Bayesian Logistic Regression

태그	ESC
날짜	@2024년 2월 3일

INDEX

1. Logistic Regression, GLM (Generalized Linear Model)

Logistic Regression :

In statistics, the **logistic model** (or **logit model**) is a statistical model that models the log-odds of an event as a linear combination of one or more independent variables. In regression analysis, **logistic regression** (or **logit regression**) is estimating the parameters of a logistic model (the coefficients in the linear combination).

The logistic regression model itself simply models probability of output in terms of input and does not perform statistical classification (it is not a classifier), though it can be used to make a classifier, for instance by choosing a cutoff value and classifying inputs with probability greater than the cutoff as one class, below the cutoff as the other; this is a common way to make a binary classifier.

Analogous linear models for binary variables with a different sigmoid function instead of the logistic function (to convert the linear combination to a probability) can also be used, most notably the probit model.

- From Wikipedia : Logistic Regression

Odds

$$\frac{p}{1-p}$$

- 일어나지 않을 확률 ($1-p$) 대비 일어날 확률 (p)
- 이길 확률이 0.5보다 높으면 → $1 < Odds < \infty$
- 이길 확률이 0.5보다 작으면 → $0 < Odds < 1$

We want to estimate it as a linear combination.

- range : $[0, \infty)$, not $(-\infty, \infty)$
- asymmetry

logit ; Log-Odds

Logistic Regression : models the Log-Odds of an event as a linear combination.

$$log(\frac{p}{1-p}) = \beta_0 + X_1\beta_1 + ... + X_{p-1}\beta_{p-1}$$

range : $(-\infty, \infty)$

- 이길 확률이 0.5보다 높으면 → $0 < logOdds < \infty$
- 이길 확률이 0.5보다 작으면 → $-\infty < logOdds < 0$

$g(p) = log(\frac{p}{1-p}) =$ ***logit function***

$g^{-1}(x) = \frac{e^X}{1+e^X} = \sigma(x) =$ ***sigmoid function***

Designated to model probability. Not to perform classification.

- But also can be used as a classifier.
    - For instance by choosing a cutoff value and classifying inputs with probability greater than the cutoff as one class, below the cutoff as the other; this is a common way to make a binary classifier.
    - Analogous linear models for binary variables with a different sigmoid function instead of the logistic function (to convert the linear combination to a probability) can also be used, most notably the probit model.

Interpretation as a generalized linear model

$logit(E[\mathbf{Y}|\mathbf{X}]) = log(\frac{p}{1-p}) = \beta\mathbf{X}$

$\Rightarrow\ E[\mathbf{Y}|\mathbf{X}] = \sigma(\beta\mathbf{X})$

GLM ; Generalized Linear Model

In a generalized linear model (GLM), each outcome **Y** of the dependent variables is assumed to be generated from a particular distribution in an **exponential family**, a large class of probability distributions that includes the normal, binomial, Poisson and gamma distributions, among others. The conditional mean ***μ*** of the distribution depends on the independent variables **X** through:

https://en.wikipedia.org/wiki/Generalized_linear_model

$$E[\mathbf{Y}|\mathbf{X}] = \mu = g^{-1}(\mathbf{X}\beta)$$

Linear Regression :

$$E[\mathbf{Y}|\mathbf{X}] = \mathbf{X}\beta$$

Generalized Linear Model :

