Bayesian Statistics

Meng-Yun Lin

mylin@bu.edu

Boston University School of Public Health
Department of Health Policy & Management

This paper was published in fulfillment of the requirements for PM931 Directed Study in Health Policy and Management under Professor Cindy Christiansen's (cindylc@bu.edu) direction. Michal Horny, Jake Morgan, Marina Soley Bori, and Kyung Min Lee provided helpful reviews and comments.

Table of Contents

Executive Summary 1

1. Difference between Frequentist and Bayesian 2

2. Basic Concepts 3

3. Bayesian Approach of Estimation and Hypothesis Testing 4

3.1. Bayesian Inference 4

3.2. Bayesian Prediction 5

3.3. Bayesian Network 6

3.4. Software 6

4. Bayesian Modeling 7

4.1. Prior Probability Distribution (Prior) 7

4.2. Posterior Probability Distribution (Posterior) 10

4.3. Bayes Estimation 10

4.4. Non-informative Bayes Models 12

4.5. Informative Bayes Models 14

5. Examples: Fitting Bayesian Logistic Regression in SAS 16

6. Useful Resources 21

Executive Summary

This report is a brief introduction of Bayesian statistics. The first section describes the basic concepts of Bayesian approach and how they are applied to statistical estimation and hypothesis testing. The next section presents the statistical modeling using Bayesian approach. It first explains the main components of Bayes model including prior, likelihood function, and posterior. Then, it introduces informative and non-informative Bayes models. The last section provides an example of fitting Bayesian logistic regression in SAS. It illustrates how to program Bayes model and how to check model convergence.

Keywords: Bayesian, Prior, Posterior, Informative Bayes Model, Non-informative Bayes Model.

1. Difference between Frequentist and Bayesian

In recent years, Bayesian approach has been widely applied to clinical trials, research in education and psychology, and decision analyses. However, some statisticians still consider it as an interesting alternative to the classical theory based on relative frequency. These frequentists argue that the introduction of prior distributions violates the objective view point of conventional statistics. Interestingly, the feature of applying priors is also the reason why Bayesian approach is superior to frequentist. The following table briefly summarizes the differences between frequentist and Bayesian approaches. Then, I simply list the cons and pros of Bayesian statistics and suggest situations to which Bayesian statistics is more applicable.

	Frequentist	Bayesian
parameter of model	• fixed, unknown constants • can NOT make probabilistic statements about the parameters	• random variables (parameters can't be determined exactly, uncertainty is expressed in probability statements or distributions) • can make probability statements about the parameters
probability	objective, relative frequency	subjective, degree of belief
main outcomes	point estimates with standard error	posterior distribution
estimate/inference	use data to best estimate unknown parameters	• pinpoint a value of parameter space as well as possible by using data to update belief • all inference follow posterior • use simulation method: generate samples from the posterior and use them to estimate the quantities of interest
interval estimate	Confidence Interval: a claim that the region covers the true parameter, reflecting uncertainty in sampling procedure. e.g: 95%CI=(a, b) implies the interval (a, b) covers the true parameter among 95% of the experiments	Credible Interval: a claim that the true parameter is inside the region with measurable probability. One can make a direct probability statement about parameters. e.g: 95%CI=(a, b) implies the chance that the true parameter falls in (a, b) is 95%.

Pros of Bayesian Statistics:

1) combine prior info with data
2) provide exact estimate w/o reliance on large sample size
3) can directly estimate any functions of parameters or any quantities of interest
4) obey the likelihood principle
5) provide interpretable answers (credible intervals have more intuitive meanings)
6) can be applied to a wide range of models, e.g. hierarchical models, missing data...

Cons of Bayesian Statistics:

1) prior selection
2) posterior may be heavily influenced by the priors (informative prior+ small data size)
3) high computational cost
4) simulation provide slightly different answers
5) no guarantee of Markov Chain Mote Carlo (MCMC) convergence

In brief, Bayesian statistics may be preferable than frequentist statistics when research wants to combines knowledge modeling (info from expert, or pre-existing info) with knowledge discovery (data, evidence) to help with decision support (analytics, simulation, diagnosis, and optimization) and risk management.

2. Basic Concepts

Bayesian probability is the foundation of Bayesian statistics. It interprets probability as an abstract concept—a quantity that one assign theoretically by specifying some prior probabilities—for the purpose of representing a state of knowledge. It is then updated in the light of new and relevant data. In Bayesian statistics, a probability can be assigned to a hypothesis that can be any quantities between 0 and 1 if the truth value is uncertain. Broadly speaking, there are two views on Bayesian probability that interpret the concept of probability in different ways. For objectivists, probability objectively measures the plausibility of propositions. The probability of a proposition corresponds to a reasonable belief. For subjectivists, probability corresponds to a personal belief. Rationality and coherence constrain the probabilities one may have but allow for substantial variation within those constraints. The objective and subjective variants of Bayesian probability differ mainly in their interpretation and construction of the prior probability[1].

Based on Bayesian probability, *Bayes' theorem*[2] links the degree of belief in a proposition before and after accounting for evidence by giving the relationship between the probabilities of A and B, p(H) and p(D), and the conditional probabilities of H given D and D given H, p(H|D) and p(D|H). The algebraic formula is given by $p(H|D) = \frac{p(D|H)*p(H)}{p(D)}$, where H and D stand for hypothesis and data (evidence) respectively; p(H) is

[1] Jaynes, E. T. Bayesian Methods: General Background, 1986.

the initial degree of belief in H (*prior*); p(H|D) is the degree of belief in H having accounted for D (*posterior*); and p(D|H)/p(D) represents the support D provides for H.

Further, *Bayes' rule*[2] relates the odds of event A1 to event A2 before and after conditioning on event B. The relationship is expressed in terms of the *Bayes factor*, Ƒ, which represents the impact of the conditioning on the odds. On the ground of Bayesian probability, Bayes' rule relates the odds on probability models A1 and A2 before and after evidence B is observed. In this case, Ƒ represents the impact of the evidence on the odds. For single evidence, the association is given by O(A1: A2 |B) = O(A1: A2)*Ƒ(A1: A2 |B), where O(A1: A2) is called *prior odds*; O(A1: A2 |B) is called *posterior odds.* In brief, Bayes' rule is preferred to Bayes' theorem when the relative probability, odds, of two events matters but the individual probabilities.

