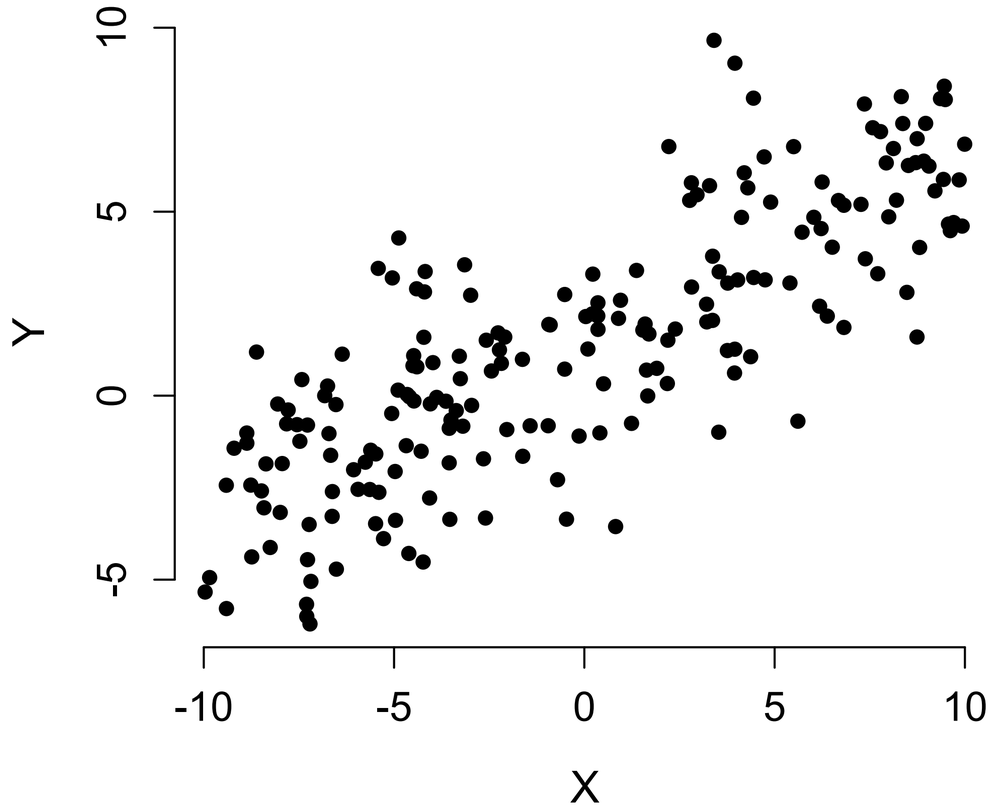
BAYESIAN GLMS

BAYESIAN STATISTICS FOR ECOLOGISTS

IGB 12. TO 19. NOVEMBER 2018

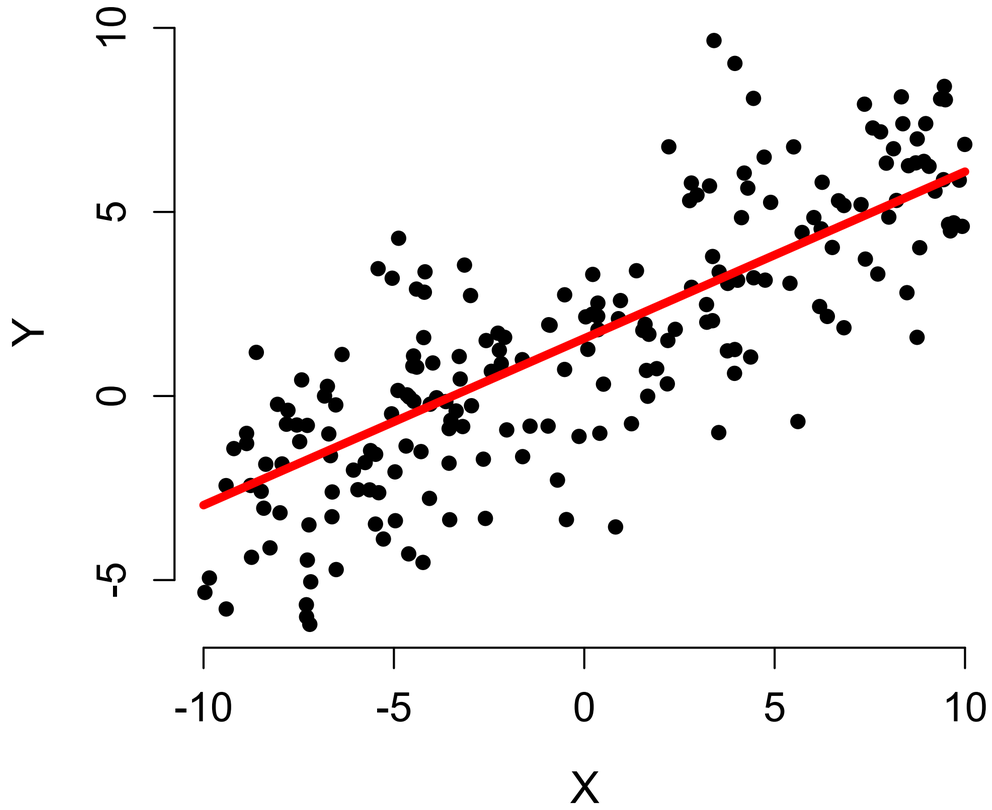
A BAYESIAN REGRESSION MODEL

- Goal: from observations of an i.i.d variable y and some covariate x, estimate