Bayesian Methods

Awol Seid Ebrie (PhD)

Wits School of Public Health
Johannesburg, South Africa
Email: awol.ebrie@wits.ac.za

08 May 2025

Bayesian Inference for Binary and Count Outcomes

Topics

Session 1. Theoretical Discussions

- Overview of Generalized Linear Models (GLMs) and Exponential Family
- Framework of Bayes' Theorem: for Two Binary Events, for Regression Models
- Binary Logistic Regression: Classical vs Bayesian
- Count Models: Poisson (and Negative Binomial) Regression

Session 2. Practical Exercises

- Estimating Posterior Distributions
- Implementation of Metropolis-Hastings (MH) Algorithm
- No-U-Turn Sampler (NUTS) using `PyMC` (and `Bambi`) Library.
- Beta–Binomial, Logistic, Poisson (and Negative Binomial) Models

Overview of Generalized Linear Models

Generalized Linear Models (GLM)

- The term ”general” linear model (GLM) usually refers to models for a continuous responses.
  - Linear regression,
  - ANOVA
- The term ”generalized” linear model (GLM) refers to a larger class of models.
  - The response variable is assumed to follow an **exponential family** distribution.
- Different authors/books use GLM to mean either ”general” or ”generalized” linear model.
  - So it is best to rely on context to determine which is meant
- We will prefer to use GLM to mean ”generalized” linear model in this course.

Exponential Family

- A probability distribution belongs to the exponential family if it can be expressed as:

$$f(y;\theta,\phi) = \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right\}$$

  where

  - y is the outcome variable.
  - θ is the natural (or canonical) parameter of the distribution
  - ϕ is the dispersion parameter.
  - $a(\phi), b(\theta)$ and $c(y,\phi)$ are known functions.

The Three Main Components of a GLM

1. **Random Component**: It specifies the probability distribution of the response variable.
   - Normal distribution for a continuous outcome Y in the classical regression model.
   - Binomial distribution for a binary outcome Y in the binary logistic regression model.
   - Poisson distribution for a count outcome Y in Poisson and Negative Binomial models.
2. **Systematic Component**: is a function of the covariates (often called linear predictor):

$$\eta_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} = \sum_{j=1}^{p} \beta_j x_{ij} = \boldsymbol{x}_i^T \boldsymbol{\beta} \qquad \text{where } x_{i0} = 1, \forall i.$$

3. **Link Function**: It specifies the link between the random and the systematic components.
   - It indicates how the expected value of the response relates to the linear combination of covariates.
   - For classical regression: $\eta_i = g\{E(Y_i)\} = g(\mu_i) = \mu_i$
   - For logistic regression: $\eta_i = g\{E(Y_i)\} = \log(\frac{\pi_i}{1-\pi_i}) = \text{logit}(\pi_i)$
   - For Poisson regression: $\eta_i = g\{E(Y_i)\} = \log\{E(Y_i)\} = \log(\mu_i)$

Review of Bayes' Theorem

Bayesian Inference

- Bayesian inference is a particular form of statistical inference based on combining probability distributions in order to obtain other probability distributions.
- For this purpose, the Bayes theorem provides us with a general recipe.

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B|A)\,P(A)}{P(B)}$$

Bayes' Theorem for the Relationship Between Two Binary Events

- Bayesian inference is a form of inference based on combining probability distributions in order to obtain other probability distributions using the idea of Baye's Theorem.
- Assume a binary outcome $Y \in \{0,1\}$ and a binary covariate $X \in \{0,1\}$. The Bayes' Theorem:

$$\underbrace{p(X|Y)}_{Posterior} = \frac{\overbrace{p(Y|X)}^{Likelihood}\,\overbrace{p(X)}^{Prior}}{\underbrace{p(Y)}_{Marginal\ Likelihood}}; \;\; p(Y) > 0.$$

  - The (known) probability distribution of the factor is called **prior** distribution.
  - The distribution of the outcome given the factor is called **likelihood** function.
  - The distribution of the outcome is called **marginal likelihood** function.
  - The distribution of the factor given the outcome is called **posterior** distribution.

