GEOG0125

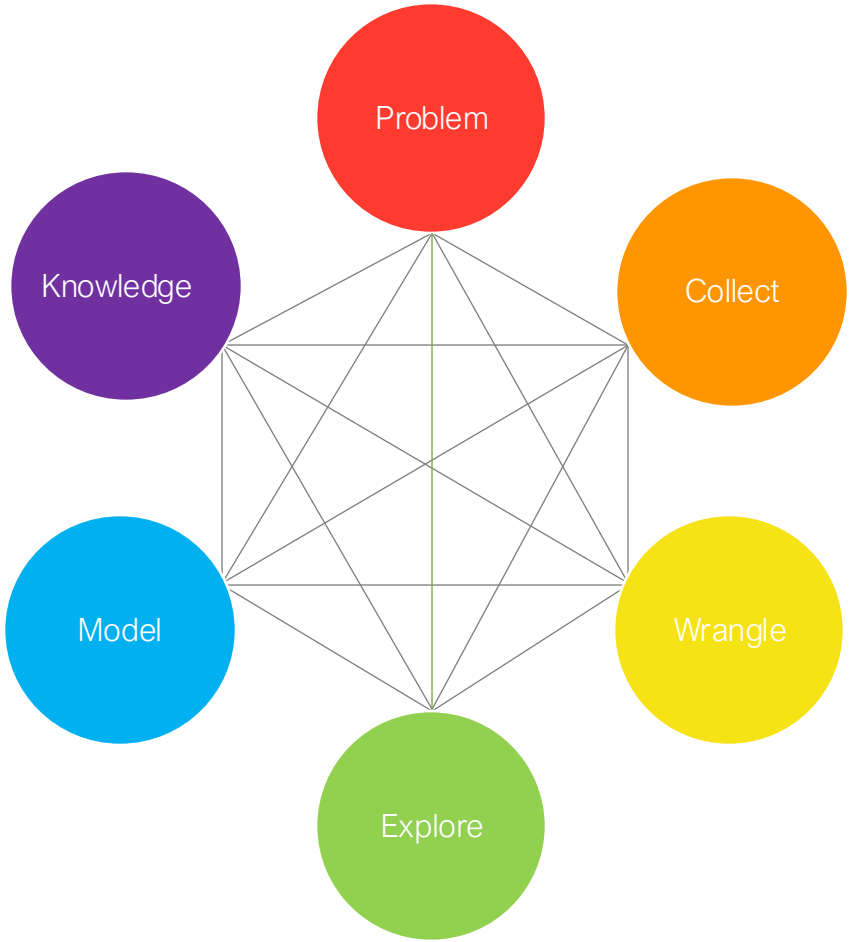
ADVANCED TOPICS IN SOCIAL AND GEOGRAPHIC DATA SCIENCE

INTRODUCTION TO BAYESIAN GENERALISED LINEAR MODELS (GLM)

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

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



Remember in Term 1…
In Week 9's PSA lecture, we said…

Multivariable Linear Regression Model

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

Notes 1: Remember, in term 1 (week 9), we described what a linear regression model was before discussing at length what spatial lag and error models were etc.

Notes 2: We mentioned that a linear regression model such as the above formula allows the user to quantify the relationship (or association) between a **continuous** outcome (i.e. dependent variable) with one, or more predictors (i.e., independent variable(s)). These models are good for making causal and predictive inference

Variables

- **y** is the dependent variable
- **x_1, x_2, x_3, …, 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

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:

Notes 1: Recall that we described how linear regression models are best suited for modelling outcomes that are only continuous measures, whereby we assumed that such continuous measures are from a Gaussian/normal distribution. Before, deep diving into spatial lag and error regression models… because are from the family of linear models but a spatial component to it.

Distribution of dependent variable	Suitable Model
Continuous measures: e.g., average income in postcode (£); concentrations of ambient particular 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:

Notes 1: Recall that we only touched on outcomes that can follow a different distribution, and models can potentially be violated if the inappropriate outcome is fitted into the wrong model!