the slope and intercept of the relationship between x and y



A BAYESIAN REGRESSION MODEL

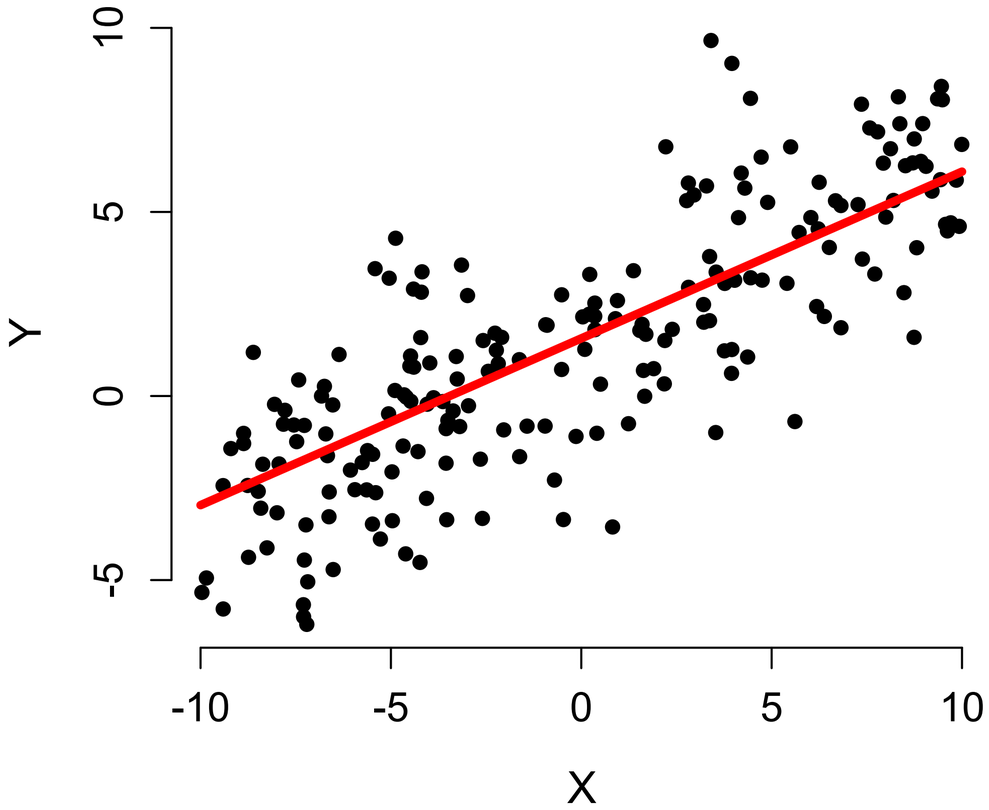
- Goal: from observations of an i.i.d variable y and some covariate x, estimate the slope and intercept of the relationship between x and y



