PROBLEM SET # 4

EC/ACM/CS 112: Bayesian Statistics

Due date	Tuesday, February 2, 10:30 am
Submission instructions	>> Create a pdf of the R Notebook with your solutions (details below) >> Submit in Canvas
Additional files included in the problem set package	>> dataset for problem set >> solutions template (.Rmd file)

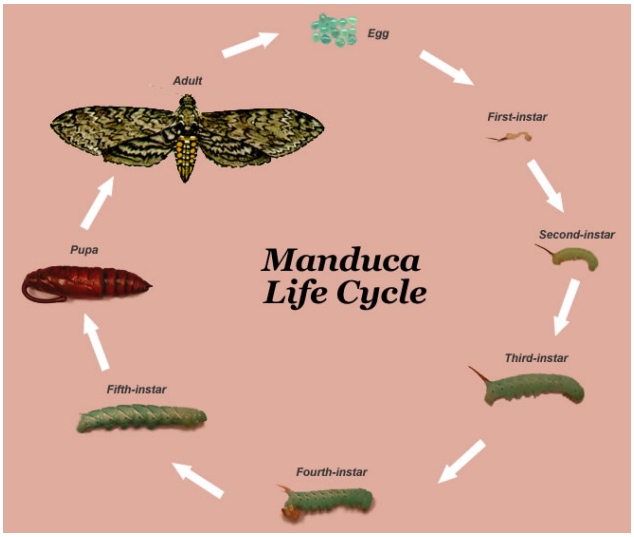
QUESTION 1. USING THE LINEAR REGRESSION MODEL TO UNDERSTAND THE METABOLIC RATE OF MANDUCA SECTA CATERPILLARS

LEARNING GOAL

- A number of variables have been shown to affect the metabolic rate of organisms, including age, weight, and activity level.
- In this problem you will use the linear regression model to investigate the role of body size and life cycle stage on the metabolic rate of manducta sexta caterpillars

BACKGROUND

- Manducta sexta caterpillars (also known as tobacco hornworms) to go through the following life-cycle:



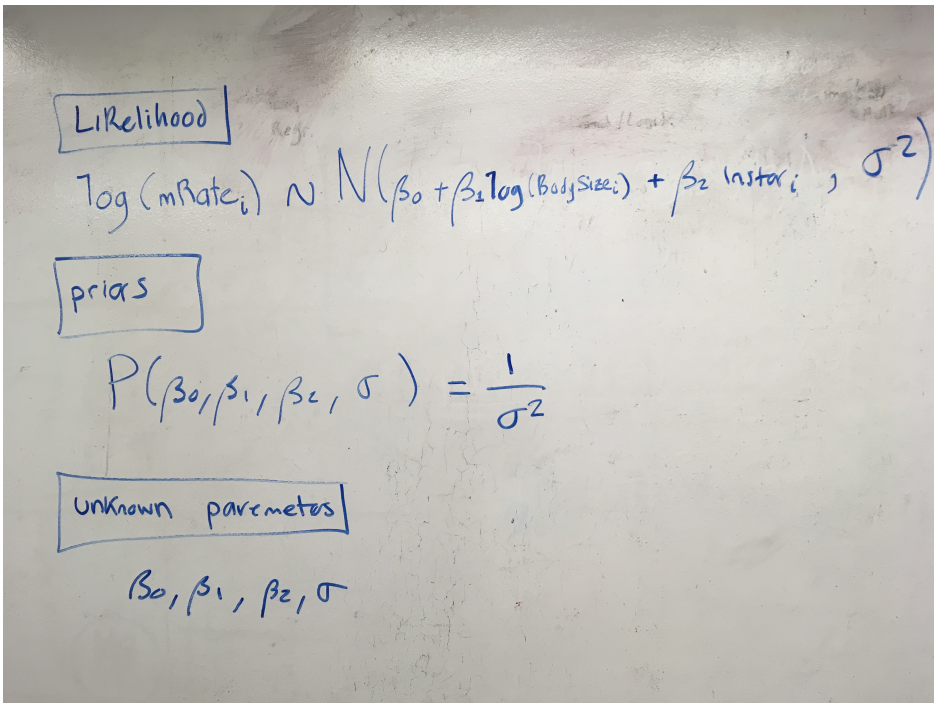
DATASET. The dataset "MetabolicRate.csv" included with the problem set package was originally compiled by Prof. Itagaki and his students http://www.kenyon.edu/directories/campus-directory/biography/harry-itagaki/).

It contains measures on the following three variables 305 caterpillars:

- BodySize = size of the caterpillar (in grams)
- Instar = number from 1 (smallest) to 5 (largest) indicating the caterpillar's life stage
- mRate = a measurement of the caterpillar's metabolic rate

The dataset has no missing observations.

STATISTICAL MODEL. In this problem we are interested in the following regression model:



STEPS

Step 1. Use a scatter plot matrix to visualize the relationship between the variables mRate, instar and BodySize.

- What do you learn from the scatter plot?
- Does it raise any potential concerns to keep in mind during the analysis and interpretation of the linear regression model?

