

# Linear Regression for a Continuous Variable

## General Principles

To study relationships between a continuous independent variable and a continuous dependent variable (e.g., height and weight), we can use linear regression. Essentially, we draw a line that passes through the point cloud of the two variables being tested. For this, we need to have:

- 1) An intercept  $\alpha$ , which represents the origin of the line—the expected value of the dependent variable (height) when the independent variable (weight) is equal to zero.
- 2) A coefficient  $\beta$ , which informs us about the slope of the line. In other words, it tells us how much Y (height) increases for each increment of the independent variable (weight).
- 3) A variance term  $\sigma$ , which informs us about the spread of points around the line, i.e., the variance around the prediction.

## Considerations

### 🔥 Caution

- Bayesian models consider model parameter uncertainty , allowing for the quantification of confidence or uncertainty through the parameters' posterior distribution . Therefore, we need to declare prior distributions for each model parameter, in this case for:  $\alpha$ ,  $\beta$ , and  $\sigma^2$ .
- Prior distributions are built following these considerations:
  - As the data is scaled (see introduction), we can use a Normal distribution for  $\alpha$  and  $\beta$ , with a mean of 0 and a standard deviation of 1.
  - Since  $\sigma$  is strictly positive, we can use any distribution that is positively defined, such as the Exponential or Uniform distribution.
- Gaussian regression deals directly with continuous outcomes, estimating a linear relationship between predictors and the outcome variable without needing a link

function (see introduction). This simplifies interpretation, as coefficients represent direct changes in the outcome variable.

## Example

Below is an example code snippet demonstrating Bayesian linear regression using the Bayesian Inference (BI) package. Data consist of two continuous variables (height and weight), and the goal is to estimate the effect of weight on height. This example is based on McElreath (2018).

## Python

```
from BI import bi

# Setup device-----
m = bi(platform='cpu')

# Import Data & Data Manipulation -----
# Import
from importlib.resources import files
data_path = files('BI.resources.data') / 'Howell1.csv'
m.data(data_path, sep=';')
m.df = m.df[m.df.age > 18] # Manipulate
m.scale(['weight']) # Scale

# Define model -----
def model(weight, height):
    a = m.dist.normal(178, 20, name = 'a')
    b = m.dist.lognormal(0, 1, name = 'b')
    s = m.dist.uniform(0, 50, name = 's')
    m.normal(a + b * weight , s, obs = height)

# Run mcmc -----
m.fit(model) # Optimize model parameters through MCMC sampling

# Summary -----
m.summary() # Get posterior distributions
```

## R

```
library(BI)
m=importbi(platform='cpu')

# Load csv file
m$data(paste(system.file(package = "BI"), "/data/Howell1.csv", sep = ''), sep=';')

# Filter data frame
m$df = m$df[m$df$age > 18,]

# Scale
m$scale(list('weight'))

# Convert data to JAX arrays
m$data_to_model(list('weight', 'height'))

# Define model -----
model <- function(height, weight){
  # Parameter prior distributions
  s = bi.dist.uniform(0, 50, name = 's')
  a = bi.dist.normal(178, 20, name = 'a')
  b = bi.dist.normal(0, 1, name = 'b')

  # Likelihood
  m$normal(a + b * weight, s, obs = height)
}

# Run mcmc -----
m$run(model) # Optimize model parameters through MCMC sampling

# Summary -----
m$summary()
```

## Mathematical Details

### *Frequentist Formulation*

The following equation allows us to draw a line:

$$Y_i = \alpha + \beta X_i + \sigma_i$$

Where:

- $Y_i$  is the dependent variable for observation  $i$ .
- $\alpha$  is the intercept term.
- $\beta$  is the regression coefficient.
- $X_i$  is the input variable for observation  $i$ .
- $\sigma_i$  is the error term for observation  $i$ .

### **Bayesian Formulation**

In the Bayesian formulation, we define each parameter with priors . We can express a Bayesian version of this regression model using the following model:

$$Y_i \sim Normal(\alpha + \beta X_i, \sigma)$$

$$\alpha \sim Normal(0, 1)$$

$$\beta \sim Normal(0, 1)$$

$$\sigma \sim Uniform(0, 50)$$

Where:

- $Y_i$  is the dependent variable for observation  $i$ .
- $\alpha$  and  $\beta$  are the intercept and regression coefficient, respectively.
- $X_i$  is the input variable for observation  $i$ .
- $\sigma$  is the standard deviation of the normal distribution, which describes the variance in the relationship between the dependent variable  $Y$  and the independent variable  $X$ .

## Notes

### Note

We can observe a difference between the *Frequentist* and the *Bayesian* formulation regarding the error term. Indeed, in the *Frequentist* formulation, the error term  $\sigma_i$  represents random fluctuations around the predicted values. This assumption leads to point estimates for  $\alpha$  and  $\beta$ , without accounting for uncertainty in these estimates. In contrast, the *Bayesian* formulation treats  $\sigma$  as a parameter with its own prior distribution. This allows us to incorporate our uncertainty about the error term into the model.

## Reference(s)

McElreath, Richard. 2018. *Statistical Rethinking: A Bayesian Course with Examples in r and Stan*. Chapman; Hall/CRC.