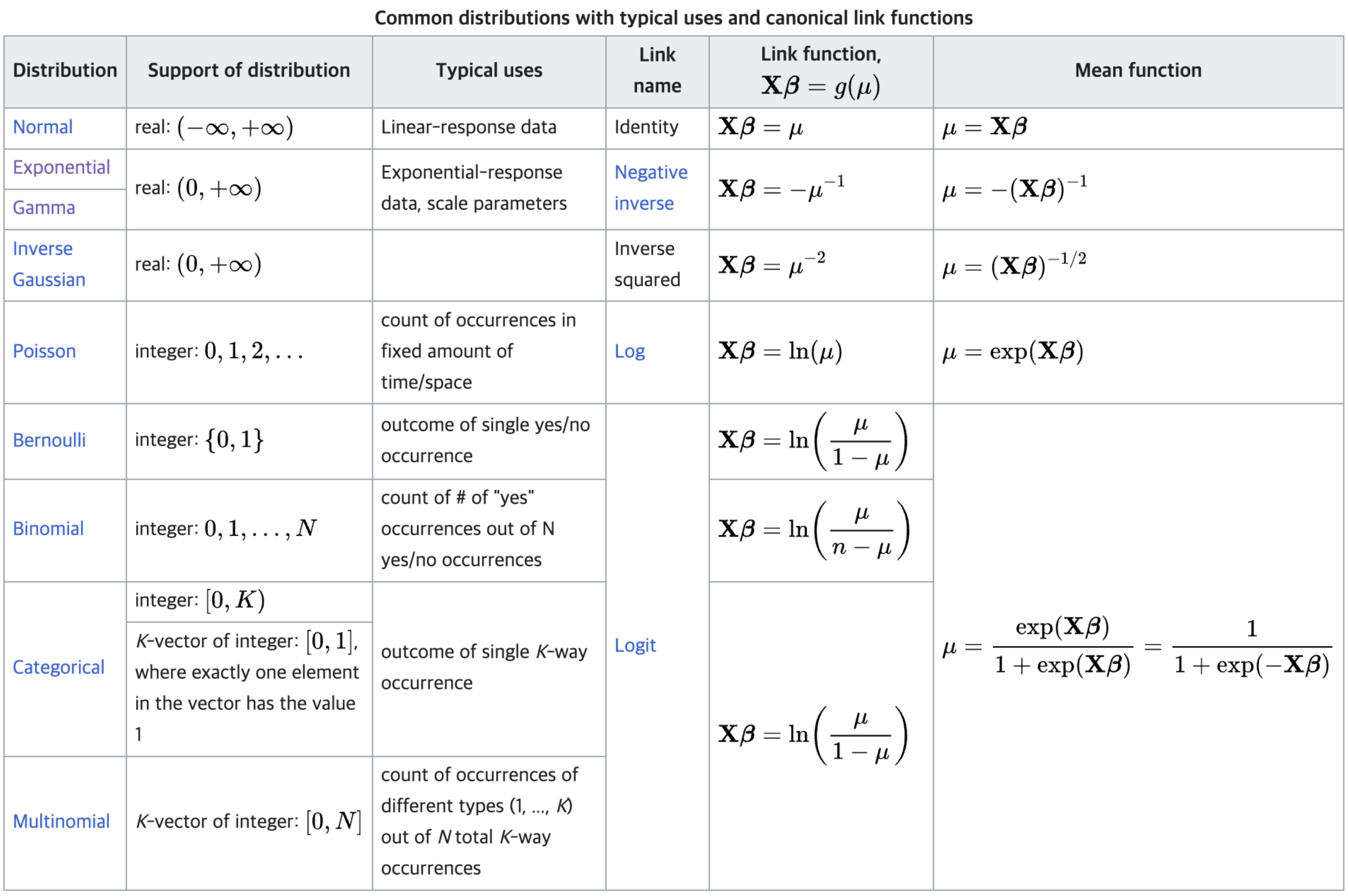
Use the link function g ; Mapping the parameters to $\mathbf{R}$

