Social Data Institute

Continuing Professional Development (CPD) course

Introduction To Bayesian Inference & Modelling (June 2025)

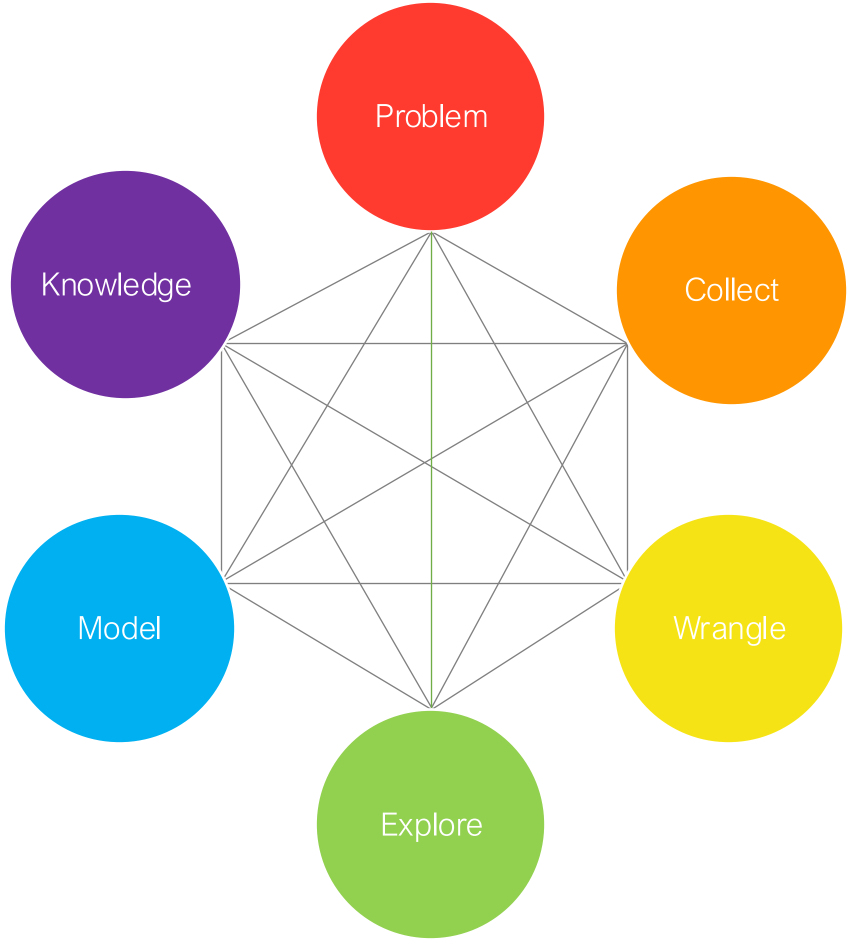
DAY 3: INTRODUCTION TO BAYESIAN GENERALISED LINEAR MODELS (GLM)

Dr Anwar Musah (a.musah@ucl.ac.uk)
Lecturer in Social and Geographic Data Science
UCL Geography

Additional details:
https://www.ucl.ac.uk/social-data

Contents

- What are Generalised Linear Models (GLMs)?
  - Link functions
- Selecting the appropriate type of statistical model
  - Linear regression model
  - Logistic regression model for Bernoulli OR Binomial
  - Poisson-based regression models (Normal, Negative Binomial & Zero-Inflated)
- What does each statistical model do?
  - Linear relationships
  - Log-odds and Odd Ratios (ORs)
  - Relative risk ratios (RRs)
- Interpretation of coefficients
- Model Specification from a Bayesian Framework



Multivariable linear regressions

Multivariable Linear Regression Model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \varepsilon$$

Definition: A regression model is a type of statistical device that measures the relationship between a dependent variable with one/more independent variables

Variables

- **y** is the dependent variable
- $\mathbf{x_1, x_2, x_3, \ldots, x_k}$ are the independent variables

Parameters

- β_0 is the intercept
- $\beta_1, \beta_2, \beta_3, \ldots, \beta_k$ are the slopes (or coefficients) for the corresponding variables x_1, x_2, x_3, …, x_k
- ε is the error term

Important notes

$$\underbrace{y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots}_{\text{Systematic part}} + \underbrace{\varepsilon}_{\text{Error part}}$$

[1] Model **with** the error term: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \varepsilon$

- This is the full, true model of how the data (y) is assumed to be generated. It reflects both **systematic part (the mean)** and the **random noise (ε)**
- Think of this model as the "factual" in the sense that it accounts for the uncertainty in real-world outcomes.
- We never actually see the **random noise (ε)**, but we know it's there.