Bayes' Theorem for the Relationship Between Two Binary Events

- Example:
  - $P(Y = 1) = 0.60$. That is, 60% cured from a disease.
  - $P(X = 1) = 0.5$. 50% of individuals were treated.
  - $P(Y = 1|X = 1) = 0.8$: Higher success rate if treated.
- Then

$$P(X = 1|Y = 1) = \frac{0.8 \times 0.5}{0.6} = 0.667$$

- Of those cured persons, there is a 66.7% chance they were treated.

Two Distinctions to be Noted

- For inference, we are interested in the posterior distribution: $P(X|Y)$.
    - What was likely true (e.g., estimating treatment effectiveness) given the observed data?
    - What is the (posterior) probability that a patient had diabetes, given their test results?
- If we are predicting outcomes, we would use the likelihood: $P(Y|X)$.
    - What will likely happen in the future or in unseen data?
    - Given a new patient's features, what is the chance they will have diabetes?

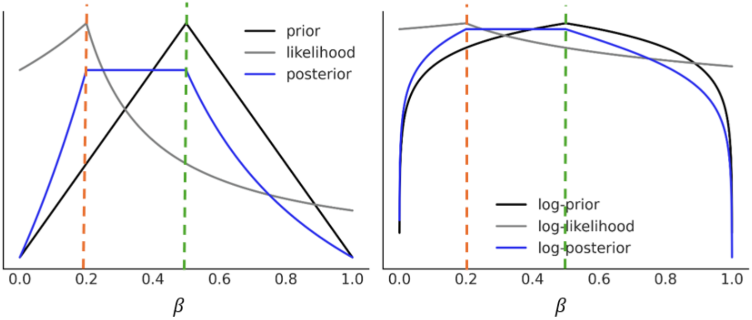
Bayes' Theorem for a Regression Model Parameters

- For a regression model, i.e., $\boldsymbol{y} = g(\boldsymbol{X}, \boldsymbol{\beta})$ where y is the outcome and $\boldsymbol{X}$ is the set of covariates.
- Bayes' Theorem provides a general recipe to estimate the parameter $\boldsymbol{\beta}$ given the data $(\boldsymbol{y}, \boldsymbol{X})$.

$$\underbrace{p(\boldsymbol{\beta}|\boldsymbol{y}, \boldsymbol{X})}_{\textit{Posterior}} = \frac{\overbrace{p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\beta})}^{\textit{Likelihood}} \; \overbrace{p(\boldsymbol{\beta})}^{\textit{Prior}}}{\underbrace{p(\boldsymbol{y}|\boldsymbol{X})}_{\textit{Marginal Likelihood}}}$$

  - The (known) probability distribution of the parameters of the model is $p(\boldsymbol{\beta})$ (i.e., **prior** distribution).
  - The probability distribution of the data given the parameters is $p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\beta})$ (i.e., **likelihood** function).
  - The distribution of the outcome given the covariates is $p(\boldsymbol{y}|\boldsymbol{X})$ (i.e., **marginal likelihood** function).
  - The distribution of the parameters given the data is $p(\boldsymbol{\beta}|\boldsymbol{y}, \boldsymbol{X})$ (i.e., **posterior** distribution).

Prior vs Likelihood vs Posterior

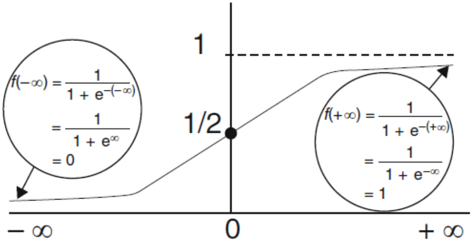


- The hypothetical prior indicates that the value $\beta = 0.5$ is more likely.
  - The plausibility of other of β decreases linearly and symmetrically (black).
- The likelihood of $\beta = 0.2$ shows it best agrees with the data (gray) and the posterior (blue).
  - There is a compromise between the prior and the likelihood.