3. Bayesian Approach of Estimation and Hypothesis Testing

3.1. Bayesian Inference

Bayesian inference is usually presented as a method for determining how scientific belief should be modified by data. Bayesian methods have a data-based core, which can be used as a calculus of evidence. This core is the Bayes factor, which in its simplest form is also called a *likelihood ratio*. The minimum Bayes factor is objective and can be used in lieu of the P value as a measure of the evidential strength. Unlike P values, Bayes factors have a sound theoretical foundation and an interpretation that allows their use in both inference and decision making. Most important, Bayes factors require the addition of background knowledge to be transformed into inferences' probabilities that a given conclusion is right or wrong. They make the distinction clear between experimental evidence and inferential conclusions while providing a framework in which to combine prior with current evidence[3].

3.1.1. Formal Bayesian Inference

Formal Bayesian inference derives the posterior probability as a consequence of two antecedents, a prior probability and a likelihood function, derived from a probability model for the data to be observed. The posterior probability is computed according to Bayes' theorem:

$$p(H|E) = \frac{p(E|H)}{p(E)} * p(H)$$

H: *hypothesis* whose probability may be affected by data (evidence). Usually, there are competing hypotheses from which one chooses the most probable

E: *evidence* (data)

[2] Lee, Peter M. Bayesian Statistics: An Introduction. John Wiley & Sons, 2012

[3] Goodman, Steven N. "Toward Evidence-Based Medical Statistics. 2: The Bayes Factor." Annals of Internal Medicine 130, no. 12 (June 15, 1999): 1005–1013

p(H): *prior probability*, the probability of H before E is observed, indicating one's preconceived beliefs about how likely different hypotheses are

p(H|E): *posterior probability*, the probability of H given E (after E is observed)

p(E|H): *likelihood,* indicating the compatibility of the evidence with the given hypothesis

p(E): *marginal likelihood*, indicating total probability of E which is the same for all possible hypothesis being considered.

$$\Rightarrow p(H|E) \propto p(E|H) * p(H)$$

The *posterior probability of a hypothesis is determined by a combination of the inherent likeliness of a hypothesis (the prior) and the compatibility of the observed evidence with the hypothesis (the likelihood); the factor* $\frac{p(E|H)}{p(E)}$ *represents the impact of evidence (data), E, on the probability of H.*

- $\frac{p(E|M)}{p(E)} > 1$: given the model is true, the evidence would be more likely than is predicted by the current state of belief. The reverse applies for a decrease in belief.
- $\frac{p(E|M)}{p(E)}$=1: the evidence is independent of the model. It would be exactly as likely as predicted by the current state of belief.

3.1.2. Informal Bayesian Inference

Informal Bayes' inference suggests one should reject a hypothesis if the evidence doesn't match up with the hypothesis, or if it is extremely unlikely a prior even the evidence does appear to match up.

Example:

Imaging that one has various hypotheses about the nature of a newborn puppy of a friend, including:

H1: the puppy has brown eyes

H2: the puppy has white fur

H3: the puppy has five legs

Then consider two scenarios:

Scenario #1: One is presented with evidence in the form of a picture of a puppy with white fur and black eyes. One finds this evidence supports H2 and opposes H1 and H3.

Scenario #2: One is presented with evidence in the form of a picture of a black-eye, black-fur, and five-leg puppy. Although the evidence supports H3, one's prior belief in this hypothesis is extremely small, so the posterior probability is nevertheless small. For more info on Bayesian inference, refer to Wikipedia page: http://en.wikipedia.org/wiki/Bayesian_inference

3.2. Bayesian Prediction

Predictions in frequentist statistics often involves finding an optimum point estimate of the parameter, e.g. by maximum likelihood, and then plugging this estimate into the formula for the distribution of a data

point. This approach has the disadvantage that it does not account for any uncertainty in the value of the parameter and therefore will underestimate the variance of the predictive distribution. Bayesian approach use the *posterior predictive distribution* to do predictive the distribution of a new, unobserved data point. It is, instead of a fixed point as a prediction, a distribution over possible points is returned.

- the **posterior predictive distribution** is the distribution of a new data point, marginalized over the posterior: $\mathrm{p}(\tilde{\mathrm{x}}|\mathrm{X},\alpha) = \int_{\theta}\ \mathrm{p}(\tilde{\mathrm{x}}|\theta) * \mathrm{p}(\theta|\mathrm{X},\alpha)\mathrm{d}\theta$
- the **prior predictive distribution** is the distribution of a new data point, marginalized over the prior: $p(\tilde{x}|\alpha) = \int_{\theta}\ p(\tilde{x}|\theta) * p(\theta|\alpha)d\theta$

Where $\tilde{\mathrm{x}}$ denotes a new data point whose distribution is to be predicted.

If evidence is simultaneously used to update belief over a set of *exclusive and exhaustive propositions*, Bayesian inference can be generalized to act on this belief distribution as a whole. Consider a process of generating iid events, E_n, the probability distribution of E_n is unknown. Let the event space Ω represent the current belief about the process. Suppose one has M_m hypothesis (models) about the process. Each model is defined by $p(E_n|M_m)$. $p(M_m)$ is degree of initial belief in M_m, and $\{p(M_m)\}$ is a set of initial prior which sum to 1. Suppose one observes event $E \in \{E_n\}$, for each model $M \in \{M_m\}$, the prior p(M) is updated to posterior p(M|E) based on Bayes' theorem:

$$\mathrm{p}(\mathrm{M}|\mathrm{E}) = \frac{\mathrm{p}(\mathrm{E}|\mathrm{M})}{\sum_{\mathrm{m}} \mathrm{p}(\mathrm{E}|\mathrm{Mm}) * \mathrm{p}(\mathrm{Mm})} * \mathrm{P}(\mathrm{M})$$