$$E[\mathbf{Y}|\mathbf{X}] = \mu(\mathbf{X}) = g^{-1}(\mathbf{X}\beta)$$

$$g(E[\mathbf{Y}|\mathbf{X}]) = \mathbf{X}\beta$$

Main Purpose : Estimate & Inference β

Link function



Common distributions with typical uses and canonical link functions

Distribution	Support of distribution	Typical uses	Link name	Link function, $\mathbf{X}\boldsymbol{\beta}=g(\mu)$	Mean function
Normal	real: $(-\infty,+\infty)$	Linear-response data	Identity	$\mathbf{X}\boldsymbol{\beta}=\mu$	$\mu=\mathbf{X}\boldsymbol{\beta}$
Exponential	real: $(0,+\infty)$	Exponential-response data, scale parameters	Negative inverse	$\mathbf{X}\boldsymbol{\beta}=-\mu^{-1}$	$\mu=-(\mathbf{X}\boldsymbol{\beta})^{-1}$
Gamma					
Inverse Gaussian	real: $(0,+\infty)$		Inverse squared	$\mathbf{X}\boldsymbol{\beta}=\mu^{-2}$	$\mu=(\mathbf{X}\boldsymbol{\beta})^{-1/2}$
Poisson	integer: $0,1,2,\ldots$	count of occurrences in fixed amount of time/space	Log	$\mathbf{X}\boldsymbol{\beta}=\ln(\mu)$	$\mu=\exp(\mathbf{X}\boldsymbol{\beta})$
Bernoulli	integer: $\{0,1\}$	outcome of single yes/no occurrence	Logit	$\mathbf{X}\boldsymbol{\beta}=\ln\left(\frac{\mu}{1-\mu}\right)$	$\mu=\frac{\exp(\mathbf{X}\boldsymbol{\beta})}{1+\exp(\mathbf{X}\boldsymbol{\beta})}=\frac{1}{1+\exp(-\mathbf{X}\boldsymbol{\beta})}$
Binomial	integer: $0,1,\ldots,N$	count of # of "yes" occurrences out of N yes/no occurrences		$\mathbf{X}\boldsymbol{\beta}=\ln\left(\frac{\mu}{n-\mu}\right)$	
Categorical	integer: $[0,K)$	outcome of single *K*-way occurrence		$\mathbf{X}\boldsymbol{\beta}=\ln\left(\frac{\mu}{1-\mu}\right)$	
	K-vector of integer: $[0,1]$, where exactly one element in the vector has the value 1				
Multinomial	*K*-vector of integer: $[0,N]$	count of occurrences of different types (1, ..., *K*) out of *N* total *K*-way occurrences			

https://en.wikipedia.org/wiki/Generalized_linear_model

How to fit?

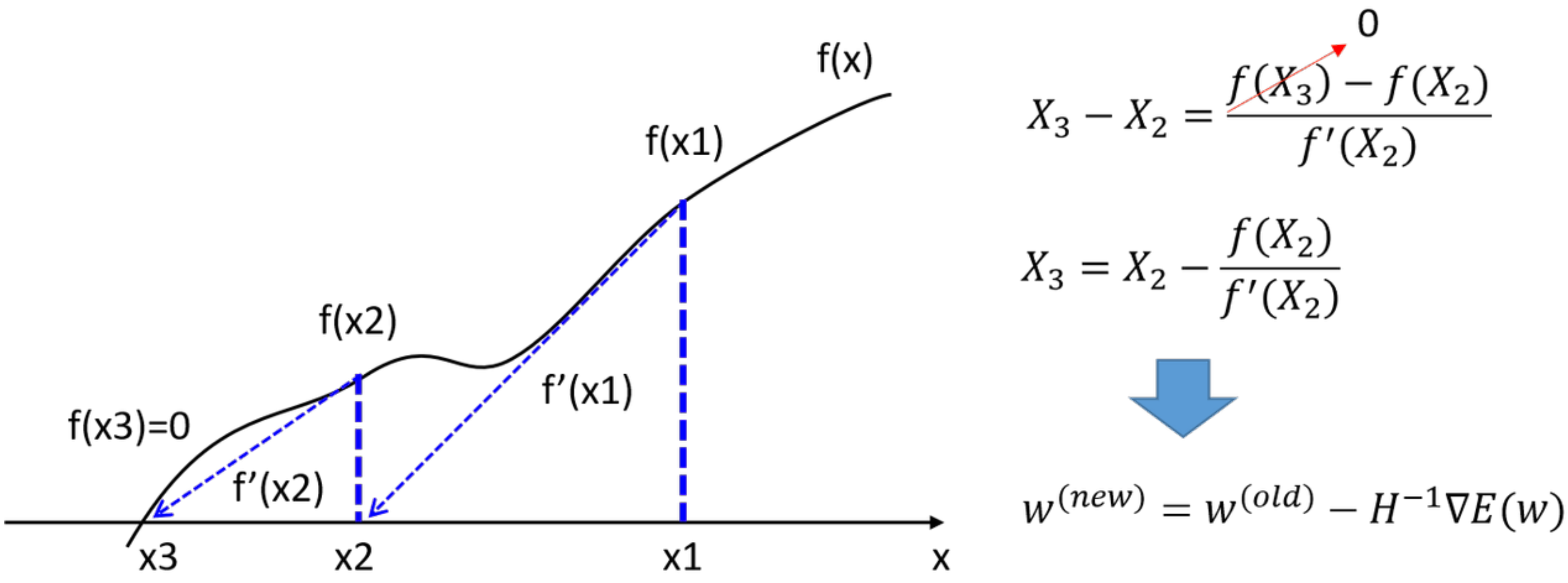
Frequentist Approach

: estimates the unknown parameter using the
Maximum Likelihood Estimate (*MLE*)

No closed form → Iterative Method. because of **NON-LINEARITY**

→ Any method could be done, but we know some nice properties.