(Expected average) Only the systematic part

[2] Model **without** the error term: $\mathbb{E}[Y] = \mu = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots$

- This is the model that we use to predict the expected value of our data (i.e., y). What we are saying is that it's best guess (on average) of what y will be, without the **random noise (ε).**
- This is not the full model – its only a **fitted/predictive model**.

In terms of regression, there are several types of models, each with there own families depending on the type distribution for the dependent variable:

Here is a board overview:

Distribution of dependent variable	Suitable Model
Continuous measures: e.g., average income in postcode (£); concentrations of ambient particulate matter (PM2.5); Normalised Vegetative Difference Index (NDVI) etc.,	Linear regression
Binary measures (1 = “present” or 0 = “absent”): e.g., Person’s voting for a candidate, lung cancer risk, house infested with rodents etc.,	Logistic Regression
Binomial measure (or proportion): e.g., prevalence of houses in a postcode infested with rodents, percentage of people in a village infected with intestinal parasitic worms, prevalence of household on a street segment victimised by crime etc.,	Logistic Regression
Counts or discrete measures: e.g., number of reported burglaries on a street segment, number of riots in a county etc.,	Poisson Regression
Time-to-event binary measures: e.g., Lung cancer risk due to chronic exposure to environmental levels of indoor radon. Risk of landslide and time dependence of surface erosion etc.,	Survival Analysis with Cox regression

In terms of regression, there are several types of models, each with there own families depending on the type distribution for the dependent variable:

Here is a board overview:

Distribution of dependent variable	Suitable Model
Continuous measures: e.g., average income in postcode (£); concentrations of ambient particulate matter (PM2.5); Normalised Vegetative Difference Index (NDVI) etc.,	Linear regression
Binary measures (1 = “present” or 0 = “absent”): e.g., Person’s voting for a candidate, lung cancer risk, house infested with rodents etc.,	Logistic Regression
Binomial measure (or proportion): e.g., prevalence of houses in a postcode infested with rodents, percentage of people in a village infected with intestinal parasitic worms, prevalence of household on a street segment victimised by crime etc.,	Logistic Regression
Counts or discrete measures: e.g., number of reported burglaries on a street segment, number of riots in a county etc.,	Poisson Regression
Time-to-event binary measures: e.g., Lung cancer risk due to chronic exposure to environmental levels of indoor radon. Risk of landslide and time dependence of surface erosion etc.,	Survival Analysis with Cox regression

What are Generalised Linear Models?

Definition:

Generalised linear model (GLMs) is a flexible generalisation of ordinary linear regression model, which allows the user to model the expected value (or average) of outcome *y that is nonlinear*, linearly with predictor variables.

$$g(\ldots) \rightarrow \eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots$$

There are many different types of outcomes that require being modelled within a GLM framework. The most common examples are:

- Bernoulli (Binary category)
- Binomial (Aggregation of binary categories)
- Multinomial (More than two categories)
- Poisson (Counts)
- Negative binomial (Counts with overdispersion in its frequency)

Notes 1: There are a tonne of them, but you really don't have to worry about any of them. You only need to concern yourself with how this link function works!

What is a link function $g(\eta)$? [1]

$$\mathbb{E}[Y] = \mu = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots$$

- **Linear regression models** assume normally distributed residuals, as well as linearity in the relationship between the expected values of the dependent and independent variables. It is best suited for continuous outcomes.