Here is a board overview:

Distribution of dependent variable	Suitable Model
Continuous measures: e.g., average income in postcode (£); concentrations of ambient particular 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 (GLMs)?

Definition:

Generalised linear model (GLMs) is a flexible generalisation of ordinary linear regression model, which allows the user to link some outcome *y*, to a link function g(η), when that outcome is characterised by distribution that is from one the exponential families of distribution.

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

Exponential family are set of parametric (i.e., discrete or continuous) probability distributions. There are many… but the most common examples are:

- Normal
- Bernoulli (binary category)
- Binomial (aggregated binary)
- Multinomial (multiple categories)
- Poisson (counts)
- Negative binomial (counts with overdispersion)

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

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

- By default, the linear regression model does not support any other outcome whose distribution is not from a Gaussian/Normal distribution.
- However, by using some **link function**, it allows the user to transform such outcome (i.e., that's considered binary, polychotomous, discrete etc.,) in something that behave like a linear function

Notes 1: Tricking the model to thinking it is linear

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

- The type of link function implemented on a model depends on the type of analysis you are going to perform

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 (log-odds)	Logistic Regression
Binomial measure (or proportion)	Binomial distribution	Logit function on aggregated outcome for successful and failures (log-odds)	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:

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

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

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

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

Logistic Regression [2]

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

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

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

- When we estimate our coefficients i.e., β_i, it 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 it into a odds ratios (OR)

This is the quantity i.e., Odds Ratios (OR), we want to report 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: In the Frequentist approach, we use p-values and 95% CIs to deem whether the odd ratios are statistically significant or not. The Bayesian framework, only 95 credibility intervals are needed for significance.

Table 2. Results of logistic regression models of iAs exposure by quintile and BCC in the ASHRAM study population [OR (95% confidence interval)].

Arsenic exposure index/quintile (range of exposure in controls)	Adjusted[a]	Additionally adjusted[b]	Trend test (*p*-value)
Lifetime average iAs concentration (μg/L)			0.001
0.00–0.68	1.00	1.00	
0.68–0.98	1.27 (0.82, 1.97)	1.39 (0.89, 2.19)	
0.98–7.00	1.02 (0.67, 1.56)	1.20 (0.77, 1.88)	
7.10–19.43	1.63 (0.93, 2.85)	1.73 (0.97, 3.11)	
19.54–167.29	2.81 (1.62, 4.87)	3.03 (1.70, 5.41)	
Peak daily iAs dose rate (μg/day)			0.001
0.00–0.73	1.00	1.00	
0.73–1.48	0.93 (0.62, 1.39)	0.91 (0.59, 1.39)	
1.48–9.09	1.29 (0.86, 1.95)	1.55 (1.00, 2.41)	
9.09–32.23	1.78 (1.05, 3.02)	1.76 (1.01, 3.07)	
32.23–242.14	2.31 (1.32, 4.03)	2.50 (1.39, 4.49)	
Cumulative iAs dose (g)			0.001
0.00–0.01	1.00	1.00	
0.01–0.03	1.02 (0.68, 1.52)	1.09 (0.72, 1.67)	
0.03–0.13	1.19 (0.78, 1.81)	1.46 (0.93, 2.27)	
0.13–0.55	1.73 (1.02, 2.91)	1.76 (1.02, 3.04)	
0.55–4.46	2.45 (1.39, 4.32)	2.63 (1.45, 4.78)	

Notes 1: An example of logistic regression model, applied to health risk assessment study determining the impacts of arsenic exposure (biomarkers) and skin cancer risk in Eastern Europe.

Interpretation for independent variable that is categorical

Giovanni et al. 2012 [Source]

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.
- It is use for dealing with aggregated units the contain information of **counts** or **rates (expressed per capita)**
- Link function:

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

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

OR

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

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

Poisson Regression [2]

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

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

- 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., $\exp(\beta_i)$, to convert it into a **risk ratios (RR)**

This quantity i.e., **Risk Ratios (RR) (interchangeable with the term Relative Risk Ratios)**, is the thing we want to report 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, it's a reduced effect, or reduced risk of the outcome

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

Notes 1: From the Frequentist approach, We use p-values and 95% CIs to deem whether the risk ratios are statistically significant or not. In the Bayesian, we don't deal with p-values, only 95% credibility intervals

Table 2

Using a negative binomial Poisson regression model to report multivariable associations between street-level exposures and residential burglaries in Kaduna, Nigeria.