A BAYESIAN REGRESSION MODEL

- Goal: from observations of an i.i.d variable y and some covariate x, estimate the slope and intercept of the relationship between x and y

$$\hat{y} = \alpha + \beta x$$

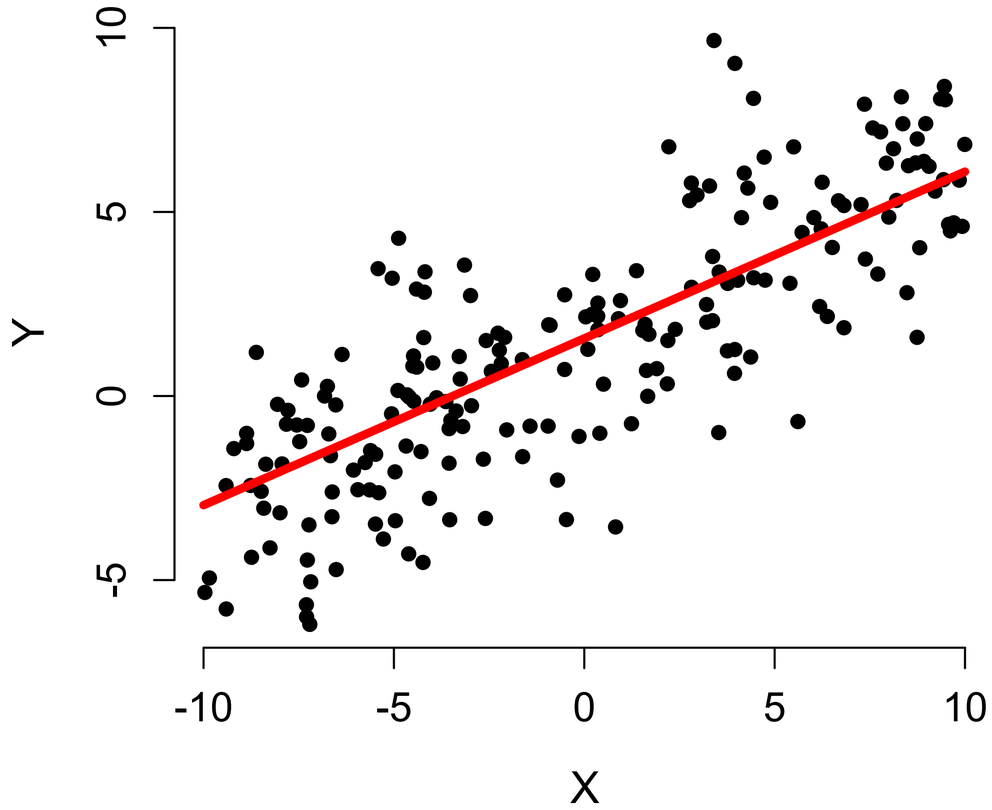


A BAYESIAN REGRESSION MODEL

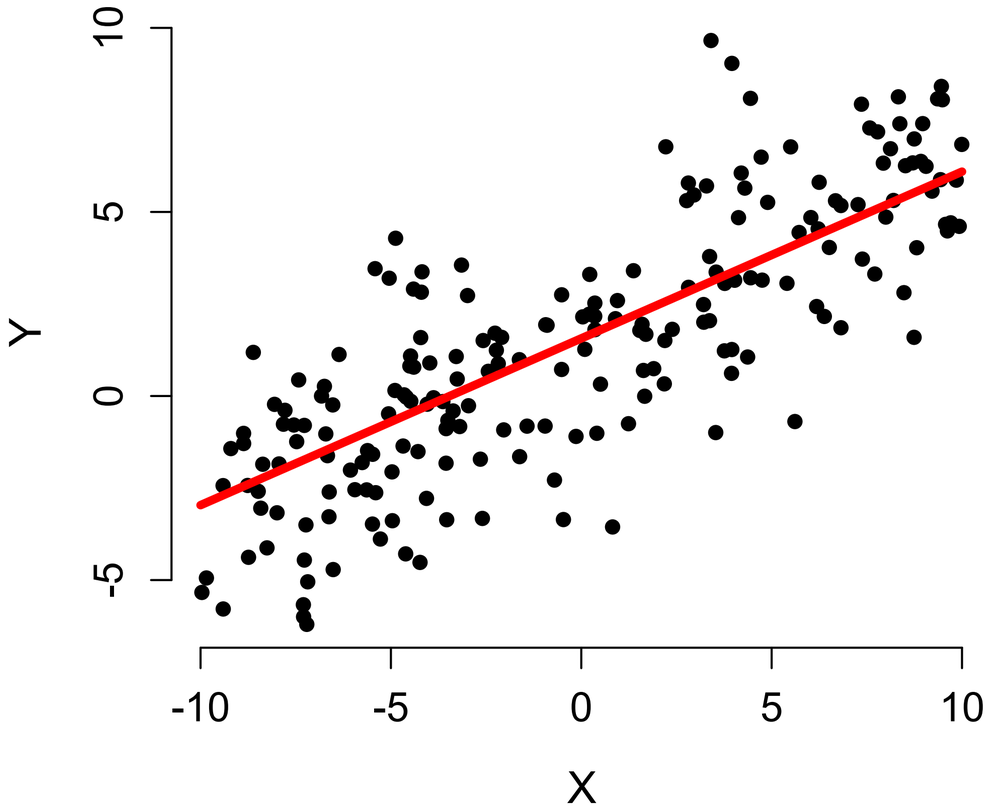
- Goal: from observations of an i.i.d variable y and some covariate x, estimate the slope and intercept of the relationship between x and y