For multiple observations, $\mathbf{E}=\{e_1,\ldots, e_n\}$:

$$p(M|\boldsymbol{E}) = \frac{p(\boldsymbol{E}|M)}{\sum_m p(\boldsymbol{E}|Mm) * p(Mm)} * P(M) = \frac{\prod p(ei|M)}{\sum_m p(\boldsymbol{E}|Mm) * p(Mm)} * P(M)$$

By parameterizing the space of models, the belief in all models can be updated in a single step. The distribution of belief over the model space can then be considered as a distribution of belief over the parameter space. Suppose vector θ span the parameter space, the initial prior distribution over θ is p(θ|α), where α is as set of parameters to the prior. Let $e_1 \sim p(e|\theta)$, the posterior distribution over θ is given by:

$p(\theta|\boldsymbol{E},\alpha) = \frac{p(\boldsymbol{E}|\theta,\alpha)}{p(\boldsymbol{E}|\alpha)} * P(\theta|\alpha) = \frac{p(\boldsymbol{E}|\theta,\alpha)}{\int_{\theta}\ p(\boldsymbol{E}|\theta,\alpha)*p(\theta|\alpha)d\theta} * P(\theta|\alpha) = \frac{\prod p(ei|\theta,\alpha)}{\int_{\theta}\ p(\boldsymbol{E}|\theta,\alpha)*p(\theta|\alpha)d\theta} * P(\theta|\alpha)$ For more info on Bayesian prediction, refer to Wikipedia page: http://en.wikipedia.org/wiki/Bayesian_inference.

3.3. Bayesian Network

Refer to Michal Horný's report on Bayesian Network.

3.4. Software

1) **SAS**: BAYES statement (specify prior and estimate the parameters by Markov chain Monte Carlo sampling approach) is available to

a. PROC GENMOD: GLM
b. PROC LIFEREG: time to event (survival)
c. PROC PHREG: cox regression

In addition, PROC MCMC is general-purpose Bayesian modeling procedure that Bayesian models with arbitrary priors and likelihood functions. The BAYES statement uses the Gibbs sampler while PROC MCMC uses a self-tuning Metropolis algorithm.

Tips: Assess the convergence of the Markov Chains. Inferences can be inaccurate and misleading if not converged properly.

http://support.sas.com/documentation/onlinedoc/stat/930/introbayes.pdf

2) **WinBUGS**: free, runs under Microsoft Windows; the latest version available is 1.4.3 but it is not developed anymore. It can be run directly from R using *R2WinBUGS*.
http://www.mrc-bsu.cam.ac.uk/bugs/winbugs/contents.shtml
3) **OpenBUGS**: free, open source variant of WinBUGS, run under Windows, Linux, and R; currently under development. It can be run directly from R using *BRugs*.
http://www.openbugs.info/w/
http://www.mrc-bsu.cam.ac.uk/bugs/documentation/Download/manual05.pdf (Manual)

4. Bayesian Modeling

The principle of Bayes model is to compute posteriors based on specified priors and the likelihood function of data. It requires researchers to appropriately specify priors given inappropriate priors may lead to biased estimates or make computation of posteriors difficult. In this section, we will briefly go through several most common priors, how posteriors are calculated based on priors, and how the selection of priors influence posterior computation.

4.1. Prior Probability Distribution (Prior)

Statistical models using Bayesian statistics require the formulation of a set of prior distributions for *any* unknown parameters. The probability distributions express one's uncertainty about an unknown quantity, *p,* before the "data" is taken into account. It is meant to attribute uncertainty rather than randomness to the uncertain quantity. The unknown quantity may be a parameter or latent variable. It is often the purely subjective assessment of an experienced expert. The following sections will introduce the most frequently used prior—conjugate prior and prior used in hierarchical models—hierarchical prior.

4.1.1. Conjugate Prior

If the posterior distributions $p(\theta|x)$ are in the same family as the prior probability distribution $p(\theta)$, the prior and posterior are then called *conjugate distributions*, and the prior is called a *conjugate prior* for the likelihood. The commonly use of conjugate prior in Bayesian modeling is probably driven by a desire for computational convenience. When researchers have limited knowledge about the

distributional attributes of prior, conjugate prior usually is the best choice because it provides a practical way to obtain the posterior distributions. From Bayes' theorem, the posterior distribution is equal to the product of the *likelihood function, θ→p(x|θ),* and *prior, p(θ)*, normalized (divided) by the probability of the *data, p(x)*. The use of prior ensures the posterior has the same algebraic form as the prior (generally with different parameter values). Further, conjugate priors may give intuition, by more transparently showing how a likelihood function updates a distribution. The following examples shows how conjugate prior is updated by corresponding data and ensure posterior possess the same distributional format.

Example[4]:

1) If the likelihood is poisson distributed, y~ poisson(μ), a conjugate prior on μ is the Gamma distribution.

Prior p(μ) ~ **gamma(v, ρ)** => μ| v, ρ= $\frac{\rho^v}{\Gamma(v)}\mu^{v-1}\exp(-\mu\rho)$

Data $p(Y_n|\mu)$ ~ poisson(μ) => y|μ= $\frac{\mu^y exp^{-\mu}}{y!} \propto \mu^{\sum y}\exp(-n\mu)$

Posterior $p(\mu|Y_n)= \frac{p(\mathrm{Yn}|\mu)*p(\mu)}{p(\mathrm{Yn})} = \frac{1}{p(\mathrm{Yn})} * \mu^{\sum y}\exp(-n\mu)* \frac{\rho^v}{\Gamma(v)}\mu^{v-1}\exp(-\mu\rho)$

$\propto \mu^{\sum y+v-1}\exp[-\mu(n+\rho)]$ ~ **gamma(v+∑y, ρ+n)**

	Prior	*Data*	*Posterior*
	μ ~ gamma(v, ρ)	*Y|μ ~ poisson(μ)*	*μ|Y ~ gamma(v+∑y, ρ+n)*
Shape	v		**v+∑y**
Rate	ρ		ρ+n
E(μ)	v/ ρ		(v+∑y)/(ρ+n)
Var(μ)	v/ ρ2		(v+∑y)/(ρ+n)2