TIPS:

- The command **pairs()** from the base R package is useful here.
- Link to related information: http://www.sthda.com/english/wiki/scatter-plot-matrices-r-base-graphs

Step 2. Compute the joint posteriors of the unknown parameters for the following regression model (i.e., compute $P(\beta_0, \beta_1, \beta_2, \sigma|data)$) using the grid method.

With four parameters, the choice of the posterior grid becomes important as it is computationally prohibitive to use fine wide grids over all dimensions.

Fortunately there is a simple heuristic for addressing this.

Start with a wide but coarsely defined grid. In particular, compute the posterior using the following grid:

- For β_0 use {-5,-4.75, …., 4.75,5 }
- For β_1 use {-2,-1.9, …., 1.9, 2}
- For β_2 use {-2,-1.9, …., 1.9, 2}
- For σ use {0.1,0.2, …, 3}

Plot the marginal posteriors for this grid (i.e., $P(\beta_0|data)$, $P(\beta_1|data)$, $P(\beta_2|data)$, and $P(\sigma|data)$) to check that the specified grid covers the region where the posterior of each parameter has positive mass.

Then repeat the computation of the posterior using a narrower grid range for each parameter, but with approximately the same number of points. The key idea is to narrow the grid to points that are likely to contain the full probability mass of the posterior. Thus, for example, next you might want to try the following grid:

- For β_0 use {2,2.05, ….,3.95 ,4 }
- For β_1 use {0.5, 0.51, …, 0.99, 1}
- For β_2 use {-0.25,-0.24, …., 0.24, 0.25}
- For σ use {0.1, 0.2, …, 1}

After you complete this second step, you can again check that the specified grid covers the region where the posterior of each parameter has positive mass. A good way of checking that you are not excluding any key regions of the grid is to verify that the marginal posteriors have regions with zero probability on the left and right regions of the grid. If you fail this test, go back and adjust the grid ranges to fix this.

Keep iterating like this until you end up with a grid region that gives you a degree of coverage and resolution that you find satisfactory.

Please describe your final grid region in the solution set.

IMPORTANT: The decision of when to stop this search is subjective. These types of choices are a key part of statistical modeling and you need to be get used to making them. Also, you will get a full score for the problem provided that your posteriors look approximately correct.

Step 3. Do the following two visualizations of your posteriors:

- Plot the marginal posteriors $P(\beta_1|data)$ and $P(\beta_2|data)$). These plots should include a vertical line indicating the mean of each marginal posterior.
- Use a heat map to visualize the joint posterior (i.e., for $P(\beta_1, \beta_2|data)$).

Step 4. Why does $P(\beta_1, \beta_2|data)$) has this shape?

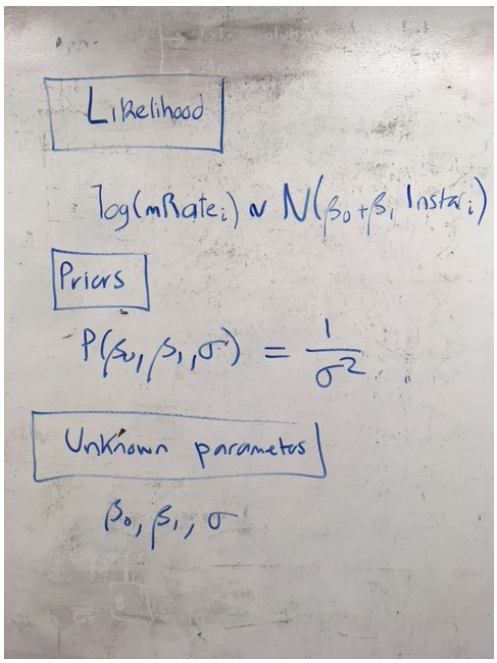
Step 5. Compute the probability that both β_1 and β_2 are greater than zero.

QUESTION 2. COMPARISON OF THE UNIVARIATE & BIVARIATE MODELS OF THE METABOLIC RATE OF MANDUCA SECTA CATERPILLARS

LEARNING GOAL

- Deepen your understanding of the effect of omitting correlated predictors in linear regression models

STATISTICAL MODEL. In this question we are interested on the following univariate version of the previous model, which looks for a linear relationship between Instar and log(mRate), ignoring the impact of body size:



STEPS

Step 1. Given your findings in question 1, would you expect that this model would lead to a posterior distribution of beta 1 that is about the same as in the bivariate model, shifted upwards (i.e., larger), or shifted downwards (i.e., smaller)? Why?

Please answer this question qualitatively and WITHOUT/BEFORE estimating the new model.

Step 2. Compute the joint posteriors of the unknown parameters for the univariate regression model (i.e., compute $P(\beta_0, \beta_1, \sigma|data)$) using the grid method. As before, choose your grid

parameters iteratively to end up with a level of parameter coverage and resolution that you find adequate.

Step 3. What are the resulting mean marginal posteriors for β_{instar} under the univariate and bivariate model? Why are they so different?

Step 4. In order to understand the data better, please make the following scatter plot:

- Plot log(bodySize) vs log(mRate), with log(mRate) in the y-axis.
- Use a different color for caterpillars in each different stage.

What does this plot suggest about the role of the instar and body weight variables on metabolic rate?