# Solutions Chapter 7 - The posterior - the goal of Bayesian inference

Alexandra Hotti

October 2019

## 7.1 Googling

Suppose you are chosen, for your knowledge of Bayesian statistics, to work at Google as a search traffic analyst. Based on historical data you have the data shown in Table 7.1 for the actual word searched, and the starting string (the first three letters typed in a search). It is your job to help make the search engines faster, by reducing the search-space for the machines to lookup each time a person types.

| | Barack Obama | Baby clothes | Bayes |
|---|---|---|---|
| **Bar** | 50 % | 30 % | 30 % |
| **Bab** | 30 % | 60 % | 30 % |
| **Bay** | 20 % | 10 % | 40 % |

Table 1: The columns give the historic breakdown of the search traffic for three topics: Barack Obama, Baby clothes, and Bayes; by the first three letters of the user's search.

### Problem 7.1.1.

Find the minimum-coverage confidence intervals of topics that are at least at 70%. The confidence intervals can be found in Table **??**.

| | Barack Obama | Baby clothes | Bayes | Credibility |
|---|---|---|---|---|
| **Bar** | [50 %] | 30 % | 30 % | 45% |
| **Bab** | 30 % | [60 % | 30 %] | 75 % |
| **Bay** | [20 % | 10 % | 40 %] | 100 % |
| **Coverage** | 70 % | 70 % | 70 % | |

Table 2: Confidence intervals of 70% ≤

**Answer:** We want a coverage equal to or larger than 70%. These intervals can be found in Table 2 Each interval is marked with one color and in every interval the true search word is contained.

### Problem 7.1.2.

Find most narrow credible intervals for topics that are at least at 70%.

**Answer:**

| | Barack Obama | Baby clothes | Bayes | Credibility |
|---|---|---|---|---|
| **Bar** | [50 % | 30 %] | 30 % | 72.7% |
| **Bab** | 30 % | [60 % | 30 %] | 75% |
| **Bay** | 20 % | [10 % | 40 %] | 71.4 % |
| **Coverage** | 50 % | 100 % | 70 % | |

Table 3: Credible intervals of 70% ≤

## 7.2 GDP versus infant mortality

The data in *posterior_gdpInfantMortality.csv* contains the GDP per capita (in real terms) and infant mortality across a large sample of countries in 1998.

### Problem 7.2.1.

A simple model is fit to the data of the form:

$$M_i \sim N(\alpha + \beta GDP_i, \sigma) \tag{1}$$

Fit this model to the data using a Frequentist approach. How well does the model fit the data?

**Answer:** In the plot in Figure 1 it is apparent that the relationship between the mortality and the GDP is not linear. The relation looks more like an inverted exponential function.

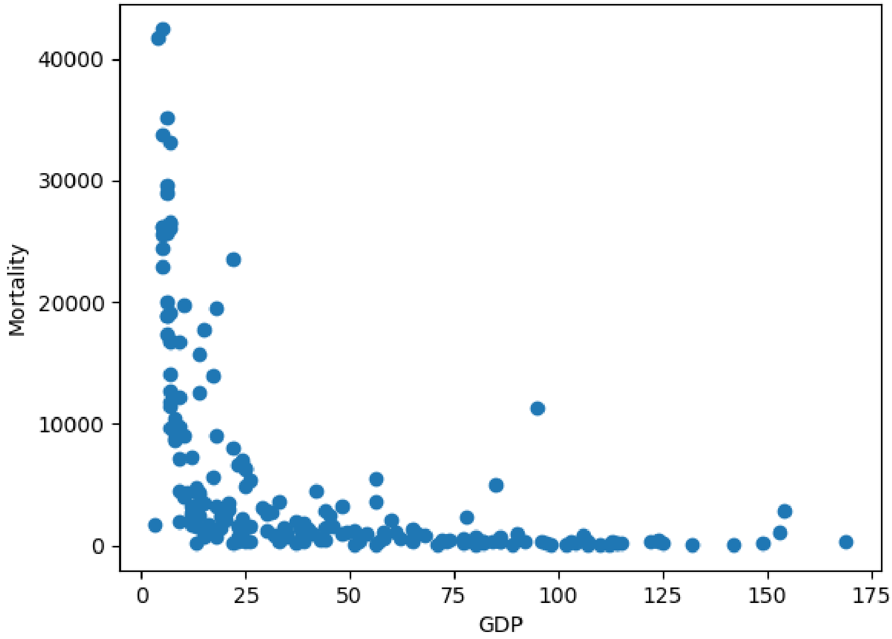

Figure 1: A scatter plot of the GDP per capita and the infant mortality across several countries in 1998.