$$\hat{y} = \alpha + \beta x$$

$$y \sim \mathbb{D}(.)$$



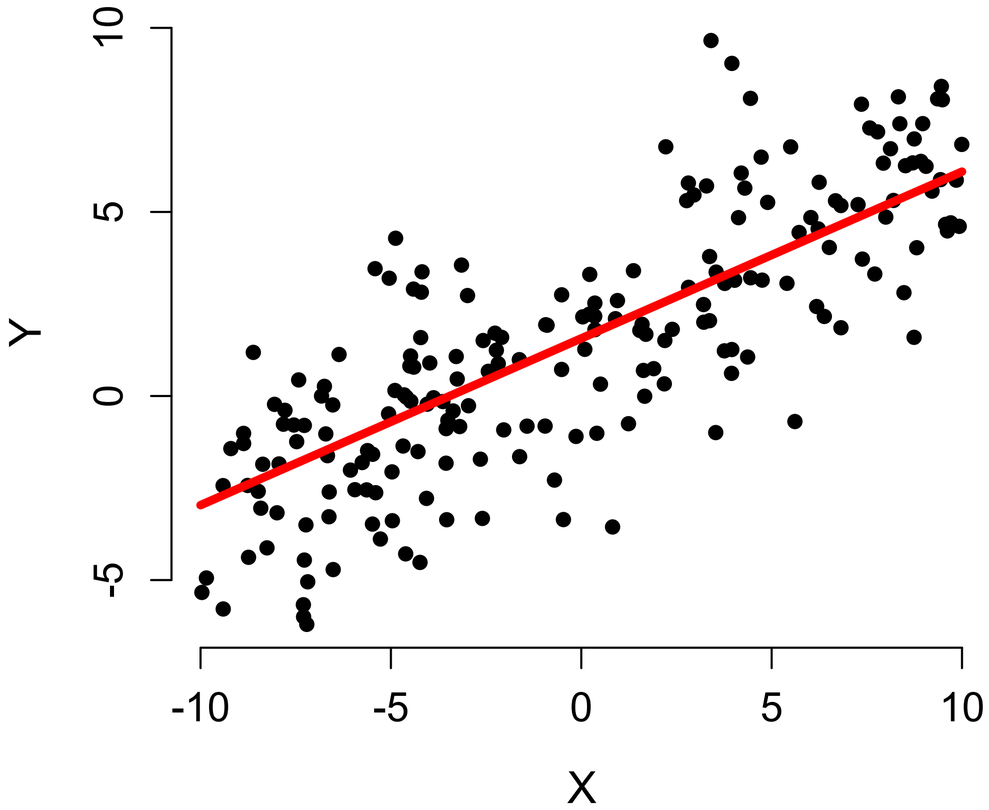
A BAYESIAN REGRESSION MODEL



A BAYESIAN REGRESSION MODEL

- We need to define some reasonable distribution for y

$$\hat{y} = \alpha + \beta x$$

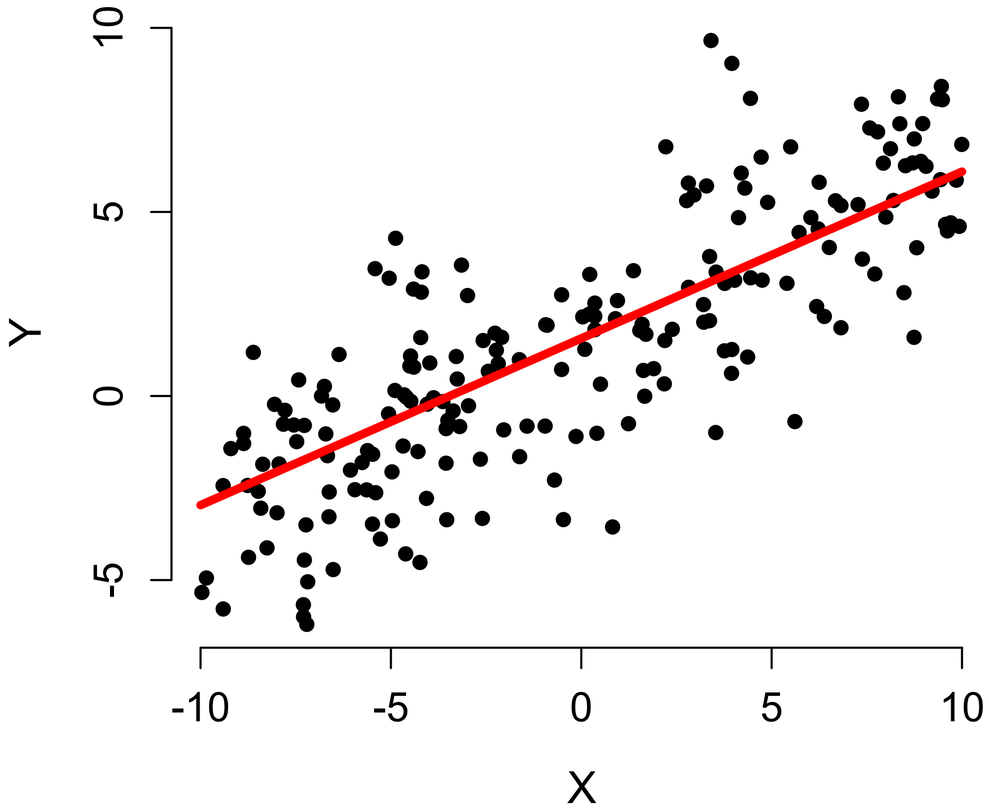


A BAYESIAN REGRESSION MODEL

- We need to define some reasonable distribution for y

$$\hat{y} = \alpha + \beta x$$

$$y \sim \mathbb{N}(\hat{y}, \sigma)$$



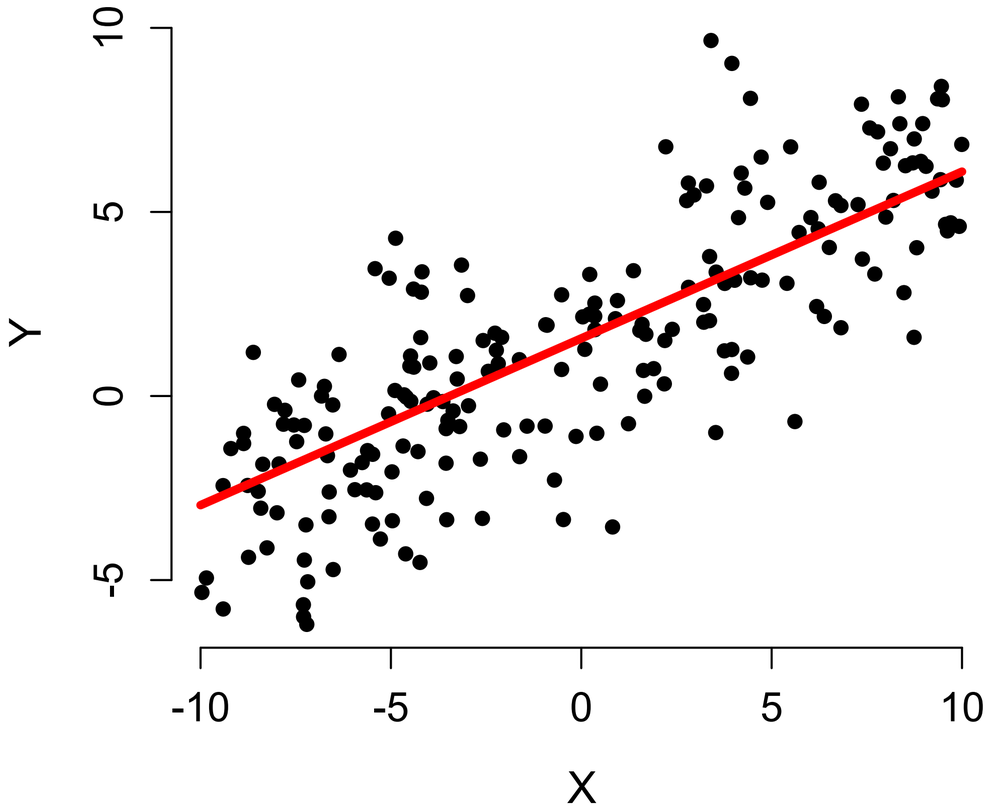
A BAYESIAN REGRESSION MODEL

- We need to define some reasonable distribution for y

