

Bayesian Linear Regression

Felipe José Bravo Márquez

July 9, 2021

Bayesian Linear Regression

- 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 .
- The response variable \mathbf{y} is again modeled with a Gaussian distribution:
 $y_i \sim N(\mu_i, \sigma^2)$.
- 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

- The Bayesian linear regression is 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 σ but not μ_i .
- This is because μ_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 (IID assumption).
- 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]
$\beta_0 \sim N(150, 50)$	$[\beta_0 \text{ prior}]$
$\beta_1 \sim N(0, 1)$	$[\beta_1 \text{ prior}]$
$\sigma \sim \text{Uniform}(0, 50)$	$[\sigma \text{ prior}]$

- Parameters β_0 and β_1 are the intercept and the slope of our linear model.
- The parameter σ is the standard deviation of all the heights.
- Note that we are setting the same σ 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 prior density $f(\beta_0, \beta_1, \sigma)$ 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 β_0 (intercept), centered on 150cm with a standard variation of 50, covers a huge range of plausible mean heights for human populations while giving very little chance for negative heights.
 - 2 The Gaussian prior for β_1 (slope), centered on 0 with a standard variation of 1, acts as a **regularizer** to prevent the model from **overfitting** the data by assigning extreme values to β_1 .¹
 - 3 The uniform prior for the standard deviation σ between 0 and 50 prohibits 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.

¹Regularization and overfitting will be discussed later in the course.

Fitting the Model

- Now we need to fit the model to the data to build the posterior distribution.
- Grid approximation is not a valid option, as setting up a grid for 3 parameters would be too computationally expensive.
- We will use Laplace approximation instead.
- In this approach we obtain the MAP estimates for each parameter using a hill-climbing **optimization** method.
- Then we fit a **multivariate Gaussian distribution** centered on these values.
- This distribution is the multidimensional extension to the standard Gaussian.

The multivariate Gaussian distribution

- The multivariate Gaussian distribution in d -dimensions defined by the following density function (PDF):

$$f_{\mathbf{x}} = \frac{1}{(2\pi)^{d/2} |\Sigma|^{1/2}} \exp \left(-\frac{1}{2} (\vec{\mathbf{x}} - \vec{\mu})^T \Sigma^{-1} (\vec{\mathbf{x}} - \mu) \right)$$

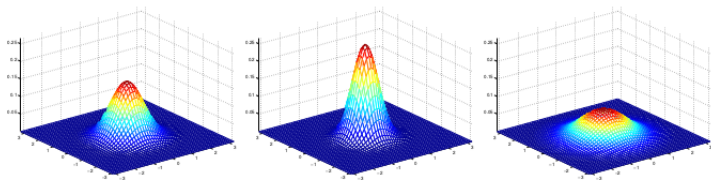
- This density function allows working with a d -dimensional vector of random variables \vec{X} .
- The first parameter of this distributions is a mean vector $\vec{\mu} \in \mathcal{R}^d$ with the mean value of each dimension.
- The second parameter is a covariance matrix $\Sigma \in \mathcal{R}^{d \times d}$,
- This matrix contains the variance of each variable in the diagonal and the covariance of variables X_i and X_j in the other cells $\Sigma_{i,j}$:

$$\text{Cov}(X) = \Sigma$$

- The matrix Σ is symmetric and positive semi-definite.
- The multivariate Gaussian $N(\vec{\mu}, \Sigma)$ is a very convenient distribution for modeling multidimensional random variables.

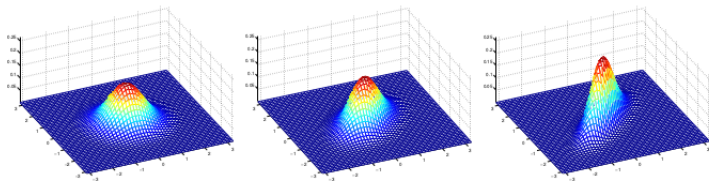
The multivariate Gaussian distribution

- Here are some examples taken from [Ng, 2008] of what the density of a multivariate Gaussian distribution looks like:



- The left-most figure shows a Gaussian with mean zero (that is, the 2×1 zero-vector) and covariance matrix $\Sigma = I$ (the 2×2 identity matrix).
- A Gaussian with zero mean and identity covariance is also called the standard normal distribution.
- The middle figure shows the density of a Gaussian with zero mean and $\Sigma = 0.6I$.
- The rightmost figure shows one with $\Sigma = 2I$.
- We see that as Σ becomes larger, the Gaussian becomes more “spread-out”, and as it becomes smaller, the distribution becomes more “compressed”.