Street exposure variables	Residential Burglary	
	CRR (95% CI)	VIF
Intercept	0.69 (0.50–0.95)*	–
Length of street segment (m)	1.01 (0.99–1.02)	1.83
Connectivity	1.05 (1.03–1.08)**	2.62
Betweeness (normalised index) (quartiles)		2.33
1st Quartile (lowest)	1.00 (referent)	
2nd Quartile	1.29 (0.99–1.66)	
3rd Quartile	1.55 (1.20–2.02)*	
4th Quartile (highest)	1.64 (1.19–2.29)*	
Closeness (normalised index) (quartiles)		1.12
1st Quartile (lowest)	1.00 (referent)	
2nd Quartile	0.62 (0.49–0.78)*	
3rd Quartile	0.78 (0.61–0.99)*	
4th Quartile (highest)	0.62 (0.48–0.81)*	
Business activity index (z-scores) (quintiles)		1.18
1st Quintile (lowest)	1.00 (referent)	
2nd Quintile	0.96 (0.72–1.29)	
3rd Quintile	0.71 (0.57–1.03)	
4th Quintile	1.47 (1.15–1.86)*	
5th Quintile (highest)	1.31 (1.01–1.68)*	
Socioeconomic status (z-scores) (quintiles)		1.15
1st Quintile (lowest)	1.00 (referent)	
2nd Quintile	1.28 (1.00–1.63)	
3rd Quintile	0.95 (0.74–1.21)	
4th Quintile	0.79 (0.61–1.02)	
5th Quintile (highest)	0.81 (0.63–1.05)	

*Significant with p-value < 0.05; Crime Rate Ratios (CRR); 95% Confidence Intervals (95% CI); Variance Inflation Factor (VIF).

Notes 1: An example of Poisson-based regression model, applied to crime victimisation study to determine the impacts of various environmental and society society risk factor (quantified on a street-level) and burglary risk in Nigeria.

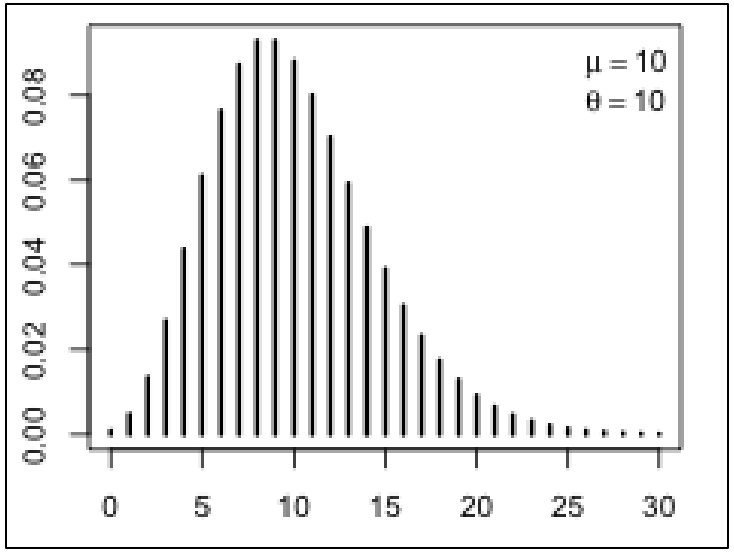
Interpretation for independent variable that are continuous as well as those that are categorical

Risk ratio which has been operationalised and termed as Crime risk ratio (CRR)

Musah et al. 2020 [Source]

