

# Bayesian Linear Models

Felipe José Bravo Márquez

July 6, 2021

# Bayesian Linear Models

- In this class, which is mostly based on chapter 4 of [McElreath, 2020], we are going to revisit the linear regression model from a Bayesian point of view.
- The idea is the same: to model the relationship of a numerical dependent variable  $\mathbf{y}$  with  $n$  independent variables  $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$  from a dataset  $d$ .
- We also maintain the assumption that each attribute has a linear relationship to the mean of the outcome.

$$\mu_i = \beta_0 + \beta_1 x_i + \dots \beta_n x_n$$

- However, we are not going to use least squares or maximum likelihood to obtain point estimates of the parameters.
- Instead, we are going to estimate the joint posterior distribution of all the parameters of the model:

$$f(\theta|d) = f(\beta_0, \beta_1, \dots, \beta_n, \sigma|d)$$

# Bayesian Linear Models

- Bayesian linear regressions more flexible than least squares as it allows incorporating prior information.
- It also allows to interpret the uncertainty of the model in a clearer way.
- Notice that the parameters of the model are  $\beta_0, \beta_1, \dots, \beta_b$  and  $\sigma$  but not  $\mu_i$ .
- This is because  $\mu_i$  it is determined deterministically from the linear model's coefficients.
- In order to build our posterior we need to define a likelihood function:

$$f(\mathbf{d}|\beta_0, \beta_1, \dots, \beta_n, \sigma) = \prod_{i=1}^m f(d_i|\beta_0, \beta_1, \dots, \beta_n, \sigma)$$

- Where  $d_i$  corresponds to each data point in the dataset containing values for  $y$  and  $x_1, \dots, x_n$ .
- The likelihood of each point is modeled with a Gaussian distribution:

$$f(d_i|\beta_0, \beta_1, \dots, \beta_n, \sigma) = N(\mu_i, \sigma^2)$$

# Bayesian Linear Models

- Now we need a joint prior density:

$$f(\theta) = f(\beta_0, \beta_1, \dots, \beta_n, \sigma)$$

- And the posterior gets specified as follows:

$$f(\theta|d) = \frac{\prod_{i=1}^m f(d_i|\beta_0, \beta_1, \dots, \beta_n, \sigma) * f(\beta_0, \beta_1, \dots, \beta_n, \sigma)}{f(d)}$$

- The evidence is expressed by a multiple integral:

$$f(d) = \int \int \dots \int \prod_{i=1}^m f(d_i|\beta_0, \beta_1, \dots, \beta_n, \sigma) * f(\beta_0, \beta_1, \dots, \beta_n, \sigma) d\beta_0 d\beta_1 \dots d\beta_n d\sigma$$

- In most cases, the priors are specified independently for each parameter, which is equivalent to assuming:

$$f(\beta_0, \beta_1, \dots, \beta_b, \sigma) = f(\beta_0) * f(\beta_1) * \dots * f(\beta_n) * f(\sigma).$$

# A model of height revisited

- To understand this more concretely, we will rebuild the linear model relating the height and weight of the !Kung San people using a Bayesian approach.
- We will refer to each person's height and weight as  $y_i$  and  $x_i$  respectively.
- Our probabilistic model specifying all components of a Bayesian model is defined as follows:

$y_i \sim N(\mu_i, \sigma)$	[likelihood]
$\mu_i = \beta_0 + \beta_1 x_i$	[linear model]
$\beta_0 \sim N(100, 100)$	$[\beta_0 \text{ prior}]$
$\beta_1 \sim N(0, 1)$	$[\beta_1 \text{ prior}]$
$\sigma \sim \text{Uniform}(0, 50)$	$[\sigma \text{ prior}]$

- Parameters  $\beta_0$  and  $\beta_1$  are the intercept and the slope of our linear model.
- The parameter  $\sigma$  is the standard deviation of all the heights.
- Note that we are setting the same  $\sigma$  for all observations, which is equivalent to the Homoscedasticity property of the standard linear regression.

# A model of height revisited

- Our priors were set independently for each parameter which implies that the joint posterior  $f(\beta_0, \beta_1, \sigma)$  can be expressed as  $f(\beta_0) * f(\beta_1) * f(\sigma)$ .
- It should be kept in mind that the choice of priors is subjective and should be evaluated accordingly.
- Let's try to justify our choice a bit:
  - 1 The Gaussian prior for  $\beta_0$  (intercept), centered on 100cm with a standard variation of 100, covers a huge range of plausible mean heights for human populations while giving very little chance for negative heights.
  - 2 The Gaussian prior for  $\beta_1$  (slope), centered on 0 with a standard variation of 1, acts as a **regularizer** to prevent the model from assigning extreme values to  $\beta_1$ .<sup>1</sup>
  - 3 The uniform prior for the standard deviation  $\sigma$  between 0 and 50 forbids obtaining negative standard deviations. The upper bound (50 cm) would imply that 95% of individual heights lie within 100cm of the average height. That's a very large range.

---

<sup>1</sup>Regularization will be discussed later in the course.

# Conclusions

- Blabla



McElreath, R. (2020).

*Statistical rethinking: A Bayesian course with examples in R and Stan.*

CRC press.