PROBLEM SET # 3

EC/ACM/CS 112: Bayesian Statistics
Caltech

Due date	Tuesday, January 26, 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)

ABOUT THE PROBLEM SET

Dataset. In this problem set you will analyze the data set "data_task_duration_difficulty.csv" that is included in the problem set package.

It contains self-report data on the duration and difficulty of the problem set 2 for this course that was collected from students who took the course in Spring 2018.

The dataset contains two variables:

- duration = reported number of hours spent doing problem set 1
- difficulty = reported difficulty of problem set 1 (scale: 1 = fairly easy to 5 = fairly hard)

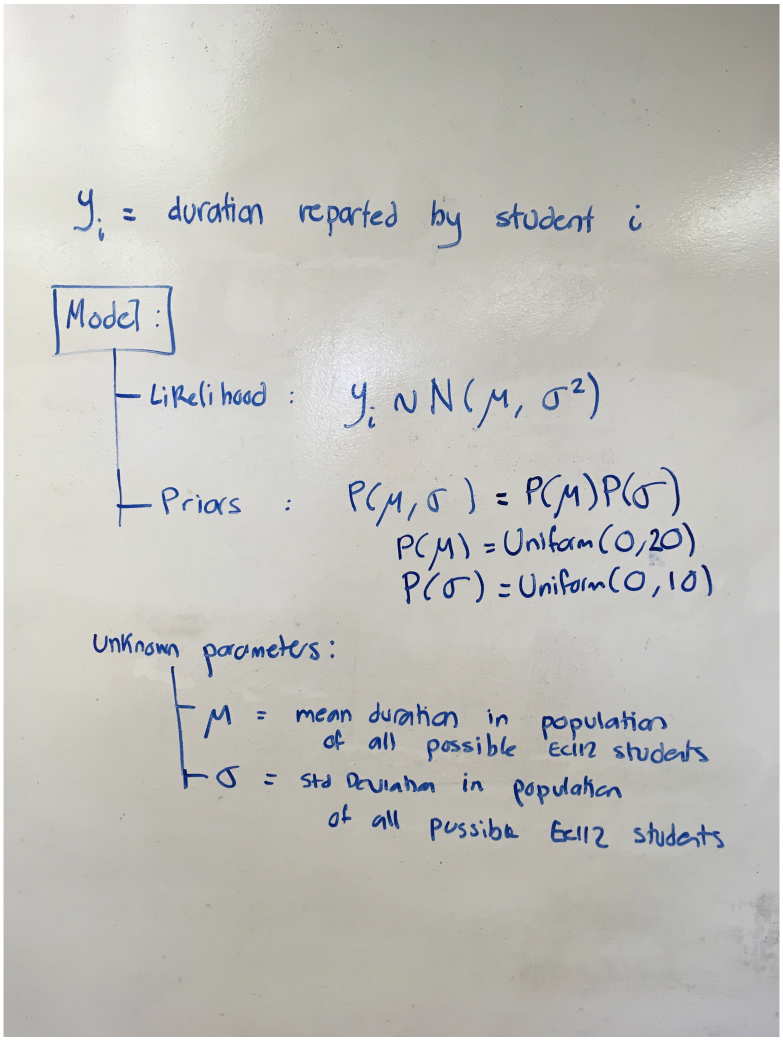
The dataset contains 65 observations. Each row of observations contains the reported duration and difficulty of a given student.

Learning goals.

- Apply the basic measurement and basic linear regression models to a real dataset
- Practice fitting simple linear regression models using the grid method.
- Gain an appreciation of how these canonical statistical models provide useful tools for understanding many real datasets.

QUESTION 1. APPLYING THE BASIC MEASUREMENT MODEL TO THE DURATION VARIABLE

The basic measurement model can be applied to the duration data as follows:



In this problem you are asked to fit this model using the grid method and to report various aspects of the results.

Step 1: Compute the joint posterior for μ and σ using the grid method (i.e., $P(\mu, \sigma|data)$)

- For μ use the grid {0, 0.1, 0.2, …., 19.9, 20}
- For σ use the grid {0.05, 0.1, …., 9.95, 10}

Step 2: Compute the marginal posteriors for μ and σ (i.e., $P(\mu|data)$ and $P(\sigma|data)$)

Step 3: Use the results of steps 1 and 2 to compute the following summary statistics of the posterior function, and report them in your solution document.