- **Generalised Linear Models (GLMs)** extend this framework to support outcomes from other distributions that typically nonlinear or not continuous such as binary, count, or categorical variables. A link function $g(\,)$ therefore relates expected values of such nonlinear response linearly to a combination of predictors.

$$g(\ldots) \rightarrow \eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots$$

Link function apply to some outcome (binary, binomial etc.)

That outcome transformed accordingly to nu

Notes 1: This link function apply applies a transformation to the dependent variable, so its nonlinear relationship can be modelled linearly – its on a transformed scale

What is a link function $g(\eta)$? [2]

Here are the most frequent examples which you will certainly encounter

Distribution of dependent variable	Exponential Family (Distribution)	Link Function	Suitable Model
Continuous measures	Normal distribution	Identity (we've been using this all this while)	Linear regression
Binary measures (1 = "present" or 0 = "absent")	Bernoulli distribution	Logit	Logistic Regression
Binomial measure (or proportion)	Binomial distribution	Logit function on aggregated outcome for successful and failures	Logistic Regression
Counts or discrete measures	Poisson distribution	Log or ln	Poisson Regression

Logistic Regression [1]

- This model allows the user to model binary outcomes linearly with other independent variables

- Examples of such outcomes can be from **Bernoulli distribution** e.g., disease status: no disease = 0 or disease = 1; Victimisation status: not burgled = 0 or burgled = 1; etc.,

- Other examples can also be from a **Binomial distribution** where binary responses are aggregated: e.g. total number of individual surveyed in a village (*N*) and number people detected to be positive (*n*)

- Link function:

$\mathrm{g}(\ldots) = \mathrm{logit}(\mathrm{p})$, where p is a probability

$\mathrm{logit}(\mathrm{p}) = \ln\left(\frac{p}{1-p}\right)$ is what we called the "log-odds"

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots$$

Notes 1: With binary outcomes, we are dealing with probabilities and not averages

Logistic Regression [2]

$\mathrm{g}(\eta) = \mathrm{logit}(\mathrm{p})$, where p is a probability

$\mathrm{logit}(\mathrm{p}) = \ln\left(\frac{p}{1-p}\right)$ is what we called the "log-odds"

$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots$

- When estimate our coefficients i.e., β_i, which shows the linear relationship between the binary or binomial response variable with independent variable x_i - they are always on the log-odds scale.

- For interpretability: we always take the exponential of our coefficient i.e., $\exp(\beta_i)$, to convert in onto the scale of **odds ratios (OR)**

This is the quantity i.e., **Odds Ratios (OR)**, we want to estimate and interpret from our logistic regression

Interpretation of Odds Ratios (OR)

OR = 1 (null value), it means that independent variable has no effect on the outcome

OR < 1, the independent variable has an impact on the outcome – in this case, its reduced effect, or reduced risk on the outcome

OR > 1, the independent variable has an impact on the outcome – and so, in this case, its increased effect, or increased risk on the outcome

Notes 1: If you want to do a risk assessment and your study design is either a cross-sectional or case-control study design with some binary/binomial outcome, use a logistic regression model.

Table: Estimated posterior odds ratios for prevalence of anaemia at baseline, using a Bayesian hierarchical logistic regression model.

Variable	Odds ratio	95% credible interval
Fixed part		
Individual-level (*n=1261*)		
Sex (reference category: Female)		
Male	1.125	(0.887 – 1.425)
Age (reference category: 7 years)		
8 years	0.799	(0.526 – 1.207)
9 years	1.091	(0.716 – 1.668)
10 years	0.730	(0.477 – 1.109)
11 years	0.653	(0.424 – 1.005)
12 years	1.016	(0.635 – 1.622)
HAZ (reference category: Not stunted)		
Stunted growth	1.419	(1.053 – 1.918)[a]
S.haematobium (reference category: Not infected)		
Low infection	1.444	(1.017 – 2.051)[a]
High infection	2.294	(1.388 – 3.834)[a]
Hookworm (reference category: Not infected)		
Light infection	0.758	(0.563 – 1.021)
Moderate/Heavy infection	1.937	(0.542 – 6.917)
Malaria parasitaemia (ref. category: Not infected)		
Infected	0.870	(0.620 – 1.221)

Notes 1: An example of logistic regression model, applied to health risk assessment study I did for my MSc dissertation in 2011, a long time ago, using data from the Schistosomiasis Control Initiative (SCI) for determining the impacts of various parasitic and other factors on risk of anaemia in school children in Tanzania.