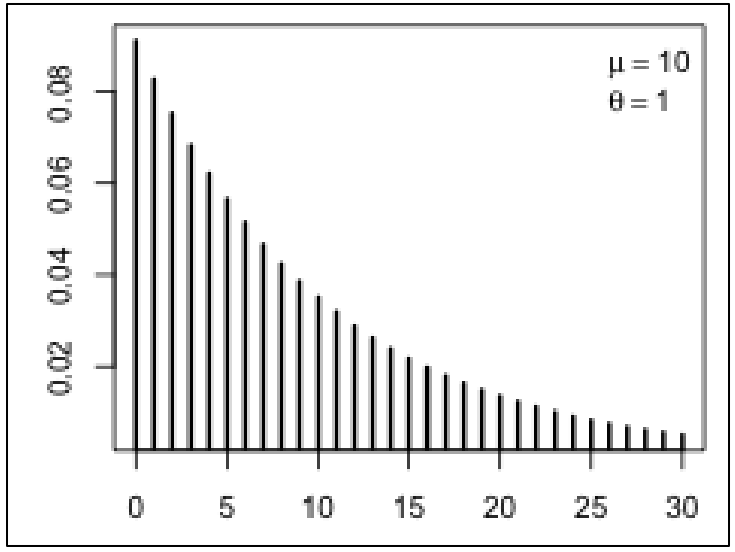
Types of Poisson Regression

Examine the frequency distribution of the count response



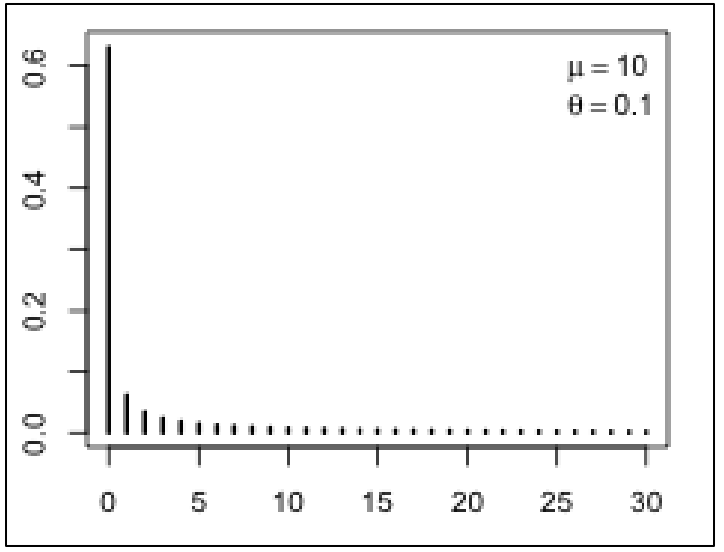
Scenario 1: Little to no dispersion

Use: Standard Poisson model



Scenario 2: Over dispersed

Use: Negative Binomial Poisson model



Scenario 3: Strong over-dispersed response

Use: Zero-inflated Poisson model

GLMs in a Bayesian Framework

How do you code a Bayesian GLM in RStudio?

Specifications for model block:

- 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

Question:

What set of sociodemographic and water usage variables have an impact of overall water expenditure bills among Syrian refugees in stationed in camps in Jordan?

Stan code

```
data {
  int<lower=0> N;          // sample size N
  int<lower=0> k;          // number of variables 3
  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, 2 and 3
  sigma ~ cauchy(0, 2.5);  // Prior for sigma
  y ~ normal(mu, sigma);  // Likelihood function
}
```

Model formulation

- Simple GLM (Linear case)

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$$

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

$$y \sim \mathbf{norm}(\boldsymbol{\mu}, \boldsymbol{\sigma})$$
$$\mu = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

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

$$\beta_0 \sim \text{Norm}(0, 20)$$
$$\beta_1 \sim \text{Norm}(0, 5)$$
$$\beta_2 \sim \text{Norm}(0, 5)$$
$$\beta_3 \sim \text{Norm}(0, 5)$$
$$\sigma \sim \text{cauchy}(0, 2.5)$$

- Build Bayesian model

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

$$P(\boldsymbol{\beta}_0, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \sigma|\mu) \propto P(\mu|\boldsymbol{\beta}_0, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \sigma)\, P(\boldsymbol{\beta}_0) P(\boldsymbol{\beta}_1) P(\boldsymbol{\beta}_2) P(\boldsymbol{\beta}_3) P(\sigma)$$