2) Beta family as conjugate priors for Binomial data

Prior p(π) ~ **beta(α, β)** => π| α, β= $B(\alpha,\beta)\pi^{\alpha-1}(1-\pi)^{\beta-1}$

Data $p(Y_n|\pi)$ ~ binomial(π) => y|π= $\binom{n}{y}\pi^y(1-\pi)^{n-y}$

Posterior $p(\pi|Y_n)= \frac{p(\mathrm{Yn}|\pi)*p(\pi)}{p(\mathrm{Yn})} = \frac{1}{p(\mathrm{Yn})} * \binom{n}{y}\pi^y(1-\pi)^{n-y} *B(\alpha,\beta)\pi^{\alpha-1}(1-\pi)^{\beta-1}$

$\propto \pi^{\alpha+y-1}(1-\pi)^{\beta+n-y-1}$ ~ **beta(α+y, β+n-y)**

	Prior	*Data*	*Posterior*
	π ~ beta(α, β)	*Y|π ~ binomial(n, π)*	*π|Y ~ beta(α+y, β+n-y)*
α	α		α+y

[4] Referred to Dr. Gheorghe Doros' lecture on "Bayesian Approach to Statistics – Discrete Case"

β	β		β+n-y
E(π)	$\frac{\alpha}{(\alpha+\beta)}$		$\frac{(\alpha+y)}{(\alpha+\beta+n)}$
Var(π)	$\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$		$\frac{(\alpha+y)(\beta+n-y)}{(\alpha+\beta+n)^2(\alpha+\beta+n+1)}$

Though conjugate prior theoretically provide considerable advantage of computing posterior, it is unrealistic to assume conjugate prior is applicable to all kinds of data. However, any data distribution belonging to members of the natural exponential families (NEFs) with quadratic variance functions has conjugate priors[5]. This family includes distributions: normal, poisson, gamma, binomial, negative binomial, and the NEFs generated by the generalized hyperbolic secant (GHS). Morris and Lock (2009) graphically illustrate the statistical properties of NEFs and their connections with corresponding conjugate distributions[6]. Another study (Morris, 1983) provides a comprehensive overview of how each member of NEFs is related to its prior and posterior accordingly[7]. Additionally, Wikipedia provides comprehensive tables summarizing various conjugate prior/likelihood combinations along with the corresponding model, prior, and posterior parameters in a less mathematical manner[8].

4.1.2. Hierarchical Prior

In the use of conjugate prior it requires researchers to indicate the parameters of prior. To distinguish them from parameters of the model of the underlying data, they are often referred as *Hyperparameter.* Furthermore, except for using a single value for a given hyperparameter, researchers can instead taking a probability distribution on the hyperparameter itself--it is then called a *hyperprior*.

Example:

One is using a beta distribution to model the distribution of the parameter p of a Bernoulli distribution:

1) The Bernoulli distribution (with parameter *p*) is the *model* of the underlying system
2) p is a parameter of the underlying system (Bernoulli distribution)
3) The beta distribution (with parameters α and β) is the prior distribution of p
4) α and β are parameters of the prior distribution (beta distribution), hence hyperparameters
5) A prior distribution of α and β is thus a hyperprior

Hierarchical prior allows one to express uncertainty in a Bayesian model: taking a fixed prior is an assumption, varying a hyperparameter of the prior allows one to do sensitivity analysis on this assumption, and taking a distribution on this hyperparameter allows one to express uncertainty in this assumption: "assume that the prior is of this form (this parametric family), but that we are uncertain as to precisely what the values of the parameters should be". For more info on hyperparameter and

[5] Gelman, Andrew, and John Carlin. Bayesian Data Analysis. 2nd ed. CRC Press, 2003.
[6] Morris, Carl N., and Kari F. Lock. "Unifying the Named Natural Exponential Families and Their Relatives." The American Statistician 63, no. 3 (August 2009): 247–253. doi:10.1198/tast.2009.08145.
[7] Carl N., Morris. "Natural Exponential Families with Quadratic Variance Functions: Statistical Theory." Institute of Mathematical Statistics 11, no. 2 (n.d.): 515–529.
[8] The Wikipedia page about Conjugate Prior at http://en.wikipedia.org/wiki/Conjugate_prior

hyperprior, refer to Wikipedia page: http://en.wikipedia.org/wiki/Hyperparameter and http://en.wikipedia.org/wiki/Hyperprior.

4.2. Posterior Probability Distribution (Posterior)

The posterior probability of a random event or an uncertain proposition is the conditional probability that is assigned after the relevant evidence is taken into account. The posterior probability distribution of one random variable given the value of another can be calculated with Bayes' theorem by multiplying the prior probability distribution by the likelihood function, and then dividing by the normalizing constant, as follows:

$$f_{X|Y=y}(x) = \frac{f_X(x) L_{X|Y=y}(x)}{\int_{-\infty}^{\infty} f_X(x) L_{X|Y=y}(x)\, dx}$$

$f_{X|Y=y(x)}$ gives the posterior probability density function for a random variable X given the data $Y = y$, where $f_X(x)$ is the prior density of X; $L_{X|Y=y(x)} = f_{Y|X=x(y)}$ is the likelihood function as a function of x; $\int f_{X(x)} * L_{X|Y=y(x)}\, dx$ is the normalizing constant. In brief, the above formula implies *posterior probability* $\propto$ *prior probability* * *likelihood.* However, Bayesian modeling is still controversial. Computation of Bayesian statistics could be very intractable. Potential solution is Markov Chain Monte Carlo.

4.3. Bayes Estimation

Generally, posterior calculations with the normal likelihood for the mean are simple as long as the prior are chosen from the normal family. Although many data themselves are not continuous, binomial and time to event for example, the estimates of the coefficients in generalized linear models (GLM) follow a normal distribution. With *sufficient data* all the usual estimates for GLM are approximated well by a normal distribution.

By contrast to classical methods which reports the maximum likelihood estimator of a parameter, Bayesian approaches is primarily based on the posterior distribution. All relevant information about the parameter (given the data and prior experience) is summarized in the posterior distribution. There are various ways, including the mean, median and mode of the parameter's posterior distribution, in which one can summarize the distribution. Mostly, point estimation is conducted though reporting the posterior mean. The following section will then introduce how Bayesian approach conduct interval estimation and hypothesis testing.