Bayesian Logistic Regression

Binary Regression

- Recall the logistic function is $\quad f(z) = \dfrac{1}{1+\exp(-z)}; \ -\infty < z < \infty.$



- The figure describes the range of $f(z)$ is between 0 and 1 (i.e., $0 \leq f(z) \leq 1$) for any value of z.
- It is suitable for use as a probability model and let us use $\pi(z)$ to indicate a probability value.

$$\pi(z) = \frac{1}{1+\exp(-z)} = P(Y=1|Z=z); \ -\infty < z < \infty$$

Binary Regression

- In logistic regression, z is expressed as a function (mostly linear) of the explanatory variables.

$$z_i = \beta_0 + \beta_1 x_{i1} + \beta_1 x_{i2} + \cdots + \beta_p x_{ip} = \sum_{j=1}^{p} \beta_j x_{ij} = \boldsymbol{x}_i^T \boldsymbol{\beta} \quad \text{where } x_{i0} = 1, \forall i.$$

- As a result, the logistic probability model is:

$$\pi(\boldsymbol{x}_i) = P(Y_i = 1 | \boldsymbol{x}_i) = \frac{1}{1 + \exp[-(\beta_0 + \beta_1 x_{i1} + \beta_1 x_{i1} + \cdots + \beta_p x_{ip})]}$$

- It can also be written as:

$$\pi(\boldsymbol{x}_i) = P(Y_i = 1 | \boldsymbol{x}_i) = \frac{\exp(\beta_0 + \beta_1 x_{i1} + \beta_1 x_{i2} + \cdots + \beta_p x_{ip})}{1 + \exp(\beta_0 + \beta_1 x_{i1} + \beta_1 x_{i2} + \cdots + \beta_p x_{ip})}.$$

- The relationship between the probability of success and the covariates is not linear.
  - However, it can be linearized by using different transformations of the probability of success.
  - The most common one is called **logit** or **log-odds** transformation.

The logit Transformaion

- An odds is the ratio of the probability of success to the probability of failure.
- Hence, the odds of successes at a particular value $\boldsymbol{x}_i$ of the explanatory variables is

$$\Omega(\boldsymbol{x}_i) = \frac{\pi(\boldsymbol{x}_i)}{1 - \pi(\boldsymbol{x}_i)}.$$

- Thus, the odds of successes is $\Omega(\boldsymbol{x}_i) = e^{\beta_0 + \beta_1 x_{i1} + \beta_1 x_{i2} + \cdots + \beta_p x_{ip}}$.
  - If $\Omega(\boldsymbol{x}_i) = 1$, a success is as likely as a failure at the particular value $\boldsymbol{x}_i$ of the explanatory variables.
  - If $\Omega(\boldsymbol{x}_i) > 1$, a success is more likely to occur than a failure at $\boldsymbol{x}_i$.
  - On the other hand, if $\Omega(\boldsymbol{x}_i) < 1$, a success is less likely than a failure.

The logit Transformaion

- The **logit** of the probability of success is the natural logarithm of the odds of successes.
- It is a linear function of the explanatory variable:

$$\text{logit}\ \pi(\boldsymbol{x}_i) = \log\left[\frac{\pi(\boldsymbol{x}_i)}{1-\pi(\boldsymbol{x}_i)}\right] = \beta_0 + \beta_1 x_{i1} + \beta_1 x_{i2} + \cdots + \beta_p x_{ip}$$

- This is particularly called the *logit* model as it uses the **log-odds** transformation.
- **Note**: From now onwards, let us use the π_i instead of $\pi(\boldsymbol{x}_i)$ for simplicity.

$$\text{logit}\ \pi_i = \log\left[\frac{\pi_i}{1-\pi_i}\right] = \beta_0 + \beta_1 x_{i1} + \beta_1 x_{i2} + \cdots + \beta_p x_{ip}$$