1. ***exponential family***
   a. Newton-Raphson Scheme


f(x)
f(x1)
f(x2)
f'(x1)
f(x3)=0
f'(x2)
x3
x2
x1
x
0
$X_3 - X_2 = \frac{\cancel{f(X_3)} - f(X_2)}{f'(X_2)}$
$X_3 = X_2 - \frac{f(X_2)}{f'(X_2)}$
$w^{(new)} = w^{(old)} - H^{-1}\nabla E(w)$

https://big-dream-world.tistory.com/33

- IRLS : Iterative Re-weighted Least Squares
  - *Iterative Reweighted Least Squares*

    - L-BFGS, etc..

  1. Fisher-Scoring
     - https://zephyrus1111.tistory.com/68
  - Nice Properties?
    - If $\mathbf{X}$ is full-rank, its Hessian matrix is Negative Definite. So, log-likelihood becomes Strictly Concave. Thus $\beta^{(k)} \to \beta^{MLE}$ as $k \to \infty$. Local Maximum Problem 발생X
      - That’s why we use the Newton-Raphson Method.
    - Computational Cost를 낮추는 등, 다양한 이점이 더 있는 것으로 알고 있음.

2. ***non exponential family***

   a. Probit Regression 등

   ※ Note : Probit Regression is **not** designated as a solution to fit models outside the exponential family.

▼ See also ; by ChatGPT

Let's clarify how GLMs utilize the properties of the exponential family and how we approach situations when the response variable doesn't neatly fit into this framework.

GLMs and the Exponential Family

GLMs are particularly powerful when the response variable's distribution is a member of the exponential family. This family includes distributions like the normal, binomial, Poisson, and gamma distributions, each of which has properties that are advantageous for modeling:

1. **Canonical Link Function**: Each member of the exponential family has a natural or canonical link function that relates the mean of the distribution to the linear predictors. This canonical link function simplifies the mathematics and computations involved in parameter estimation, making models more interpretable and efficient to fit.
2. **IRLS and Fisher Scoring**: The Iteratively Reweighted Least Squares (IRLS) method, which is a specific form of the Newton-Raphson algorithm, and the closely related Fisher Scoring method, are often used to estimate the parameters of a GLM. These methods exploit the nice properties of the exponential family, such as the relationship between the mean and variance and the form of the log-likelihood function, to facilitate convergence to the optimal parameters. These methods are particularly effective due to the exponential family's mathematical properties, allowing for efficient computation of updates to the parameter estimates.

Non-Exponential Family Situations

When the response variable's distribution does not fall within the exponential family, or when the data exhibit characteristics like overdispersion that are not well captured by the standard exponential family models, alternative approaches are needed:

1. **Generalized Estimating Equations (GEEs)**: For correlated or clustered data that might not fit the standard GLM assumptions, GEEs provide a way to estimate the parameters by focusing on the mean structure without fully specifying the likelihood.
2. **Quasi-likelihood and Generalized Linear Mixed Models (GLMMs)**: These extend GLMs to handle overdispersion and correlated data by introducing random effects and using a quasi-likelihood approach instead of a full likelihood, which does not require the full specification of the distribution of the response variable.
3. **Link Function Flexibility**: Even when the response variable's distribution is not in the exponential family, GLMs can still be applied by carefully choosing an appropriate link function. For example, **probit regression** uses the probit link function suitable for binary data modeled with a latent variable approach, where the latent variable follows a normal distribution.
4. **Non-Canonical Link Functions**: In some cases, a non-canonical link function may be more appropriate due to the nature of the data or interpretability requirements. While this might not leverage all the advantageous properties of the exponential family, it can still provide a meaningful model for the data.