### Problem 7.2.2.

An alternative model is:

$$log(M_i) \sim N(\alpha + \beta log(GDP_i), \sigma) \tag{2}$$

Fit this model to the data using a Frequentist approach. Which model do you prefer, and why?

**Answer:**

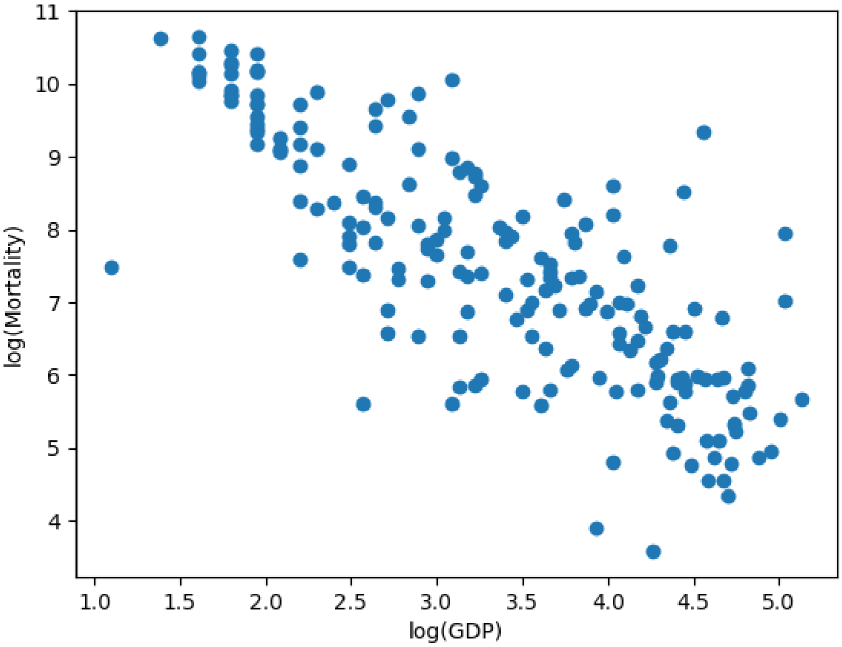

Figure 2: A scatter plot of the log(GDP) per capita and the log(infant mortality) across several countries in 1998.

It is clear that the data in Figure 2 is linear while the data in Figure 1 is not. Therefore, the second model is preferred.

Here in Figure 3 I used least squares to fit the second model. It looks as though we have several far off outliers to the least-squares line. Thus, it would not be optimal to use a Normal Distribution to model the error as this would deem far away outliers very improbable or even impossible. A better option is to use a Student T distributions since this distribution has heavier tails than the normal distribution, thus allowing for larger errors. Also, our populations standard deviation is unkown, but we can still calculate the sample standard deviation.

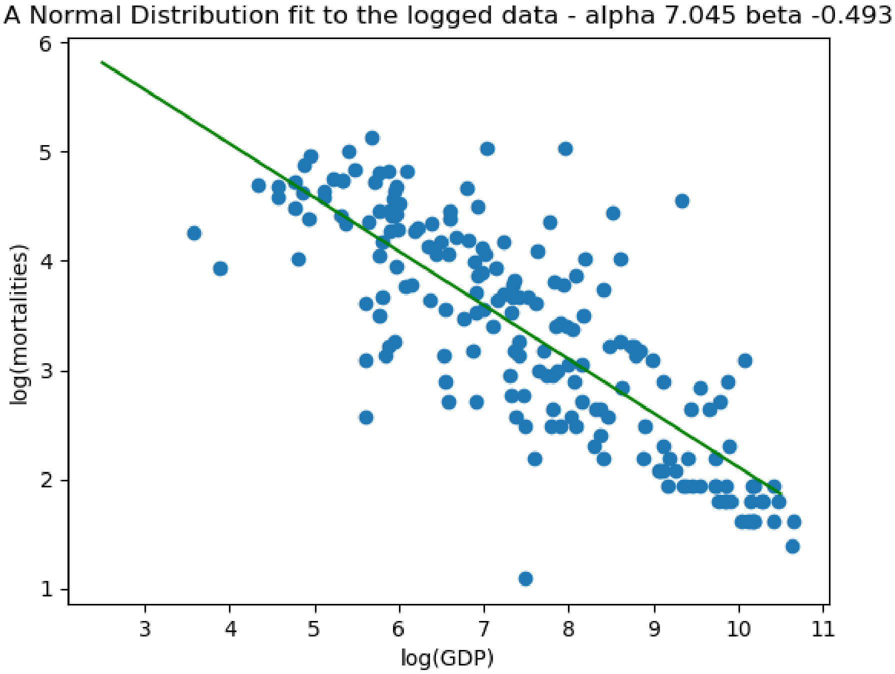

Figure 3: A model fit to the log(GDP) per capita and the log(infant mortality) across several countries in 1998.