The multivariate Gaussian distribution



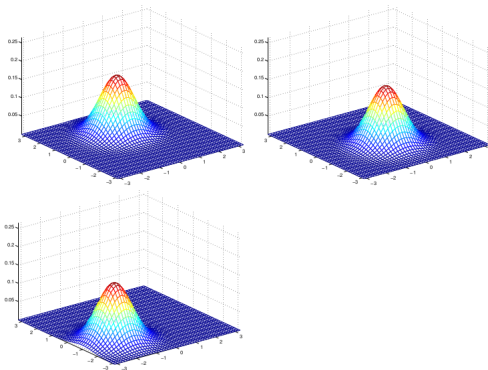
- The figures above show Gaussians with mean 0, and with covariance matrices respectively

$$\Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}; \quad \Sigma = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}; \quad \Sigma = \begin{bmatrix} 1 & 0.8 \\ 0.8 & 1 \end{bmatrix}.$$

- The leftmost figure shows the familiar standard normal distribution, and we see that as we increase the off-diagonal entry in Σ , the density becomes more “compressed” towards the 45° line (given by $x_1 = x_2$).

The multivariate Gaussian distribution

- As our last set of examples, fixing $\Sigma = I$, by varying $\vec{\mu}$ we can also move the mean of the density around.



- The figures above were generated using $\Sigma = I$, and respectively

$$\mu = \begin{bmatrix} 1 \\ 0 \end{bmatrix}; \quad \mu = \begin{bmatrix} -0.5 \\ 0 \end{bmatrix}; \quad \mu = \begin{bmatrix} -1 \\ -1.5 \end{bmatrix}.$$

Laplace approximation

- In Laplace approximation we assume that the joint posterior follows a multivariate Gaussian distribution $f(\theta_1, \dots, \theta_n) = N(\vec{\mu}, \Sigma)$.
- This approximation is convenient for unimodal and roughly symmetric posterior distributions [Gelman et al., 2013].
- The Bayesian asymptotic theory says that if the dataset is large enough, a posterior distribution can be approximated by a Gaussian [Gelman et al., 2013].
- The values of $\vec{\mu}$ are obtained from the posterior mode of each parameters (MAP):

$$\vec{\mu} = \vec{\theta}_{MAP}$$

- These is done using numeric optimization techniques.

Laplace approximation

- The values of Σ are obtained from the curvature near these values, which are obtained from the second derivatives of the posterior:

$$\Sigma = [I(\theta_{MAP})]^{-1}$$

where

$$I(\theta) = -\frac{d^2}{d\theta^2} \log f(\theta|d)$$

- Notice that both $\vec{\mu}$ and Σ can be calculated from the unnormalized posterior $f(d|\theta) * f(\theta)$
- We don't need to calculate the evidence $f(d)$ to perform Laplace approximation [?].

Fitting the Model

- Laplace approximation is implemented in the **quap** function from the **rethinking** package.
- The model for height defined above can be implemented as follows:

```
library(rethinking)
data(Howell1)
d <- Howell1
d2 <- d[ d$age >= 18 , ]

b.reg1 <- quap(
  alist(
    height ~ dnorm( b0 + b1*weight, sigma ),
    b0 ~ dnorm( 150 , 50 ) ,
    b1 ~ dnorm( 0 , 1) ,
    sigma ~ dunif( 0 , 50 )
  ) , data=d2 )
```

Fitting the Model

- We can summarize this posterior with the command **precis**:

```
> precis( b.reg1, prob=0.95 )  
      mean    sd    2.5%   97.5%  
b0    113.99  1.90  110.26  117.71  
b1      0.90  0.04    0.82    0.98  
sigma   5.07  0.19    4.70    5.45
```

- These numbers provide Gaussian approximations for each parameter's marginal posterior distribution.
- This is the distribution of any parameter θ_j regardless of the values of the others.
- In mathematical terms this is obtained, after integrating (averaging) over $\theta_j \forall j \neq i$:

$$f(\theta_i|d) = \int \cdots \int f(\theta_1, \dots, \theta_n|d) d\theta_1, \dots, d\theta_{i-1}, d\theta_{i+1}, \dots, d\theta_n$$

- We can compare the mean and standard deviation of β_0 and β_1 with what we obtained using least squares in previous lecture.
- The values are almost the same!

Conclusions

- Blabla

References I



Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013).

Bayesian data analysis.

CRC press.



McElreath, R. (2020).

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

CRC press.



Ng, A. (2008).

Generative learning algorithms.

Stanford Lecture Notes, 5(4).