Interpretation of the Parameters

- The sign of each $\beta_j; j = 1, 2, \cdots, p$ determines whether the probability of success is increasing or decreasing as the value of the corresponding explanatory variable increases.
- When the parameter β_j is zero, Y is independent of X_j.
- The slope parameters can be interpreted in terms of odds ratio.
  - From logit $\pi_i = \beta_0 + \beta_1 x_{i1} + \beta_1 x_{i2} + \cdots + \beta_p x_{ip}$, an odds is an exponential function of $\boldsymbol{x}_i$.
  - This provides a basic interpretation for the **magnitude** of the slope parameter β.
  - Thus, the odds ratio is associated with each covariate is given as:

$$\theta_j = \frac{\Omega(x_{i1}, x_{i2}, \cdots, x_{ij} + 1, \cdots, x_{ip})}{\Omega(x_{i1}, x_{i2}, \cdots, x_{ij}, \cdots, x_{ip})} = e^{\beta_j}.$$

  - For every one unit increase in x_{ij}, the odds of success changes by a factor of e^{β_j}.
  - Similarly, for an m units increase in x_{ij}, the corresponding odds ratio becomes $e^{m\beta_j}$.

Estimation of Parameters

- Recall the binary response probability π_i given the values of the explanatory variables $\boldsymbol{x}_i$ is

$$\pi_i = \frac{e^{\beta_0+\beta_1 x_{i1}+\beta_1 x_{i2}+\cdots+\beta_p x_{ip}}}{1+e^{\beta_0+\beta_1 x_{i1}+\beta_1 x_{i2}+\cdots+\beta_p x_{ip}}} = \frac{e^{\sum\limits_{j=0}^{p}\beta_j x_{ij}}}{1+e^{\sum\limits_{j=0}^{p}\beta_j x_{ij}}} = \frac{e^{\boldsymbol{x}_i^T\boldsymbol{\beta}}}{1+e^{\boldsymbol{x}_i^T\boldsymbol{\beta}}} \tag{1}$$

- Equivalently using the logit transformation, it can be written as

$$\log\left[\frac{\pi_i}{1-\pi_i}\right] = \beta_0+\beta_1 x_{i1}+\beta_1 x_{i2}+\cdots+\beta_p x_{ip} = \sum_{j=0}^{p}\beta_j x_{ij} = \boldsymbol{x}_i^T\boldsymbol{\beta}. \tag{2}$$

- The goal in the logistic regression model is to estimate the $p+1$ (unknown) parameters.
- This is done with maximum likelihood estimation (MLE).
  - Entails finding the value of parameters for which the probability of the observed data is maximum.

Estimation of Parameters

- Consider $Y_1, Y_2, \cdots, Y_n$ is an **independent** sample from an **identical** Bernoulli distribution.

$$Y_i \sim Bernoulli(\pi) \qquad \text{where } \pi = P(Y = 1).$$

- The probability mass function (pmf) of Y_i is:

$$P(Y_i = y_i) = p(y_i) = \pi^{y_i}(1-\pi)^{1-y_i}; \quad i = 1, 2, \cdots, n.$$

- The likelihood function is defined as:

$$L(\pi) = p(\boldsymbol{y}|\pi) = \prod_{i=1}^{n} \pi^{y_i}(1-\pi)^{1-y_i}.$$

Estimation of Parameters

- The above likelihood function represents an intercept-only model (or there are no covariates).
- When there are covariates, we need to express the parameter π as a function of the covariates.
- The distribution of each Y_i is no longer identical, i.e.,

$$Y_i \sim Bernoulli(\pi_i) \qquad \text{where } \pi_i = P(Y = 1|\boldsymbol{x}_i).$$

- The probability mass function (pmf) of Y_i becomes

$$P(Y_i = y_i) = p(y_i) = \pi_i^{y_i}(1 - \pi_i)^{1-y_i}; \quad i = 1, 2, \cdots, n.$$

- Note that

$$\pi_i = P(Y_i = 1|\boldsymbol{x}_i) = \frac{1}{1 + e^{-\boldsymbol{x}_i^T\boldsymbol{\beta}}} = \frac{e^{\boldsymbol{x}_i^T\boldsymbol{\beta}}}{1 + e^{\boldsymbol{x}_i^T\boldsymbol{\beta}}}.$$