Summary

GLMs leverage the properties of the exponential family through canonical link functions and efficient parameter estimation methods like IRLS and Fisher Scoring. When dealing with data or distributions that do not fit neatly into the exponential family framework, alternative approaches such as using different link functions, quasi-likelihood methods, or extending the model to include random effects or account for overdispersion, can be employed. The choice of method depends on the nature of the data, the goals of the analysis, and the specific challenges posed by the data distribution and structure.

2. Bayesian GLM using Sampling

Bayesian Approach

: estimates the posterior distribution of the unknown random parameters

→ **Let’s estmate the posterior distribution.**

But, how?

Method 1 : Sampling

Given a dataset $(\mathbf{X}_i, y_i)_{i=1}^n$, select the likelihood and the prior distribution.

Our goal is to estimate the posterior distribution.

We have the kernel of the posterior distribution $f(\theta|\mathbf{X}, \mathbf{Y}) \propto L(\mathbf{Y}|\theta, \mathbf{X})f(\theta)$,

with the dataset $(\mathbf{X}_i, y_i)_{i=1}^n$. $\theta = [\ \beta \ \ \sigma^2\ ]^T$

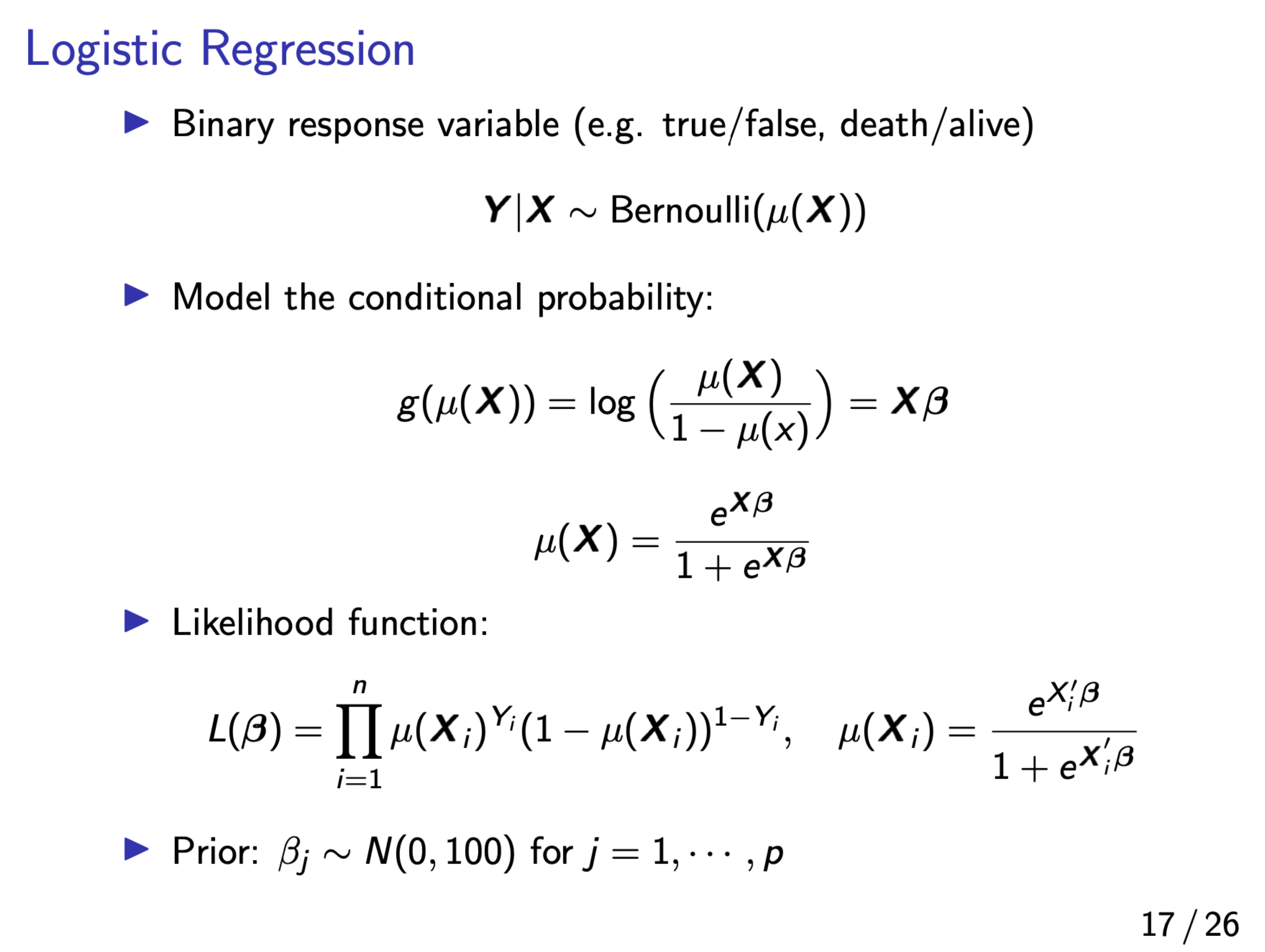
Using MCMC Sampling, we can draw samples of β, σ^2 from the posterior distribution.

