the conditional MLEs of the intercept and sigma for each plotted value of the slope parameter (see p. 173 of Hilborn and Mangel).



Q7: Plot the joint likelihood surface for the intercept and slope parameters. Is there evidence of confounding between these two parameters (i.e., a ridge rather than a mountain top)?



Yes, there is evidence of confounding because we have a ridge instead of a mountaintop (between slopes of 1-2, there is a minimum log likelihood for intercepts from 2-7).

Q8: How different are the estimated coefficients from Q1, Q2, and Q5 and how do they compare to the true values?

The estimates from Q1 and Q2 were closer to the true values than the estimates for Q3, but all are very close (Q4 was actually the closest to the true values).

question	intercept	slope
Q1	4.375962	1.342488
Q2	4.375962	1.342488
Q5	4.382545	1.342130

Bonus Q2: Calculate the standard errors of the intercept, slope, and sigma using the Hessian matrix. Standard errors are the square roots of the diagonal of the inverse Hessian matrix. How do these standard errors compare to those from lm() in Q1?

Did not have time to do the bonus questions.

Bonus Q3: How does the computational speed compare between using lm(), the normal equation, and optim() to estimate the coefficients? Note, you can get the computation time placing the following code around your regression code:

Did not have time to do the bonus questions.