## Problem 7.2.3.

Construct 80% confidence intervals for ($\alpha$, $\beta$) for the log-log model.

**Answer:** Now, we calculate a two sided (we are spread in both directions) confidence interval for our parameters using the Student t distribution. Since we want a 80% confidence interval and have 193 samples we get a t value of 1.28.

$$\alpha = 7.0452 \pm \frac{1.28 \cdot s_{\alpha}}{\sqrt{n}} \tag{3}$$

$$\beta = -0.493 \pm \frac{1.28 \cdot s_{\beta}}{\sqrt{n}} \tag{4}$$

Using the standard error for each coefficient from *statsmodels.api* in python, I get the 80% confidence intervals:

$$\alpha = 7.0452 \pm 1.28 \cdot 0.199 = [6.79028, 7.29972] \tag{5}$$

$$\beta = -0.4932 \pm 1.28 \cdot 0.026 = [-0.52648, -0.45992] \tag{6}$$

## Problem 7.2.4.

We have fit the log-log model to the data using MCMC. Samples from the posterior for $(\alpha, \beta, \sigma)$ are contained within the file $posterior\_posteriorsGdpInfantMortality.csv$. Using this data find the 80% credible intervals for all parameters (assuming these intervals to be symmetric about the median). How do these compare with the confidence intervals calculated above for ($\alpha$, $\beta$)? How does the point estimate of $\sigma$ from the Frequentist approach above compare?

**Answer:** First, $\sigma$ for a linear regression model is the standard deviation of the residuals (prediction errors). i.e. the Root Mean Square Error (RMSE). Residuals are a measure of how far from the regression line data points are; RMSE is a measure of how spread out these residuals are. In other words, it tells you how concentrated the data is around the line of best fit.

Calculating the RMSE by comparing the actual and modeled log(mortality) values using $\alpha$ and $\beta$ gives us.

$$RMSE = \hat{\sigma}_F = 0.5907630300187096 \approx 0.59 \tag{7}$$

For comparison with the frequentist $\sigma_F$ the mean of the posterior distribution over $\sigma$ is calculated as:

$$\hat{\sigma}_B = 0.5966279950195218 \approx 0.60 \tag{8}$$

In the MCMC model the credible intervals for $\sigma$, $\beta$ and $\alpha$ are found below. They where calculated using a 80% predictive interval around the highest density region. The package used to achieve this was pymc3.

$$6.79658858 \leq \alpha \leq 7.27955343 \tag{9}$$

$$-0.52550682 \leq \beta \leq -0.46289348 \tag{10}$$

$$0.5537347 \leq \sigma \leq 0.63361329 \tag{11}$$

Overall, the two approached yields very similar results in this particular case.

## Problem 7.2.5.

The following priors were used for the three parameters:

$$\alpha \sim N(0, 10) \tag{12}$$

$$\beta \sim N(0, 10) \tag{13}$$

$$\sigma \sim N(0, 5), \sigma \geq 0. \tag{14}$$

Explain any similarity between the confidence and credible intervals in this case.

**Answer:** The specified priors are very broad and diffuse compared to our parameter ranges and are thus almost ”flat”. Therefore, the posterior and the likelihood will contain roughly the same information since their shapes become similar. Therefore, it is more likely that the credible and confidence intervals numerically will coincide. However, since the intervals are computed in different ways this is not always the case.

## Problem 7.2.6.

How are the estimates of parameters ($\sigma$, $\beta$) correlated? Why?

**Answer:** They have a negative correlation as can be seen in Figure 2.

## Problem 7.2.7.

Generate samples from the prior predictive distribution. How does the min and max of the prior predictive distribution compare with the actual data?

**Answer:**
The prior predictive distribution can in general be written as:

$$p(x) = \int_0^1 p(x, \theta)d\theta = \int_0^1 p(x|\theta)p(\theta)d\theta \tag{15}$$

where the first part of the expression is the likelihood and the second the prior. So, to estimate the prior predictive distribution we would have to sample parameter values from the priors and then condition the likelihood on these values to generate a data point in the histogram for the prior predictive distribution.

The priors are given in the assignment description problem 7.2.5.

Assuming that are samples are iid the log likelihood becomes:
$(https://en.wikipedia.org/wiki/Maximum\_likelihood\_estimation)$

$$log(p(M|\theta)) = \frac{-n}{2}log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_i^n(GDP_i - (\alpha + \beta log(GDP_i)))^2 \tag{16}$$

# References