Estimation of Parameters

- The likelihood function is now expressed as:

$$
\begin{aligned}
L(\boldsymbol{\beta}) &= p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\beta}) \\
&= \prod_{i=1}^{n} \pi_i^{y_i} (1 - \pi_i)^{1-y_i} \\
&= \prod_{i=1}^{n} \left[\frac{1}{1 + e^{-\boldsymbol{x}_i^T \boldsymbol{\beta}}} \right]^{y_i} \left[1 - \frac{1}{1 + e^{-\boldsymbol{x}_i^T \boldsymbol{\beta}}} \right]^{1-y_i} \\
&= \prod_{i=1}^{n} \left[\frac{e^{\boldsymbol{x}_i^T \boldsymbol{\beta}}}{1 + e^{\boldsymbol{x}_i^T \boldsymbol{\beta}}} \right]^{y_i} \left[1 - \frac{e^{\boldsymbol{x}_i^T \boldsymbol{\beta}}}{1 + e^{\boldsymbol{x}_i^T \boldsymbol{\beta}}} \right]^{1-y_i}
\end{aligned}
$$

Bayesian Approach

- Recall Bayes' Theorem: $p(A|B) = \dfrac{p(A)\,p(B|A)}{p(B)}; \quad p(B) > 0.$
- Given
  - The probability distribution of the parameters of the model $p(\boldsymbol{\beta})$.
  - The probability distribution of the data given the parameters $p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\beta})$ (i.e., likelihood function).
- The probability distribution of the parameters given the data $p(\boldsymbol{\beta}|\boldsymbol{y}, \boldsymbol{X})$ is given by:

$$\underbrace{p(\boldsymbol{\beta}|\boldsymbol{y}, \boldsymbol{X})}_{Posterior} = \frac{\overbrace{p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\beta})}^{Likelihood}\ \overbrace{p(\boldsymbol{\beta})}^{Prior}}{\underbrace{p(\boldsymbol{y}|\boldsymbol{X})}_{Marginal Likelihood}}.$$

  - The distribution of the parameters, which will be known, is called **prior** distribution.
  - The distribution of the parameters given the data is called **posterior** distribution.

Bayesian Approach

- The denominator in this formula, i.e., $p(\text{data})$ does not depend on the parameter and is fixed.
- Hence,

$$\begin{aligned} p(\boldsymbol{\beta}|\boldsymbol{y}, \boldsymbol{X}) &\propto p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\beta})\, p(\boldsymbol{\beta}) \\ &\propto L(\boldsymbol{\beta})\, p(\boldsymbol{\beta}) \\ &\propto \text{Likelihood} \times \text{Prior} \end{aligned}$$

- The task in the Bayesian approach is to find the parameters that maximize the probability $p(\boldsymbol{\beta}|\boldsymbol{y}, \boldsymbol{X})$ which is proportional to Likelihood $\times$ Prior.

Bayesian Approach

- Suppose we have a normal prior distribution for all the parameters:

$$\beta_j \sim N(0, \sigma^2), \quad j = 1, 2, \cdots, p \qquad \text{or} \qquad \boldsymbol{\beta} \sim N(\mathbf{0}, \sigma^2 \boldsymbol{I})$$

- That is, the joint pdf of all β_j is:

$$f(\boldsymbol{\beta}) = \prod_{j=1}^{p} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}\beta_j^2} = \left[\frac{1}{\sqrt{2\pi}\sigma}\right]^p e^{-\frac{1}{2\sigma^2}\sum_{j=1}^{p}\beta_j^2} = \left[\frac{1}{\sqrt{2\pi}\sigma}\right]^p e^{-\frac{1}{2\sigma^2}\boldsymbol{\beta}^T\boldsymbol{\beta}}.$$

- The posterior distribution is:

$$p(\boldsymbol{\beta}|\boldsymbol{y}, \boldsymbol{X}) = \prod_{i=1}^{n} \left[\frac{e^{\boldsymbol{x}_i^T\boldsymbol{\beta}}}{1 + e^{\boldsymbol{x}_i^T\boldsymbol{\beta}}}\right]^{y_i} \left[1 - \frac{e^{\boldsymbol{x}_i^T\boldsymbol{\beta}}}{1 + e^{\boldsymbol{x}_i^T\boldsymbol{\beta}}}\right]^{1-y_i} \left[\frac{1}{\sqrt{2\pi}\sigma}\right]^p e^{-\frac{1}{2\sigma^2}\boldsymbol{\beta}^T\boldsymbol{\beta}}$$