4.3.1. Interval Estimation

In Bayesian approach the interval estimate for a parameter of interest (θ) is called *Credible Interval* or *Posterior Intervals*. Any interval that has 95% posterior probability is called a 95% *Credible Interval* (CI). Given a posterior distribution $p(\theta|y)$, A is a credible set for θ if $p(\theta \in A|y) = \int_A p(\theta|y) d\theta$. For

example,

1) One-sided lower 95% CI for θ: p(θ<$\boldsymbol{\theta_U}$|data)=0.95 or p(θ≥$\boldsymbol{\theta_U}$|data)=0.05
2) One-sided upper 95% CI for θ: p(θ>θL|data)=0.95 or p(θ≤θL|data)=0.05
3) Two-sided 95% CI for θ: p(θL<θ< θU|data)=0.95 or p(θ≥θU or θ≤θL|data)=0.05

One can construct credible sets that have equal tails. *Equal-Tail-Area intervals* divide the probability of the complement in two equal areas: p(θ≤$\boldsymbol{\theta_L}$|data)=p(θ≥$\boldsymbol{\theta_U}$|data)=0.025; another frequently used Bayesian credible set is called the *Highest Posterior Density (HPD) intervals.* A HPD interval is a region that contains parameters with highest posterior density values. HPS ensures that the posterior density value at the ends of the interval is the same. For symmetric posteriors: HPD=equal-tail-area interval. For skewed posteriors: HPD is shorter than equal-tail-area interval; however, HPD interval is more difficult to construct[9]. Figure 1 graphically illustrates the difference between equal-tail-area and HPD intervals.

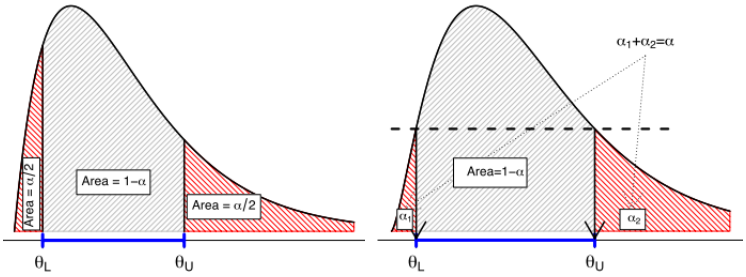


Figure 1: Equal-Tail-Area Intervals (left) vs. HPD Interval (right)

(Source: Dr. Gheorghe Doros' lecture on "Estimation and Hypothesis Testing in Clinical", page 14)

4.3.2. Hypothesis Testing

Bayesian modeling is considered *Hypotheses driven research* meaning that researchers are asked to state certain hypotheses (beliefs) before data collection and then use the data to support or dismiss the hypotheses. Three main types of hypotheses[9]:

1) Superiority: the new intervention is better than the control (standard). It implies
   H_A: θ_T(experiment)>θ_C(control)+δ(superiority margin)
2) Equivalency: two interventions are equivalent. It implies
   H_A: θ_C- δ<θ_T< θ_C+ δ(equivalency margin)
3) Non-inferiority: the new intervention is as good as the control (standard). It implies
   H_A: θ_T>θ_C(control)-δ(non-inferiority margin)

Let θ_s (θ_C+δ) and θ_I (θ_C-δ) represent superiority and inferiority thresholds respectively. Figure 2 shows: the equivalency region (grey) implies no statistical difference between control and intervention; the superiority region (green) suggests intervention surpasses control; the inferiority region (orange)

[9] Referred to Dr. Gheorghe Doros' lecture on "Estimation and Hypothesis Testing in Clinical Trials – Bayesian Approach "

indicates control in fact is better than intervention. In sum, conclusions about hypothesis test are reached by relating the info on θ, including prior, data, and posterior, to the three intervals: inferiority interval $(-\infty, \theta_I)$, equivalency interval (θ_I, θ_S), and superiority interval (θ_S, ∞).

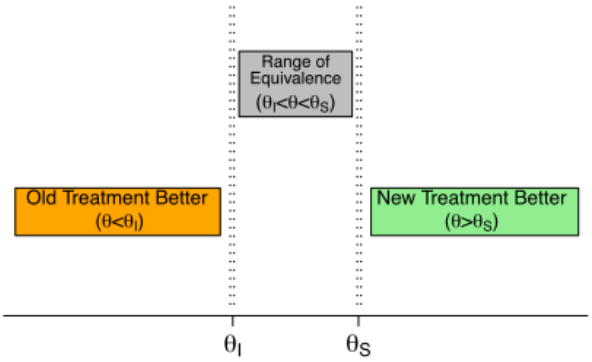


Figure 2: Main Types of hypothesis testing using Bayesian approach

(Source: Dr. Gheorghe Doros' lecture on "Estimation and Hypothesis Testing in Clinical Trials", page 21)

4.4. Non-informative Bayes Models

If no prior information is available, one needs to specify a prior that will not influence the posterior distribution and "let the data speak for themselves". This section first lists several frequently used non-informative prior. Next, it briefly explains the connection of non-informative Bayes model to random effect model. Last, it compares non-informative Bayes model to generalized estimating equation (GEE) model.

4.4.1. Non-informative Prior

Priors express vague or general information about a parameter (objective prior). The simplest and oldest rule for determining a non-informative prior is the *principle of indifference*, which assigns equal probabilities to all possibilities.

1) Uniform prior: one knows a ball is hidden under one of three cups, A, B or C, but no other information is available about its location. A uniform prior is $p(A) = p(B) = p(C) = 1/3$.

2) Logarithmic prior: to estimate an unknown proportion between 0 and 1, one may assume all proportions are equally likely and use a uniform prior. Or one may assume all orders of magnitude for the proportion are equally likely. The logarithmic prior is the uniform prior on the logarithm of proportion.