Estimate the posterior distribution using these samples.

Steps to Selecting a Bayesian GLM

1. Identify the support of the response distribution
2. Select the likelihood by picking a parametric family of distributions with this support
3. Choose a link function $g(\cdot)$ that transforms the range of parameters to the whole real line
4. Specify a linear model on the transformed parameters
5. Select priors for the regression coefficients

Exercise : Logistic Regression



Logistic Regression

- Binary response variable (e.g. true/false, death/alive)

$$\boldsymbol{Y} | \boldsymbol{X} \sim \text{Bernoulli}(\mu(\boldsymbol{X}))$$

- Model the conditional probability:

$$g(\mu(\boldsymbol{X})) = \log\left(\frac{\mu(\boldsymbol{X})}{1-\mu(x)}\right) = \boldsymbol{X}\boldsymbol{\beta}$$

$$\mu(\boldsymbol{X}) = \frac{e^{\boldsymbol{X}\boldsymbol{\beta}}}{1+e^{\boldsymbol{X}\boldsymbol{\beta}}}$$

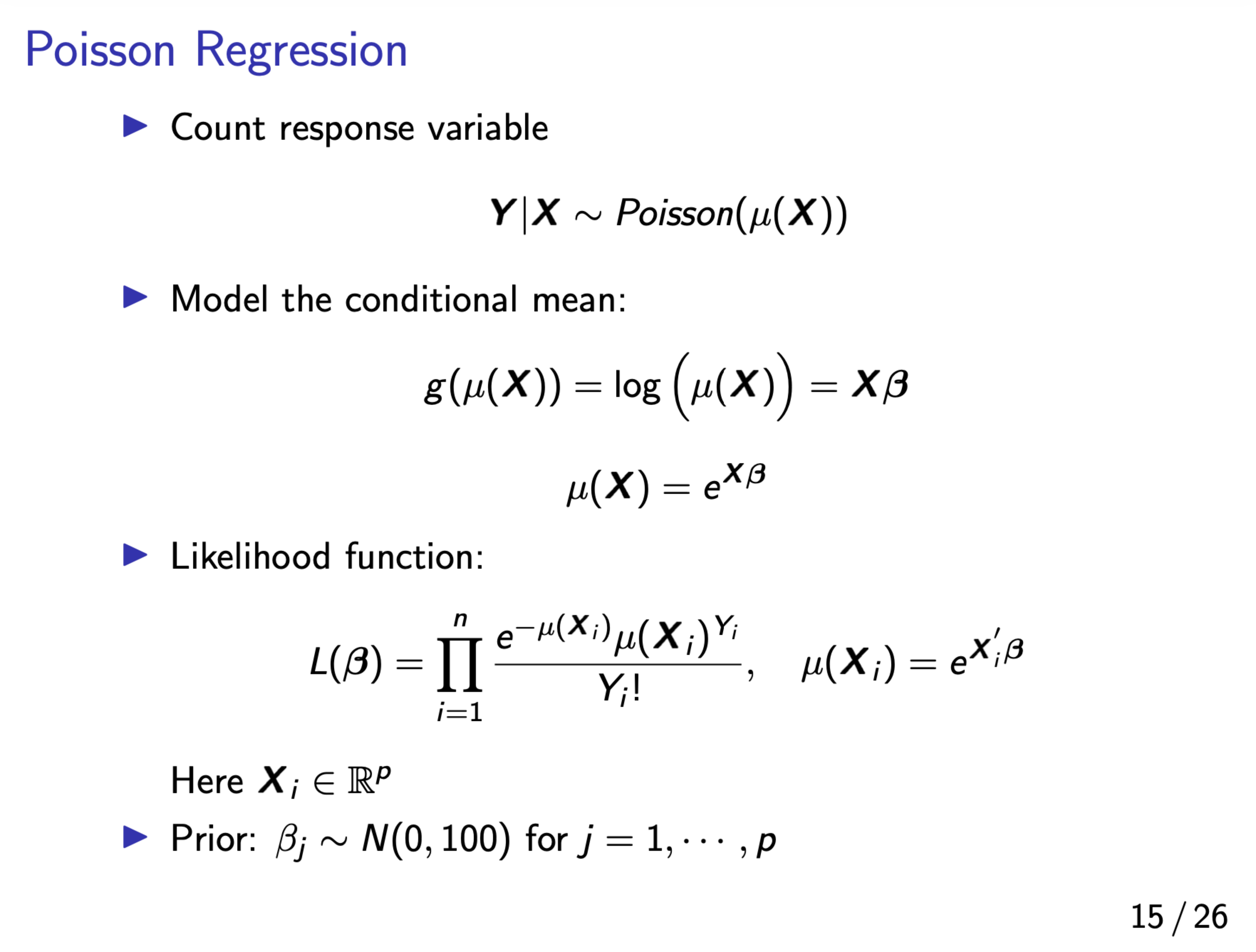
- Likelihood function:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \mu(\boldsymbol{X}_i)^{Y_i}(1-\mu(\boldsymbol{X}_i))^{1-Y_i}, \quad \mu(\boldsymbol{X}_i) = \frac{e^{X_i'\boldsymbol{\beta}}}{1+e^{\boldsymbol{X}_i'\boldsymbol{\beta}}}$$