- Mean of marginal posterior for μ
- Standard deviation of marginal posterior for μ
- Mean of marginal posterior for σ
- Standard deviation of marginal posterior for σ
- Covariance of μ and σ

Step 4. Plot a heat map to visualize the joint posterior and copy it to your solution document

Step 5. Plot the marginal posterior distributions for μ and σ in two different plots

Step 6. Professor Rangel's best guess when he created the problem set was that the average problem set duration would be under 5 hours. Given the data, what is the probability that his hypothesis was correct (i.e., compute $P(\mu < 5|data)$).

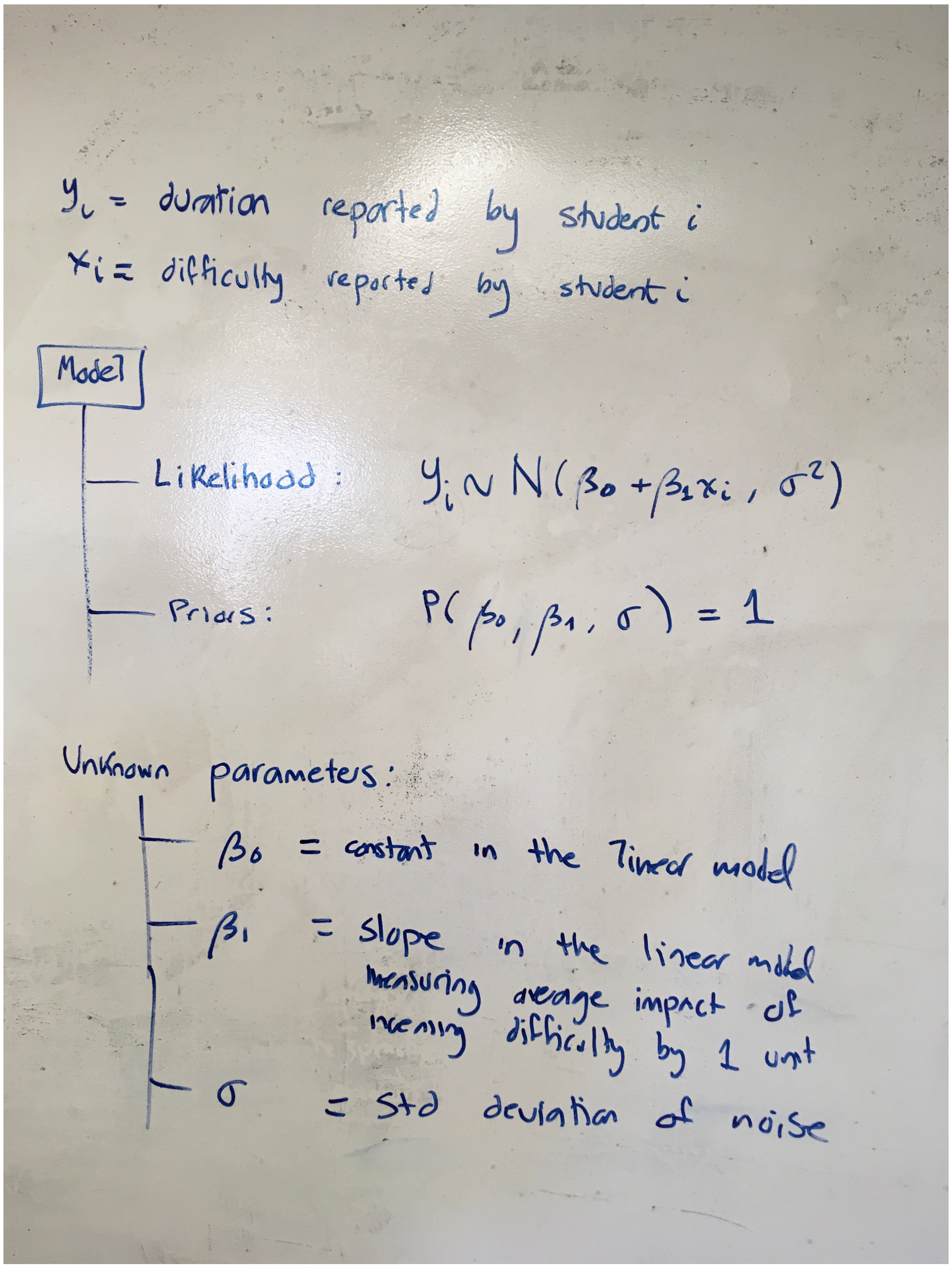
Tips:

- Double check your work by making sure that the results in steps 3, 4, and 5 are consistent
- You might find it useful to look at the code for lectures in Unit 3

QUESTION 2. APPLYING THE BASIC LINEAR REGRESSION MODEL TO THE DATASET

In this question you are asked to apply the basic linear regression model to investigate if there is a linear relationship between the reported problem set difficulty and the amount of time that it took students to complete it.

Here is the model that you should work with:



Step 1: Compute the joint posterior for $\beta_0,\ \beta_1$ and σ using the grid method (i.e., $P(\beta_0, \beta_1, \sigma|data)$)

- For β_0 use the grid {-10,-9.9, , …., 9.9, 10}
- For β_1 use the grid {-10,-9.9, , …., 9.9, 10}
- For σ use the grid {0.05, 0.1, …., 4.95, 5}

Step 2: Compute the marginal posteriors for $\beta_0,\ \beta_1$ and σ (i.e., $P(\beta_0|data)$, $P(\beta_1|data)$ and $P(\sigma|data)$)

Step 3: Use the results of steps 1 and 2 to compute the following summary statistics of the posterior function, and report them in your solution document.

- Mean of marginal posterior for β_0
- Standard deviation of marginal posterior for β_0
- Mean of marginal posterior for β_1
- Standard deviation of marginal posterior for β_1
- Mean of marginal posterior for σ
- Standard deviation of marginal posterior for σ
- Covariance of β_0 and β_1

Step 4. Plot the marginal posterior distributions for $\beta_0,\ \beta_1$ and σ in three different plots

Step 5. Plot three different heat maps to visualize the joint posterior (i.e., for $P(\beta_0, \beta_1|data)$, $P(\beta_0, \sigma|data)$ and $P(\beta_1, \sigma|data)$)

Step 6. In order to understand better the fit of the model make a plot similar to the one in p. 19 of the slides for lecture 4.1. Note that:

- You should add a "saliently displayed" line with the regression line given by the $\hat{\beta}_0,\ \hat{\beta}_1$ estimates.
- You should plot 1000 other regression lines determined by randomly sampling form the joint posterior $P(\beta_0, \beta_1|data)$.
- These last set of lines should be semi-transparent to facilitate the interpretability of the plot.

Tips:

- You need to clean the database to eliminate observations with missing data. The command **is.na()** is useful here.
- Double check your work by making sure that the results in steps 3, 4, and 5 are consistent
- You might find it useful to look at the code for lectures 4.1-4.5.

QUESTION 3. TESTING THE NORMALITY ASSUMPTIONS IN THE MODEL

In this question you are asked to test of the normality assumptions in the models used in the previous two questions.

Step 1. Look at the basic measurement model

- Plot a histogram of the duration variable and overlay an estimated density line on it (you can use the R functions **lines(density(variableName), …)** to accomplish this.
- Standardize the duration variable
- Make a q-q plot of the standardized duration variable (tip: look at the **qqnorm()** function)
- Are these plots consistent with the normality assumption of the model?

Step 2. Look at the linear regression model.

- Compute the vector of residuals associated with the $\hat{\beta}_0$, $\hat{\beta}_1$ estimates. Each residual is given by $y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$
- Plot a histogram of the residuals and overlay an estimated density line on it
- Standardize the residuals
- Make a q-q plot of the standardized residuals.
- Are these plots consistent with the normality assumption of the model?

Tips:

- The command **lm()** can be used to quickly fit linear regressions using non-Bayesian methods.
- Since $\hat{\beta}_0$, $\hat{\beta}_1$ are the estimates generated by these methods, you can use the **lm()** command to compute them for this part of the problem set.
- You might find it useful to look at the code for lectures 4.2-4.4.