3) Jeffrey's prior: a prior distribution on parameter space that is proportional to the square root of the determinate of the Fisher information. It expresses the same belief no matter which metric is used: $p(\theta) \propto \sqrt{\det \mathrm{I}(\theta)}$.

For more info about prior, refer to Wikipedia page:
http://en.wikipedia.org/wiki/Prior_probability#Uninformative_priors

Several issues that researchers need to consider when adopt non-informative priors: (1) in parameter estimation problems, the use of a non-informative prior typically yields results which are not too different from conventional statistical analysis, as the likelihood function often yields more information than the non-informative prior. (2) Non-informative priors are frequently *improper*, that is the area under the prior density is not unity—the sum or integral of the prior values are not equal to 1. In most case, improper priors, for instance beta(0, 0), uniform distribution on an infinite interval, and Logarithmic prior on positives reals, can be used in Bayesian analyses without major problems. However, things to watch out for are:

- use of improper priors might result in improper posteriors
- use of improper priors makes model selection and hypothesis testing difficult
- WinBUGS does not allow the use of improper priors

4.4.2. Connection to Random Effect Model (Hierarchical Model)

Suppose one have data in which i=1,...,n observations that belong to one of j=1,…,J groups (e.g. patients within hospitals). If one have covariates on multiple levels, fitting a fixed effects model which estimates dummy variables for J-1 groups controls for the effects inherent in groups on dependent variable. However it involves estimating a bunch of dummies, and also cannot take into account group-level covariates. An alternative is to use a random effects model. Instead of assuming a completely different intercept for each group, it assumes that the intercepts are drawn from a common distribution. It is relatively easier to fit random effects models and estimate parameters using Bayesian methods[10].

Example:

Fixed effects model: $y_i = \alpha_{j\,[i]} + x_{i1}\beta_1 + x_{i2}\beta_2 + \varepsilon_i$ controls for the time-invariant variables by letting the intercept α vary by group (j). It takes into account ALL group-level variables but can't estimate the effect of individual group-level covariate.

Random effects model: $y_i = \alpha_{j\,[i]} + x_{i1}\beta_1 + x_{i2}\beta_2 + \varepsilon_i$, $\alpha_j \sim N(\alpha, \sigma^2_\alpha)$ assumes intercepts are drawn from a normal distribution. One can incorporate group-level covariates in the following way:

$y_i = \alpha_{j\,[i]} + x_{i1}\beta_1 + x_{i2}\beta_2 + \varepsilon_i$

$\alpha_j \sim N(\gamma_0 + \mu_{j1}\gamma_1, \sigma^2_\alpha)$, γ represents group-level covariates

Then one can fit hierarchical models easily using Bayesian methods:

$p(\alpha, \beta, \gamma \mid \mathbf{y}) \propto p(\mathbf{y} \mid \alpha, \beta, \gamma) * p(\alpha|\gamma)p(\gamma)p(\beta)$

Last, solve for the joint posterior using Gibbs Sampling or Metropolis Hasting.

[10] Referred to "Bayesian Statistics in One Hour" by Patrick Lam, accessible at http://www.people.fas.harvard.edu/~plam/teaching/methods/bayesianhour/bayesianhour_print.pdf

4.4.3. Compared with GEE Model

The generalized estimating equation (GEE) assumes the variance of the cluster level random effect as a constant. It permits the specification of a working correlation matrix accounting for the form of within-cluster correlation of the outcomes. Bayesian random-effects regression assumes the variance of the cluster level random effect is an unknown parameter. It takes the uncertainty into account by assuming a prior distribution which presents the researcher's pre-belief or external information about uncertainty. If one does not have strong belief or have simply limited information about the extent to which outcomes are correlated with each other within cluster, non-informative prior is desirable and the results from non-informative Bayes model should be comparable to the results from the classical statistical methods, says GEE models. Though Bayesian approach is more flexible, it is critical to assess the non-convergence of the Markov Chain by examining the estimated Monte Carlo error for the posterior and visually checking the dynamic trace plots, times series plots, and density plots of all covariates, if applicable.

Ma and Thabane et. al's study (2009) compared three cluster-level and six individual-level statistical analysis methods in the analysis of binary outcomes from a cluster randomized trials[11]. The individual level analyses included (1) standard logistic regression, (2) robust standard errors approach, (3) generalized estimating equations, (4) random-effects meat-analytic approach, (5) random-effects logistic regression, and (6) Bayesian random-effects regression. They found Bayesian random-effects logistic regression yielded the widest 95% interval estimate for the odds ratio and led to the most conservative conclusion, though the results remained robust under all methods. The individual-level standard logistic regression is the least appropriate method as it ignores the correlation of the outcomes for the individuals within the same cluster.

4.5. Informative Bayes Models

If prior information is available, it should be appropriately summarized by the prior distribution. Such priors are just adding pseudo observations to the data.

4.5.1. Informative Prior

Priors express specific, definite information about a parameter (subjective prior). Usually, specification of the prior mean and variance is emphasized. The prior mean provides a prior point estimate for the parameter of interest, while the variance expresses ones uncertainty concerning this estimate. When one has strong belief this estimate is accurate, the variance must be set low. For example, to set up the prior distribution for the temperature at noon tomorrow, one can make the prior

[11] Ma, Jinhui, Lehana Thabane, Janusz Kaczorowski, Larry Chambers, Lisa Dolovich, Tina Karwalajtys, and Cheryl Levitt. "Comparison of Bayesian and Classical Methods in the Analysis of Cluster Randomized Controlled Trials with a Binary Outcome: The Community Hypertension Assessment Trial (CHAT)." BMC Medical Research Methodology 9, no. 1 (June 16, 2009): 37.

a normal distribution with expected value equal to today's noontime temperature and variance equal to the day-to-day variance of atmospheric temperature.