Interpretation:

Children at baseline with light *S.haematobium* intensities were 1.444 times more likely to be anaemic compared to an uninfected children, (95% credible interval: 1.017 – 2.051).

Furthermore, those with heavy *S.haematobium* intensities are at more than twice the risk of being anaemic (Odds ratio: OR = 2.294, 95% credible interval: 1.388 – 3.834).

Poisson Regression [1]

- This model allows the user to model count or discrete outcomes linearly with other independent variables
- Examples of such outcomes can be from **Poisson distribution** e.g., number of COVID cases in postcodes across London; Number of houses on street segments that were victims to burglary etc.
- With particular scenario we are dealing with aggregated units and its either **counts** or **rates**
- Link function:

$g(\dots) = \ln(\lambda_i)$ i.e., log-link function (log of some mean rate λ_i).

$$\ln(\lambda_i) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots$$

OR

$$\ln(\lambda_i) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \text{offset}$$

Often an offset is include to adjust for denominators if the outcome was measured as a rate.

Poisson Regression [2]

$g(\dots) = \ln(\lambda_i)$ i.e., log-link function (log of some mean rate λ_i).

$\ln(\lambda_i) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots$

- When we estimate our coefficients i.e., β_i, which shows the linear relationship between the counts or discrete response variable with independent variable x_i - they are always on the log-scale.

- For interpretability: we always take the exponential of our coefficient i.e., exp(β_i), to convert it onto the scale of **risk ratios (RR)**

This is the quantity i.e., **Risk Ratios (RR) (interchangeable with Relative Risk)**, we want to estimate and interpret from our Poisson regression

Interpretation of Risk Ratios (RR)

RR= 1 (null value), it means that independent variable has no effect on the outcome

RR < 1, the independent variable has an impact on the outcome – in this case, its reduced effect, or reduced risk on the outcome

RR > 1, the independent variable has an impact on the outcome – and so, in this case, its increased effect, or increased risk on the outcome

Notes 1: Use the relative risks - if you want to do a risk assessment for ecological study design where you have counts as the outcome (and denominators for offset – for rates). Also use this as well for individual-level cohort studies when following groups of participants prospectively

Table: Estimated posterior relative risks exploring the overall association for area-level socioeconomic deprivation with fire-related casualty rates in residential dwellings in England

Notes 1: An example of Poisson-based regression model, applied to fire-related casualty risk within home dwellings, spatially, across Fire Service Areas in England.

IMD Domain	Unadjusted Relative Risk (95% Credibility Intervals)		
	RR Estimates	Percentage	
LE	1.248 (1.175–1.329)	+24.8%	(+17.5% to +32.9%)
EST	1.237 (1.163–1.316)	+23.7%	(+16.3% to +31.6%)
BHS	0.918 (0.863–0.975)	−8.2%	(–13.7% to −2.5%)
Crime	1.134 (1.069–1.201)	+13.4%	(+6.90% to +20.1%)

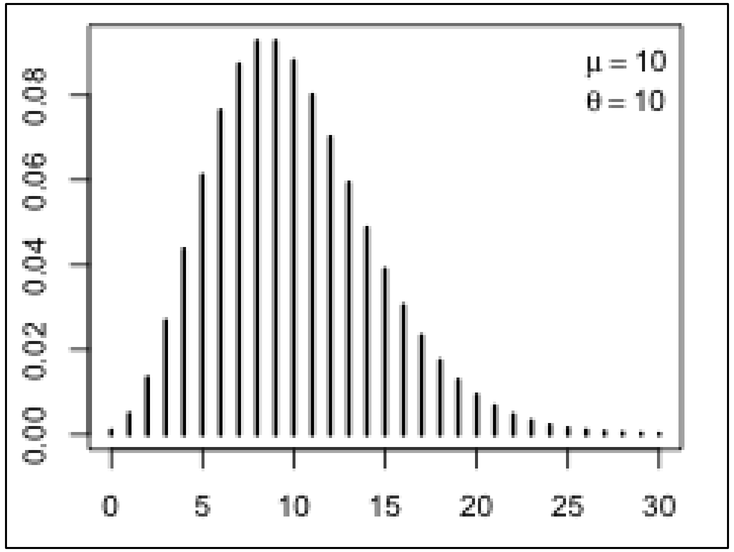
Interpretation:

When the levels of deprivation at an FSA-level increase, we found the following for the domains: Living Environment and Education, Skills & Training, significantly increases the risk of fire-related dwelling casualties by up to 25.0% in England. For elevated levels of crime, the risk of fire-related dwelling casualties increases 13.4%.

Li et al. 2022 [Source]

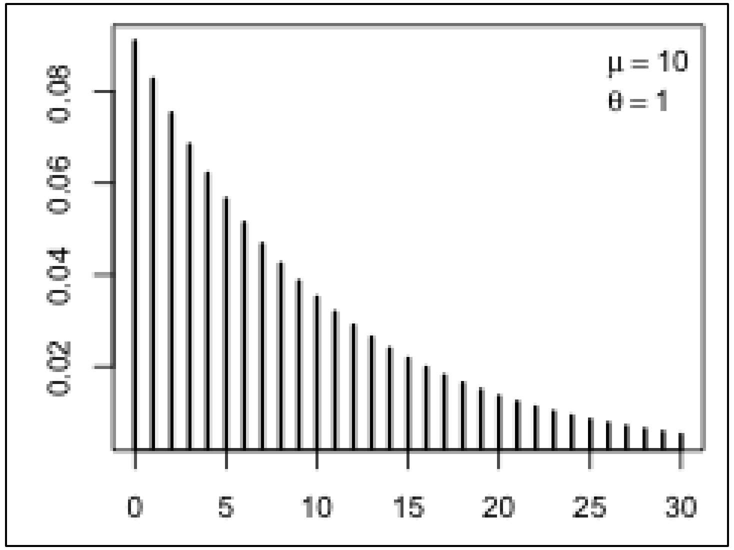
Types of Poisson Regression

Examine the frequency distribution of the count response



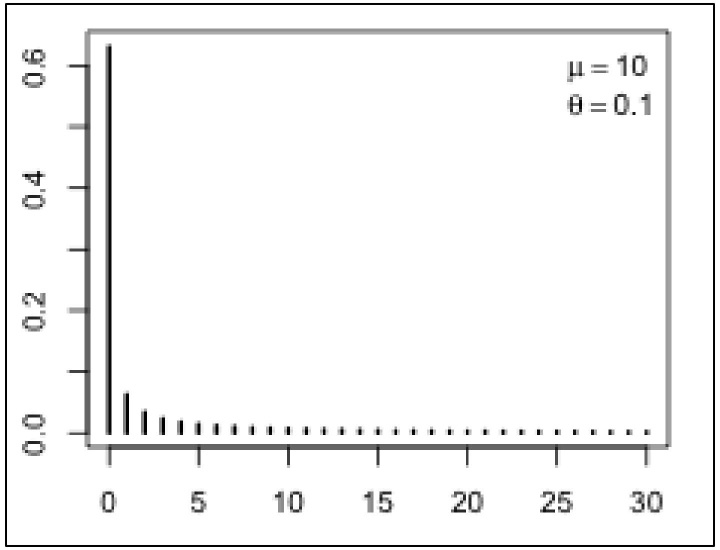
Scenario 1: Little to no dispersion

Use: regression Poisson model



Scenario 2: Over dispersed

Use: Negative Binomial Poisson model



Scenario 3: Strong over-dispersed response

Use: Zero-inflated Poisson model

Use as last restore if scenario is gives you unstable results.