Bayesian Approach

- Assume a normal prior with mean μ_j and variance σ_j^2 for each parameter β_j:

$$\beta_j \sim N(\mu_j, \sigma_j^2), \quad j = 1, 2, \cdots, p.$$

That is, $\boldsymbol{\beta} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where $\boldsymbol{\mu} = (\mu_1, \mu_2, \cdots, \mu_p)^T$ and $\boldsymbol{\Sigma} = \text{diag}(\sigma_1^2, \sigma_2^2, \cdots, \sigma_p^2)$.

- That is, the joint pdf of all β_j is:

$$f(\boldsymbol{\beta}) = \frac{1}{\left(\sqrt{2\pi}\right)^p \prod_{j=1}^{p} \sigma_j} e^{-\frac{1}{2}\sum_{j=1}^{p}\left(\frac{\beta_j - \mu_j}{\sigma_j}\right)^2} = \frac{1}{\left(\sqrt{2\pi}\right)^p \sqrt{|\boldsymbol{\Sigma}|}} e^{-\frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{\beta}-\boldsymbol{\mu})}.$$

- It is a (independent) multivariate normal distribution (i.e., with a diagonal covariance matrix).

Bayesian Approach

- Then, the posterior distribution is given by:

$$p(\beta|\boldsymbol{y}, \boldsymbol{X}) = \prod_{i=1}^{n} \left[\frac{e^{\boldsymbol{x}_i^T \boldsymbol{\beta}}}{1 + e^{\boldsymbol{x}_i^T \boldsymbol{\beta}}} \right]^{y_i} \left[1 - \frac{e^{\boldsymbol{x}_i^T \boldsymbol{\beta}}}{1 + e^{\boldsymbol{x}_i^T \boldsymbol{\beta}}} \right]^{1-y_i} \frac{1}{(\sqrt{2\pi})^p \sqrt{|\boldsymbol{\Sigma}|}} e^{-\frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\beta}-\boldsymbol{\mu})}$$

Choice of Priors

- **Expert Knowledge/Domain Context/ Effect Sizes from literature**
  - What range of effects is reasonable or expected for each predictor?
  - If we are modeling a risk factor like smoking in a health model:
    - If the odds ratio is around 2 in prior studies, it means $\beta_{smoking} = \log(2) \approx 0.693$.
    - Then, the prior can be $\beta_{smoking} \sim N(0.693, 0.2^2)$.
  - If a study based on meta-analysis report a treatment effect with mean log-odds = 0.5, SD = 0.3.
    - We can set $\beta_{treatment} \sim N(0.5, 0.3^2)$.
- **Reasonable Odds Ratios**:
  - We believe a predictor likely increases or decreases the odds by at most a factor of 3.
  - The OR range is $[\frac{1}{3}, 3]$ and its log-odds is $[-1.1, 1.1]$
  - We can choose $\beta \sim N(0, 1^2)$ that centers the prior at no effect but allows moderate effect.

Choice of Priors

- **Weakly Informative Defaults**
  - When we do not want the prior to dominate, but still want regularization, it is suggested:

$$\beta \sim \text{Student } t \text{ distribution}(df = 3, \mu = 0, \sigma = 2.5).$$

  - This is a default for logistic regression in `PyMC` (and `Bambi`) library of Python.
  - It is often enough to rule out extreme values unless the data supports them.