4.5.2. Soliciting Informative Prior

A critical part of the Bayesian paradigm is the prior distribution. For the Bayesian approach we need to specify a prior distribution on the possible values of the parameter and then, using the data, update the prior distribution and construct the posterior using Bayes Theorem. Prior distributions are a subjective representation of belief. Posterior probabilities depend on prior probabilities and thus they are subjective too! 3 important properties of the PRIOR-POSTERIOR transformation[12]:

1) Strong prior opinions: when a subjective opinion assigns a probability of 1 to a single value, it has a strong effect on the posterior. In this case, the, posterior is the same as the prior and the data is practically ignored. The effect of strong prior on posterior is illustrated in Figure 3 (right panel).
2) Strong data diminishes the effects of the prior on the posterior.
3) Weak prior opinions: when a subjective opinion includes a wide range of values which are 'equally' probable, it results in a posterior mostly shaped by the data, shown in Figure 3 (left panel).

When a prior distribution might dominate the data, researchers need to be cautious about the appropriateness of prior derivation and whether the selected prior is justification. This report simply points the most important issues that researchers should keep in mind: (1) prior should be chosen before one sees the data. Usually, there are some prior information available, e.g. previous studies; (2) assign non-informative prior if one know nothing about the parameter. In brief, the more data is available, the less the posterior distribution would be influenced by the prior, vice versa. For further guidance on selection prior, refer to the "Useful Resources" at end of this report.

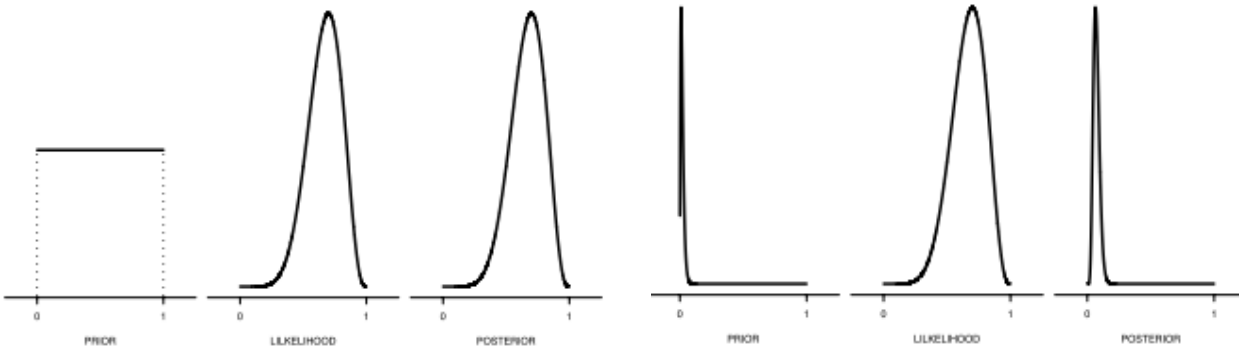


Figure 3: Effect of informative prior on posterior [left: weak prior; right: strong prior]

(Source: Dr. Gheorghe Doros' lecture on "Prior Elicitation", page 2)

[12] Referred to Dr. Gheorghe Doros' lecture on "Prior Elicitation"

5. Examples: Fitting Bayesian Logistic Regression in SAS

Prior to the launch of MCMC (Markov Chain Monte Carlo) procedure in SAS/STATA 9.2, fitting Bayesian modes is primarily carried through WinBugs or OpenBugs. The MCMC procedure is a general procedure that fits Bayesian models with arbitrary priors and likelihood functions. In addition to the MCMC procedure, GENMOD, LIFEREG, and PHREG procedures also provide Bayesian analysis along with the standard frequentist analyses that have always performed. This section demonstrates how to fit Bayesian logistic regression in PROC GENMOD and PROC MCMC.

Pediatricians have reported that less than 50% of teen girls in the US had initialed the human papillomavirus (HPV) vaccination[13]. However, it is not clear whether the low rate of immunization is associated with teen's demographic characteristics and provider attributes. This example uses data of the National Immunization Survey-Teen (year 2009) from the CDC to explore what demographic and provider characteristics predict HPV vaccination rates among teen girls aged 13-17 in NY. Outcome of interest is the probability of receiving at least one dose of HPV vaccine. The example builds a logistic model that regresses the probability of initiating HPV vaccination on a set of covariates including teen's race and age, mom's age and education, provider recommendation of HPV, Vaccine for Children (VFC) program[14], and immunization registry, and facility type.

Let Y_i denote the response variable (subscript i represents individual teen girl), p_udthpv, flagging if teens received at least one HPV vaccines; **X** denotes a vector representing the set of covariates described above. Applying a generalized linear model (GLM), we can fit the data points Y_i with a binary distribution, $Y_i \sim \text{binary}(P_i)$ where P_i is the probability of Y_i equal to 1, and links it to the regression covariates, **X**, through a logit transformation. The Bayesian model is given by $\Pr(\boldsymbol{\beta} \mid \text{logit}(P_i), \mathbf{X}) \propto \Pr(\text{logit}(P_i) \mid \mathbf{X}, \boldsymbol{\beta}) * \Pr(\boldsymbol{\beta})$ where $\text{logit}(P_i)$ is the likelihood function of data. The main advantage of GLM is that it allow for using any distribution of the exponential family. In this example, GLM assumes $\text{logit}(P_i)$ is normally distributed with a mean of $\boldsymbol{\beta}\mathbf{X}$ where $\boldsymbol{\beta}$ is a vector representing regression coefficients.

PROC GENMOD offers convenient access to Bayesian analysis for GLM. We can specify a model essentially the same way as we do from a frequentist approach, but add a BAYES statement to request Bayesian estimation. A sample code for fitting a Bayesian logistic linear regression is provided below:

```
proc genmod data=ads desc;
    class white/ param=glm order=internal desc;
    model p_utdhpv=white age momage … …/ dist=bin link=logit;
    bayes seed=1 coeffprior=normal nbi=1000 nmc=20000 outpost=posterior;
```

[13] "HPV Vaccination Initiation, Completion Rates Remain Low among Teen Girls | Infectious Diseases in Children." Accessed April 22, 2013. http://www.healio.com/pediatrics/news/print/infectious-diseases-in-children/%7Bba020c72-98de-4c1c-928e-89902eda921f%7D/hpv-vaccination-initiation-completion-rates-remain-low-among-teen-girls.