Example with exceedance probabilities

Linear model

Characteristics	Coefficients (95% CrI)	Uncertainty $P(\beta > 0)$	ESS	$\hat{R}$
Intercept	+14.50 (95% CrI: +5.78 to +21.66)	0.990	8,085	< 1.05
Sociodemographic attributes				
Nationality				
Jordanian (referent)				
Syrian	+1.01 (95% CrI: -3.24 to +5.28)	0.675	39,863	< 1.05
Educational attainment (highest)				
None (referent)				
Primary	+0.13 (95% CrI: -3.29 to +3.53)	0.531	30,575	< 1.05
Secondary	+2.26 (95% CrI: -1.76 to +6.35)	0.864	49,269	< 1.05
University	-4.68 (95% CrI: -9.91 to +0.72)	0.044	11,335	< 1.05
Total number household members	+0.78 (95% CrI: +0.22 to +1.33)	0.997	52,924	< 1.05
Household income				
0 to 350 (referent)				
351 to 700	+2.30 (95% CrI: -2.14 to +6.79)	0.850	29,834	< 1.05
Prefer not to say	+1.26 (95% CrI: -2.02 to +4.52)	0.774	7,171	< 1.05
Water Usage				
Weekly Water Supply (No of times)				
Supplied twice a week (referent)				
Supplied once a week	-0.73 (95% CrI: -3.68 to +2.19)	0.314	26,988	< 1.05
Supplied once every two weeks	+2.97 (95% CrI: -5.64 to +11.55)	0.751	15,520	< 1.05
Supplied Water meeting our needs				
Yes (referent)				
No	+0.04 (95% CrI: -3.19 to +3.32)	0.513	16,669	< 1.05
Use other alternative water sources				
No (referent)				
Yes	+1.73 (95% CrI: -1.79 to +5.14)	0.842	15,385	< 1.05
Recycling of Water				
No (referent)				
Yes	+1.59 (95% CrI: -1.03 to 4.27)	0.885	9,278	< 1.05

*Indicates that the coefficient is significant on the grounds that excludes the null value of 0.00 between the lower and upper bounds of the 95% credibility interval (95% CrI); Effective Sample Size (ESS) for each parameter has enough statistical power as sampling from the posterior distribution yielded at least 3,000 samples per chain after 30,000 iterations. Each estimate is valid since the $\hat{R}$ is lower than 1.05.

How do you code a Bayesian GLM in RStudio? [1]

Specifications for model block:

- Linear regression: norm()
- Logistic regression (Y : 1 or 0): bernoulli_logit()
- Logistic regression (Y : numerator & denominators): binomial_logit()
- Poisson regression (Y : counts or rates; normal): poisson_log()
- Poisson regression (Y : counts or rates; over-dispersed or zero-inflated): neg_binomial_2_log()

Let’s look at a simple linear regression case

Stan code

```
data {
  int<lower=0> N;          // sample size N
  int<lower=0> k;          // number of variables 2
  matrix[N, k] X;          // matrix: independent variables
  vector[N] y;             // vector/array for outcome
}

parameters {
  real beta0;               // Intercept
  vector[k] beta;           // beta coefficients
  real<lower=0> sigma;      // standard deviation
}

transformed parameters {
   vector[N] mu;
   mu = beta0 + X*beta;
}

model {
  beta0 ~ normal(0, 20);   // Prior for beta0
  beta ~ normal(0, 5);    // Prior for beta1 and 2
  sigma ~ cauchy(0, 2.5);   // Prior for sigma
  y ~ normal(mu, sigma);   // Likelihood function
}
```

Model formulation

- Simple GLM (Linear case)

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

- Specify likelihood function. The outcome is continuous – thus it normal (so no link function is need here).

$$y \sim \mathbf{norm}(\boldsymbol{\mu}, \boldsymbol{\sigma})$$
$$\mu = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

- Define the priors for the intercept, coefficients and other parameters, e.g., standard deviation

$$\beta_0 \sim \mathrm{Norm}(0, 20)$$
$$\beta_1 \sim \mathrm{Norm}(0, 5)$$
$$\beta_2 \sim \mathrm{Norm}(0, 5)$$
$$\sigma \sim \mathrm{cauchy}(0, 2.5)$$