Bayesian Inference

- Inferences about $\boldsymbol{\beta}$ are based on the marginal posterior distribution of each parameter β_j.

$$p(\beta_j|\boldsymbol{y},\boldsymbol{X}) \propto \int_{\beta_0} \cdots \int_{\beta_{j-1}} \int_{\beta_{j+1}} \cdots \int_{\beta_p} p(\boldsymbol{\beta}|\boldsymbol{y},\boldsymbol{X})\, d\beta_0 \cdots d\beta_{j-1}\, d\beta_{j+1} \cdots d\beta_p = \int_{\mathcal{R}^p} p(\boldsymbol{\beta}|\boldsymbol{y},\boldsymbol{X}) d\beta_{-j}$$

- This is not analytically tractable to get the marginal distribution by numerical integration:
  - The non-linearity (exponential component) $e^{\boldsymbol{x}_i\boldsymbol{\beta}}$.
  - The high-dimensionality (when p is large).
- We can approximate using the family of **Markov chain Monte Carlo** (MCMC) methods:
  - Metropolis-Hastings (M–H), No-U-Turn Sampler (NUTS), Gibbs Sampling, etc.
- MCMC methods approximate the posterior distribution using simulated samples.

Metropolis-Hastings Algorithm

Let the number of iterations run from $t = 0, 1, 2, \cdots$.

1. Initialize the value of the parameter $\boldsymbol{\beta}$ at $\boldsymbol{\beta}_0$ at $t = 0$.
2. Generate a new value $\boldsymbol{\beta}_{t+1}$ from $\boldsymbol{\beta}_t$ using a (symmetric) proposal distribution $q(\boldsymbol{\beta}_{t+1}|\boldsymbol{\beta}_t)$.
3. Compute the probability of accepting the new value as:

$$p_{accept} = p(\boldsymbol{\beta}_{t+1}|\boldsymbol{\beta}_t) = \min\left[1, \frac{q(\boldsymbol{\beta}_t|\boldsymbol{\beta}_{t+1})\ p(\boldsymbol{\beta}_{t+1})}{q(\boldsymbol{\beta}_{t+1}|\boldsymbol{\beta}_t)\ p(\boldsymbol{\beta}_t)}\right] = min\left[1, \frac{\text{posterior of } \beta_{t+1}}{\text{posterior of } \beta_t}\right]$$

4. Save new value $\boldsymbol{\beta}_{t+1}$ if $p_{accept} > r$ where $r \sim U(0, 1)$. Otherwise, save the old value $\boldsymbol{\beta}_t$.
5. Repeat steps $2 - 4$ until a sufficiently large sample of values has been generated.

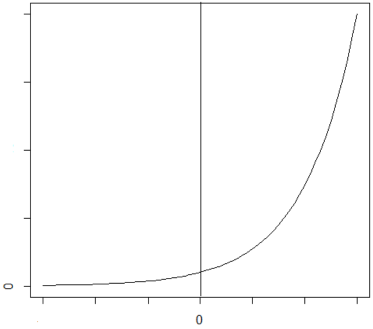
Bayesian Poisson Regression

Poisson Regression

- Count regression models are used for modelling **count (discrete)** response variables.
- Example: the number of hospital admissions, the number of accidents over some period.
- The unit of analysis could be:
  - a person (e.g., number of infections per patient per year),
  - an institution (e.g., number of admissions per hospital per month) or
  - a place (e.g., number of car accidents per city per day).
- As a first pass, such a dependent variable could be analyzed as a continuous outcome.
- However, unlike a continuous variable, there cannot be negative numbers for counts.
- Also, the distribution of counts is often right skewed and does not fit a normal distribution.

Exponential Function

- Count regression models are modeled based on the exponential function.
- The exponential function is $f(z) = \exp(z)$ is nonnegative for any value of z.



- The figure also shows that the range of f is $0 \leq f(z) < \infty$.

Poisson Regression Model

- To obtain the Poisson regression model, z should be expressed as a function (mostly linear function) of the explanatory variables.
- Here, since $f(x_i)$ represents the mean response, let us use the notation $\mu(\boldsymbol{x}_i) = \mu_i$.

$$\mu_i = e^{\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\cdots+\beta_k x_{ip}} \quad \Rightarrow \log \mu_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ip}.$$

- The slope parameters are commonly interpreted in terms of incidence rate ratio (IRR).
  - A one unit increase in x_{ij} has a multiplicative impact of e^{β_j} on the mean response assuming all other covariates constant.
  - If $\beta_j = 0$, then the multiplicative factor is 1, the mean of Y_i does not change as x_{ij} changes.
  - If $\beta_j > 0$, then $e^{\beta_j} > 1$ and the mean of Y_i increases as x_{ij} increases.
  - If $\beta_j < 0$, the mean decreases as x_{ij} increases.