- Prior: $\beta_j \sim N(0, 100)$ for $j = 1, \cdots, p$

17 / 26

Exercise : Poisson Regression



Poisson Regression

- Count response variable

$$Y|X \sim Poisson(\mu(X))$$

- Model the conditional mean:

$$g(\mu(\boldsymbol{X})) = \log\left(\mu(\boldsymbol{X})\right) = \boldsymbol{X}\boldsymbol{\beta}$$

$$\mu(\boldsymbol{X}) = e^{\boldsymbol{X}\boldsymbol{\beta}}$$

- Likelihood function:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \frac{e^{-\mu(\boldsymbol{X}_i)}\mu(\boldsymbol{X}_i)^{Y_i}}{Y_i!}, \quad \mu(\boldsymbol{X}_i) = e^{\boldsymbol{X}_i'\boldsymbol{\beta}}$$

Here $\boldsymbol{X}_i \in \mathbb{R}^p$

- Prior: $\beta_j \sim N(0, 100)$ for $j = 1, \cdots, p$

15 / 26

Exercise : Categorical(Multinomial) Regression

Do it yourself!

3. Bayesian GLM without using Sampling

Problem ; Too much computational cost of using the Sampling Method.

→ ***Let's Approximate it***

Method 2 : Approximate the posterior distribution

- ex) Laplace Approximation

Comments by ChatGPT

QUESTION : **Sampling 기법 대신 Laplace approximation을 쓰는 이유?**

In Bayesian logistic regression, why is the Laplace approximation used? I typically fit the Bayesian logistic regression model using MCMC sampling, which I find to generalize quite well. However, why is the Laplace approximation often preferred over the sampling method?