$$\hat{y} = \alpha + \beta x$$

$$y \sim \mathbb{N}(\hat{y}, \sigma)$$

- y is a random variable, distributed normally with a mean equal to the regression line



BAYESIAN REGRESSION (R)

- The likelihood simply translates this idea into code
- Remember to use logs and to sum across the entire dataset
- What prior makes sense?

$$\hat{y} = \alpha + \beta x$$

$$y \sim \mathbb{N}(\hat{y}, \sigma)$$

```
log_lik <- function(alpha, beta, sigma, x, y) {
    y_hat <- alpha + beta * x
    ll <- sum(dnorm(y, y_hat, sigma, log=TRUE))
    return(ll)
}

log_prior <- function(alpha, beta, sigma) {

}
```

$$pr(y|x, \alpha, \beta, \sigma)$$

BAYESIAN REGRESSION (R)

- The likelihood simply translates this idea into code
- Remember to use logs and to sum across the entire dataset
- What prior makes sense?

```
log_lik <- function(alpha, beta, sigma, x, y) {
    y_hat <- alpha + beta * x
    ll <- sum(dnorm(y, y_hat, sigma, log=TRUE))
    return(ll)
}

log_prior <- function(alpha, beta, sigma) {
    lp <- dnorm(alpha, 0, 50, log = TRUE)
    lp <- lp + dnorm(beta, 0, 25, log=TRUE)
    lp <- lp + dexp(sigma, 0.1, log=TRUE)
    return(lp)
}
```