Inference

- Inference for the model follows exactly the same approach as used for logistic regression.
- Like other models, the goal of Poisson regression is to estimate the $p+1$ unknown parameters.
- The method of maximum likelihood estimation is used to estimate the parameters.
- Consider a vector of n Poisson random variables.
- Each response Y_i; $i = 1, 2, \cdots, n$ has an independent Poisson distribution with parameter μ_i:

$$P(Y_i = y_i) = \frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!}$$

where $\mu_i = \mu(\boldsymbol{x}_i) = e^{\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\cdots+\beta_k x_{ip}}$.

Likelihood Function

- The likelihood function of $Y_1, Y_2, \cdots, Y_n$ is:

$$\begin{aligned} L(\boldsymbol{\beta}) &= \prod_{i=1}^{n} \frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!} \\ &= \frac{e^{-\sum_{i=1}^{p}\mu_i}\prod_{i=1}^{n}\mu_i^{y_i}}{\prod_{i=1}^{n} y_i!} \\ &\propto e^{-\sum_{i=1}^{p}\mu_i}\prod_{i=1}^{n}\mu_i^{y_i} \\ &\propto e^{-\sum_{i=1}^{p}e^{x_i\boldsymbol{\beta}}}\prod_{i=1}^{n}(e^{x_i\boldsymbol{\beta}})^{y_i} \end{aligned} \tag{3}$$

Posterior

- Using vague (non-informative, flat) priors for $\boldsymbol{\beta}$, the posterior distribution in Poisson regression is approximately proportional to the likelihood function – $\beta_j \sim U(-\infty, \infty)$:

$$p(\boldsymbol{\beta}|y) \propto e^{-\sum_{i=1}^{n} e^{\boldsymbol{x}_i \boldsymbol{\beta}}} \prod_{i=1}^{n} (e^{\boldsymbol{x}_i \boldsymbol{\beta}})^{y_i}$$

- Inferences about $\boldsymbol{\beta}$ are based on the marginal posterior distributions.
- We can obtain marginal distributions using Markov chain Monte Carlo (MCMC) simulation.

Negative Binomial Regression

Negative Binomial Regression

- Often count data vary more than the expected (it is called **over-dispersion**).
- But, over-dispersion is not an issue in ordinary regression models assuming normally distributed response, because the normal distribution has a separate parameter.
- In the presence of over-dispersion, a negative binomial model is should be applied.
- But a negative binomial model has an additional parameter called a *dispersion parameter*.
- That is, because, the negative binomial distribution has mean $E(Y) = \mu$ and variance $\text{Var}(Y) = \mu + \psi\mu^2$ where $\psi > 0$.
- The index ψ is a dispersion parameter.
- As $\psi \approx 0$, $Var(Y)$ goes to μ and the NB distribution converges to the Poisson distribution.
- The farther ψ falls above 0, the greater the over-dispersion relative to Poisson variability.

Negative Binomial Regression

- Let us assume $y_1, y_2, \cdots, y_n$ are distributed according to the negative binomial distribution.
- That is, $y_i \sim NB(p_i, r)$
- The likelihood function:

$$p(y_1, y_2, \cdots, y_n | p_i, r) = \prod_{i=1}^{n} \frac{y_i + r - 1}{y_i(r-1)} p_i^r (1-p_i)^{y_i}$$

where $p_i = \frac{r}{r+\mu_i}$ and $\log(\mu_i) = \beta_0 + \boldsymbol{\beta x}_i$

- The parameter r quantifies the amount of extra Poisson variation and we could assume a gamma prior distribution for it.
- For the coefficients, we could use uniform or non-informative normal prior distributions.

Practical Session

Practical Session

- Software: Python
- Download Notebook.

Thank You!