ANSWER :

The Laplace approximation is often used in Bayesian logistic regression for a few key reasons, even though sampling methods like Markov Chain Monte Carlo (MCMC) are widely used and can provide more accurate results. Here's why the

Laplace approximation can be preferred in certain contexts:

1. **Computational Efficiency**: The Laplace approximation tends to be computationally more efficient than MCMC, especially for models that are not too complex or when computational resources are limited. The approximation converts the problem into one of optimization rather than sampling, which can often be solved more quickly.
2. **Analytical Convenience**: The Laplace approximation provides an analytical form for the posterior distribution, approximating it with a Gaussian centered at the mode of the posterior. This can make subsequent calculations, like predictive distributions or model comparisons, more straightforward and less computationally intensive than using samples from MCMC.
3. **Large Data Scenarios**: In situations where the dataset is large, MCMC methods can become computationally expensive and slow, as they require many iterations to converge and each iteration involves evaluating the likelihood over the entire dataset. The Laplace approximation can offer a quicker alternative, although it might be less accurate.
4. **Interpretability**: The Gaussian approximation of the posterior in the Laplace method can make the results more interpretable in certain contexts, providing a clear mean and covariance structure that describes the uncertainty around the parameter estimates.
5. **Good Approximation in Certain Cases**: For many models, especially those where the posterior is relatively well-behaved and unimodal (like in some logistic regression problems), the Laplace approximation can provide a sufficiently good approximation to the true posterior, capturing the main aspects of parameter uncertainty.
6. **Integration with Optimization**: Since Bayesian logistic regression often involves finding the mode of the posterior (the maximum a posteriori estimate), those already performing this optimization might find it convenient to extend their analysis with a Laplace approximation to quantify uncertainty around these estimates.

However, it's important to note that the Laplace approximation has limitations, particularly in cases where the posterior is not well-approximated by a Gaussian distribution (e.g., multimodal, highly skewed, or heavy-tailed distributions). In such cases, sampling methods like MCMC might be preferred despite their higher computational cost, as they can capture the full complexity of the posterior distribution more accurately. The choice between using the Laplace approximation and sampling methods depends on the specific context, including the complexity of the model, the size of the dataset, computational resources, and the desired accuracy of the posterior approximation.

QUESTION : Laplace Approximation 말고도 Bayesian model fitting에 쓰이는 방법은?

Got it. Are there any popular methods for fitting Bayesian GLMs, or is the Laplace approximation overwhelmingly preferred?

ANSWER :

In fitting Bayesian Generalized Linear Models (GLMs), a variety of methods are employed, depending on the specific requirements of the problem, such as the complexity of the model, the size of the dataset, and the need for computational efficiency. While the Laplace approximation is a popular choice for its simplicity and computational speed, it is not the only method used, nor is it overwhelmingly popular across all use cases. Here are some of the most commonly used methods:

1. **Markov Chain Monte Carlo (MCMC)**: MCMC is a cornerstone method in Bayesian statistics, widely used for fitting complex models where exact inference is not feasible. Techniques such as Gibbs sampling, Metropolis-Hastings, and more advanced variants like Hamiltonian Monte Carlo (HMC) and the No-U-Turn Sampler (NUTS) are popular for their ability to sample from complex posterior distributions without making strong assumptions about their shape.
2. **Variational Inference (VI)**: VI is a family of techniques that turn the problem of inference into an optimization problem. It is generally faster than MCMC and is particularly useful for large datasets or models where the speed of inference is crucial. VI approximates the posterior with a simpler, parameterized distribution, trading off some accuracy for speed and scalability.
3. **Integrated Nested Laplace Approximation (INLA)**: Specifically designed for latent Gaussian models, which many GLMs fall under, INLA provides a deterministic approach to Bayesian inference that is often faster than MCMC while still providing accurate results. It is particularly popular in spatial and temporal modeling.
4. **Expectation Propagation (EP)**: EP is another approximate inference technique that iterates over local approximations to refine a global approximation of the posterior. While not as widely used as MCMC or VI, EP can be advantageous in specific models, offering a good balance between speed and accuracy.
5. **Hamiltonian Monte Carlo (HMC) and No-U-Turn Sampler (NUTS)**: As extensions of classical MCMC, HMC and NUTS leverage gradient information to make more efficient proposals, which can lead to faster convergence and better

exploration of the posterior, especially in high-dimensional spaces.

The choice among these methods depends on the trade-offs one is willing to make between computational efficiency, ease of implementation, and the accuracy of the posterior approximations. MCMC, despite being computationally intensive, is often preferred for its generality and robustness, particularly in complex models or when the utmost accuracy is required. Variational Inference and INLA offer faster alternatives that are useful in large-scale applications or when rapid iterations are needed. The Laplace approximation, due to its simplicity and speed, remains a useful tool, especially for simpler models or as a component of a more complex Bayesian workflow.