STA365: Homework 1

Daniel Simpson

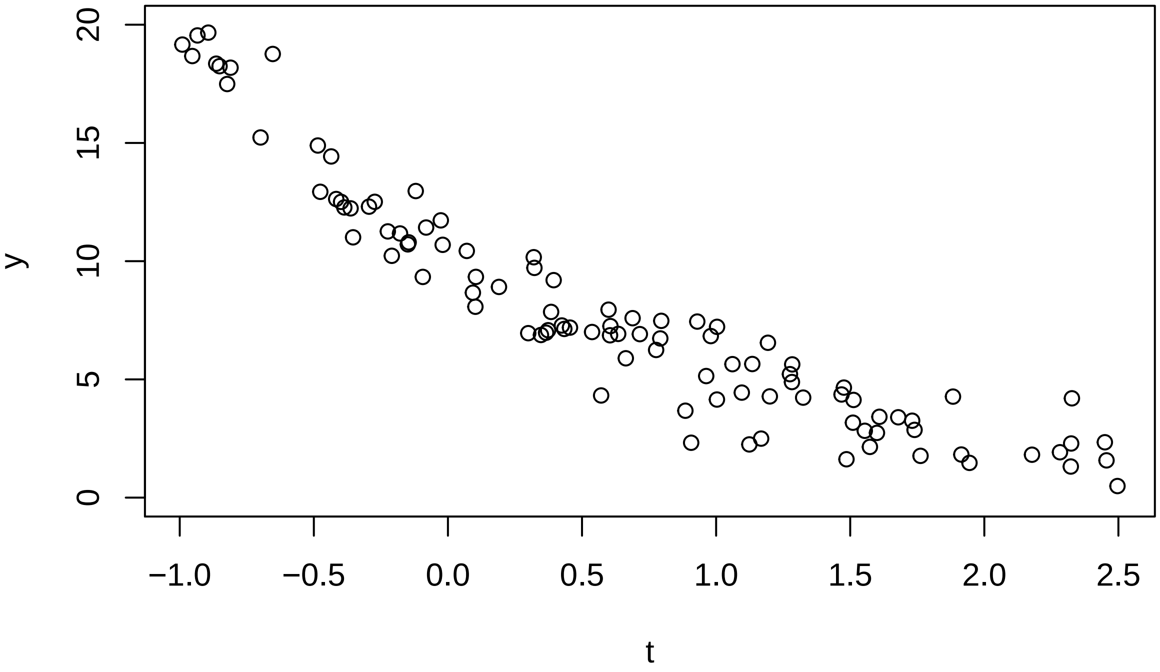
27/01/2021

Instructions

This should be prepared using RMarkdown and **all R and Stan code should be included in the document**. A large portion of the marks are for the explanations accompanying the questions, and if these are missing you will get very few marks.

The assignment should be submitted via Quercus as a single pdf file. It is due at 12pm (Midday) on 5 February, 2020.

Transformations of data



Not all regression is linear. Sometimes, we are faced with a regression problem where the relationship between the data and covariate is not linear.

For this assignment, we are going to consider a very simple data set, which contains an obervation y_i and a single covariate t_i. We assume that the relationship between the mean $\mu_i = \mathbb{E}(y_i)$ and t_i is exponential, that is

$$\mu_i = \exp(\alpha + \beta t_i).$$

We are going to consider the following two likelihoods:

1. $y_i \sim N(\exp(\alpha + \beta t_i), \sigma^2)$
2. $\log(y_i) \sim N(\alpha + \beta t_i, \sigma^2)$

Although both of these models have the same parameters names, **they mean different things and should possibly have different priors**.

The data for this problem is available on the quercus and can be read with the command

```r
data <- readRDS("hw1_data.RDS")
```

Part 1: Priors (10%)

Please set and justify weakly informative priors for both models. You should write acompanying text for each model that explains what assumptions are encoded in your priors and why they are weakly informative for the model in question.

Full marks will be given for an answer tht includes useful, well-plotted graphs; appropriate simulations; and a ~~full and detailed example for each model~~ **a full and detailed explanation for the priors in each model**.

You may use the following information about the data when setting your priors:

- While it is not physically impossible to get a negative measurement, it should be fairly unlikely.
- We expect y and t to be negatively correlated.
- It would be very unlikely to see a data value bigger than 50.
- The covariate t_i will alway be in the interval $[-1, 2.5]$.

Part 2: Posteriors (10%)

1. Write Stan programs to fit each model with the priors chosen in the previous part of the quesiton.
2. Write a detailed critique of each model is the superior model for this particular data set. This critique should include interpretations of posterior predictive checks and leave-one-out cross validation to criticize the two models individually.
3. Use leave one out cross validation to choose the more appropriate model and write an explanation of why this model is better for this particular data set.

Note: All comparisions should take place on the natural data scale (aka the scale of y_i rather than $\log(y_i)$). This will require you transform model 2 appropriately in order to obtain the predictive distributions for the leave-one-out cross validation.

Note: All Stan code must be included in the markdown document. Missing Stan code will get zero marks.