This is my model

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

$$\mu = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

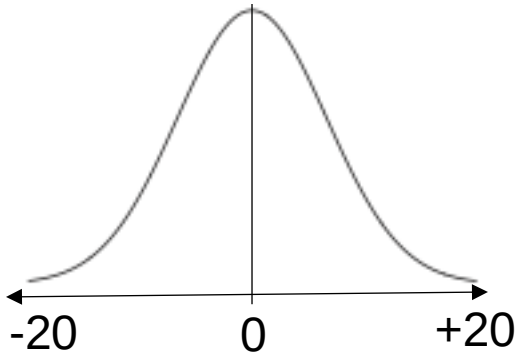
These are my priors for the coefficients and standard deviation

$$\beta_0 \sim \text{norm}(0, 20)$$
$$\beta_1 \sim \text{norm}(0, 5)$$
$$\beta_2 \sim \text{norm}(0, 5)$$
$$\beta_3 \sim \text{norm}(0, 5)$$
$$\sigma \sim \text{cauchy}(0, 2.5)$$

What are we saying?

Intercept

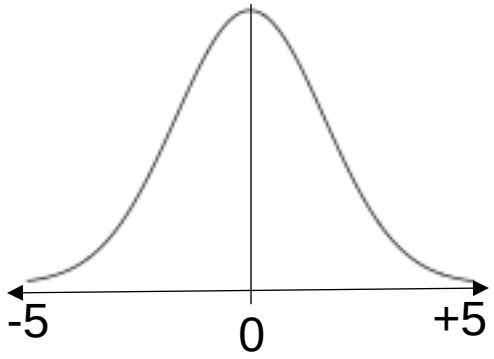
$\beta_0 \sim \text{norm}(0, 20)$



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

Coefficient for variable 1

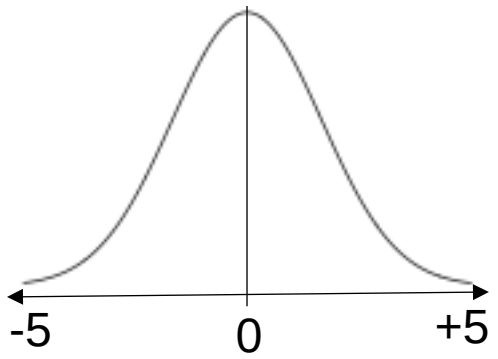
$\beta_1 \sim \text{norm}(0, 5)$



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

Coefficient for variable 2

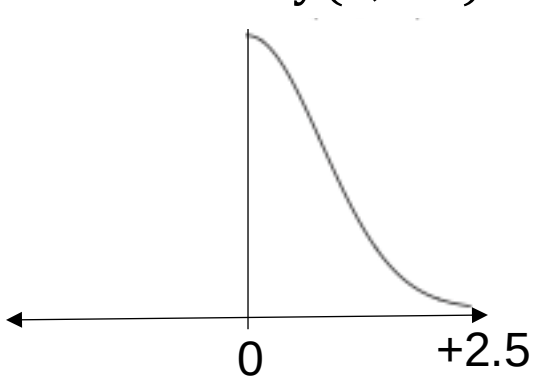
$\beta_2 \sim \text{norm}(0, 5)$



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

Standard deviation

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



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.

Note: β_3 is normal and also centred at 0 but it's distribution or value can vary ± 5

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 (cat)				
Jordanian (referent)				
Syrian	+1.01 (95% CrI: -3.24 to +5.28)	0.675	39,863	< 1.05
Educational attainment (highest) (cat)				
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 (cont)	+0.78 (95% CrI: +0.22 to +1.33)	0.997	52,924	< 1.05
Household income (cat)				
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) (cat)				
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 (cat)				
Yes (referent)				
No	+0.04 (95% CrI: -3.19 to +3.32)	0.513	16,669	< 1.05
Use other alternative water sources (cat)				
No (referent)				
Yes	+1.73 (95% CrI: -1.79 to +5.14)	0.842	15,385	< 1.05
Recycling of Water (cat)				
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.

Any questions?