- Build Bayesian model

Recall the Bayes' Rule: $P(\theta|Y) \propto P(Y|\theta)P(\theta)$

$$P(\boldsymbol{\beta}_0, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \sigma|\mu) \propto P(\mu|\boldsymbol{\beta}_0, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \sigma)\ P(\boldsymbol{\beta}_{0,})P(\boldsymbol{\beta}_1)P(\boldsymbol{\beta}_2)P(\sigma)$$

This is my model

$$y \sim \mathbf{norm}(\boldsymbol{\mu}, \boldsymbol{\sigma})$$

$$\mu = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

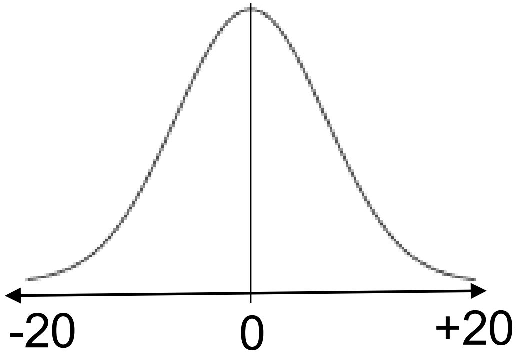
These are my priors for the coefficients and standard deviation

$$\beta_0 \sim \mathrm{norm}(0, 20)$$
$$\beta_1 \sim \mathrm{norm}(0, 5)$$
$$\beta_2 \sim \mathrm{norm}(0, 5)$$
$$\sigma \sim \mathrm{cauchy}(0, 2.5)$$

What are we saying?

Intercept

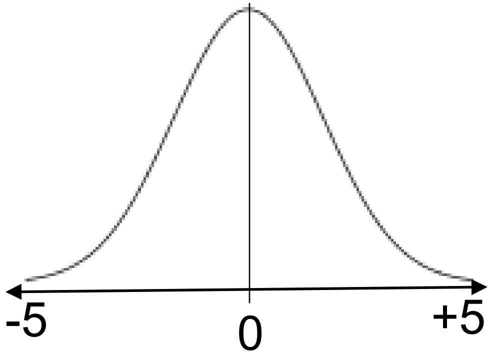
$$\beta_0 \sim \mathrm{norm}(0, 20)$$



β_0 is centred at 0 but it's distribution or value can vary ± 20

Coefficient for variable 1

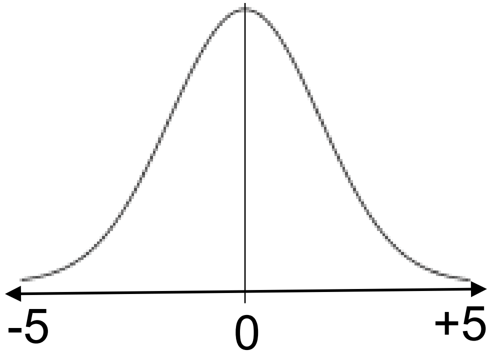
$$\beta_1 \sim \mathrm{norm}(0, 5)$$



β_1 is centred at 0 but it's distribution or value can vary ± 5

Coefficient for variable 2

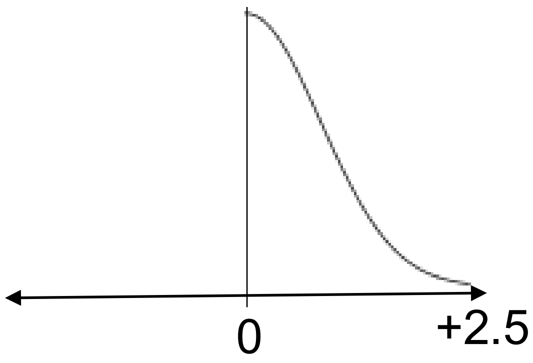
$$\beta_2 \sim \mathrm{norm}(0, 5)$$



β_2 is centred at 0 but it's distribution or value can vary ± 5

Standard deviation

$$\sigma \sim \mathrm{cauchy}(0, 2.5)$$



σ is small and can only take positive values from 0 to 2.5. Always use a half-Cauchy for the SD. Else, throw-in a uniform.

Any questions?