BAYESIAN REGRESSION LIKELIHOOD (STAN)

- The likelihood simply translates this idea into code
- In Stan, the we have to declare all variables
- No need to worry about the log likelihood, Stan takes care of this

```
data {
    int<lower=0> n; // number of data points
    vector[n] x;
    vector[n] y;
}
parameters {
    real<lower=0> sigma;
    real alpha;
    real bet;
}
transformed parameters {
    vector[n] y_hat;
    y_hat = alpha + bet * x;
}
model {
    y ~ normal(y_hat, sigma);
```

BAYESIAN REGRESSION LIKELIHOOD (STAN)

- The likelihood simply translates this idea into code
- In Stan, the we have to declare all variables
- No need to worry about the log likelihood, Stan takes care of this
- Prior as before from R

```
transformed parameters {
    vector[n] y_hat;
    y_hat = alpha + bet * x;
}
model {
    y ~ normal(y_hat, sigma);
    alpha ~ normal(0, 50);
    bet ~ normal(0, 25);
    sigma ~ exponential(0.1);
}
```

THE GENERALISED LINEAR MODEL

$$y \sim \mathbb{N}(\hat{y}, \sigma)$$

THE GENERALISED LINEAR MODEL

$$y \sim \mathbb{N}(\hat{y}, \sigma)$$

- If we have two x-variables, the model becomes

THE GENERALISED LINEAR MODEL

$$y \sim \mathbb{N}(\hat{y}, \sigma)$$

- If we have two x-variables, the model becomes

$$\hat{y} = \alpha + \beta_1 x_1 + \beta_2 x_2$$

THE GENERALISED LINEAR MODEL

$$y \sim \mathbb{N}(\hat{y}, \sigma)$$

- If we have two x-variables, the model becomes

$$\hat{y} = \alpha + \beta_1 x_1 + \beta_2 x_2$$

- Which generalises to a **matrix** of x-variables and a **vector** of βs

THE GENERALISED LINEAR MODEL

$$y \sim \mathbb{N}(\hat{y}, \sigma)$$

- If we have two x-variables, the model becomes

$$\hat{y} = \alpha + \beta_1 x_1 + \beta_2 x_2$$

- Which generalises to a **matrix** of x-variables and a **vector** of βs

$$\hat{y} = \alpha + \boldsymbol{\beta}\mathbf{x}$$

- …take a minute to review matrix multiplication

THE GENERALISED LINEAR MODEL

$$\mathbb{L}(\hat{y}) = \alpha + \boldsymbol{\beta}\mathbf{x}$$

$$y \sim \mathbb{D}(\mu = \hat{y}, \dots)$$

THE GENERALISED LINEAR MODEL

- Further, there is no requirement that the observed y's be distributed **normally,** nor that the relationship between y-hat and x be **strictly linear**

$$\mathbb{L}(\hat{y}) = \alpha + \boldsymbol{\beta}\mathbf{x}$$

$$y \sim \mathbb{D}(\mu = \hat{y}, \dots)$$

THE GENERALISED LINEAR MODEL

$$\mathbb{L}(\hat{y}) = \alpha + \boldsymbol{\beta}\mathbf{x}$$

$$y \sim \mathbb{D}(\mu = \hat{y}, \dots)$$

THE GENERALISED LINEAR MODEL

- The choice of **error distribution** and **link function** will depend on the problem
- For y's representing counts, a Poisson error distribution (likelihood) combined with a **log-link** can make sense
- For counts of successes given a number of trials, we might use **Binomial** with the **logit** or **probit** link

$$\mathbb{L}(\hat{y}) = \alpha + \boldsymbol{\beta}\mathbf{x}$$

$$y \sim \mathbb{D}(\mu = \hat{y}, \dots)$$

POISSON REGRESSION IN STAN/MCMC

- How does temperature affect the abundance of sugar maple?
- We will use only plots where the species occurs

```
trees <- readRDS("data/trees.rds")
dat <- trees[grepl("ACE-SAC", species) & n + born > 0]
```