[14] VFC program offers vaccines at no cost for eligible children through VFC-enrolled providers.

```
run;
```

We first specify the GLM model in the MODEL statement as usual. In the following BAYES statement, SEED option specifies an integer seed for the random number generator in the simulation. It enables researcher to reproduce identical Markov chains for the same specification. COEFFPRIOR=NORMAL specifies a non-informative independent normal prior distribution with zero mean and variance of 10^6 for each parameter. NBI=1000 option specifies the number of burn-in iterations before the chain was saved. Burn-in refers to the practice of discarding an initial portion of a Markov chain sample so that the effect of initial values on the posterior inference is minimized. NMC=20000 option specifies the iterations after burn-in. The OUTPOST option names a SAS dataset for the posterior samples for further analysis.

Maximum likelihood estimates of the model parameters are computed by default (Output 1, right panel). The GLM model shows white race is negatively related to HPV initialization rate; while teen's age and provider recommendation is associated with increased rate of HPV vaccination. Summary statistics for the posterior sample are displayed in the left panel of Output 1. Since non-informative prior distributions for the regression coefficients were used, the mean, standard deviations, and the intervals of the posterior distributions for the model parameters are close to the maximum likelihood estimates and standard errors.

Output 1: Posterior Descriptive and Interval Statistics of Regression Coefficients (L: Bayesian; R: GLM)

Posterior Summaries							
Parameter	N	Mean	Standard Deviation	Equal-Tail Interval		HPD Interval	
Intercept	20000	-3.9978	1.5803	-7.1126	-0.9106	-7.0831	-0.8946
white1	20000	-0.7529	0.3210	-1.3889	-0.1373	-1.3972	-0.1474
AGE	20000	0.2111	0.1021	0.0106	0.4135	0.0136	0.4155
momage	20000	0.0612	0.2963	-0.5170	0.6409	-0.5070	0.6466
college	20000	0.4985	0.2953	-0.0809	1.0749	-0.0996	1.0504
poverty	20000	0.6044	0.4351	-0.2504	1.4681	-0.2290	1.4824
registry1	20000	0.0103	0.4120	-0.8069	0.8072	-0.8110	0.8002
private	20000	0.2485	0.3492	-0.4373	0.9320	-0.4382	0.9308
vfc	20000	-0.1383	0.4105	-0.9466	0.6662	-0.9391	0.6706
recommend	20000	1.4555	0.2980	0.8786	2.0484	0.8635	2.0282

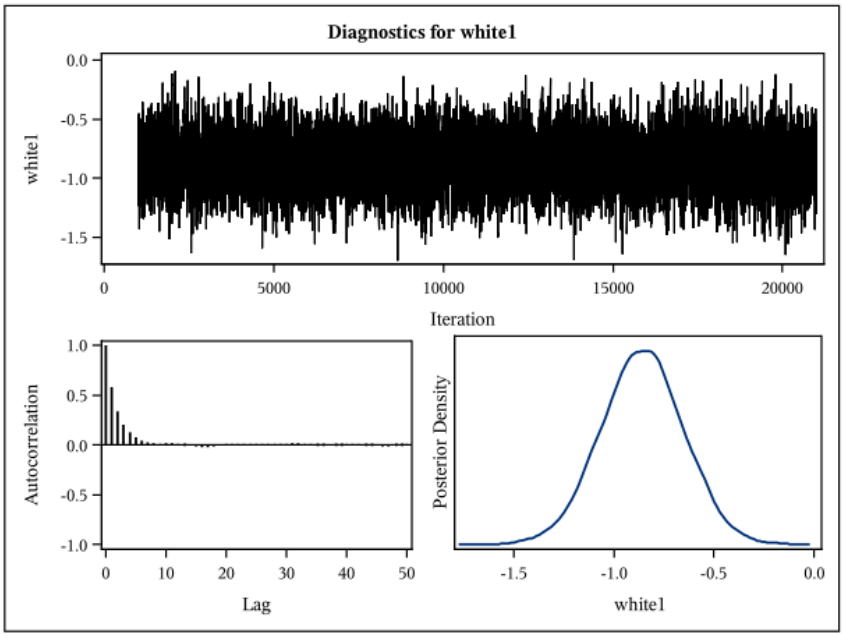
Analysis Of Maximum Likelihood Parameter Estimates						
Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits	
Intercept		1	-3.8554	1.5358	-6.8655	-0.8453
white	1	1	-0.7164	0.3130	-1.3299	-0.1029
white	0	0	0.0000	0.0000	0.0000	0.0000
AGE		1	0.2031	0.0994	0.0083	0.3980
momage		1	0.0605	0.2910	-0.5099	0.6309
college		1	0.4790	0.2913	-0.0918	1.0499
poverty		1	0.5799	0.4259	-0.2550	1.4147
registry1		1	0.0135	0.4055	-0.7813	0.8083
private		1	0.2415	0.3437	-0.4322	0.9152
vfc		1	-0.1321	0.4013	-0.9187	0.6546
recommend		1	1.3978	0.2940	0.8215	1.9741

Simulation-based Bayesian inference requires using simulated draws to summarize the posterior distribution or calculate any relevant quantities of interest. Therefore, researchers are suggested to treat the simulation draws with care. SAS performs various convergence diagnostics to help researchers determine whether the Markov chain has successfully converged—reached its stationary, or the desired posterior, distribution. One can assess Markov chain convergence by visually checking a number of diagnostic graphs automatically produced by SAS, including trace, correlation, and kernel density plots.

The trace plot (Output 2, upper panel) shows the mean of the Markov chain has stabilized and appears constant over the graph. Also, the chain has good mixing and is dense—it traverses the posterior space

rapidly. The autocorrelation plot (Output 2, bottom left panel) indicates no high degree of autocorrelation for each of the posterior samples, implying good mixing. The kernel density plot (Output 2, bottom right panel) estimates the posterior marginal distribution of parameters. In sum, these plots conclude the Markov chain has successfully converged to the desired posterior. Though, this example only displays diagnostic graphics for the covariate of interest (white race), it is essential that one visually examines the convergence of ALL parameters, not just those of interest. One cannot get valid posterior inference for parameters that appear to have good mixing if the other parameters have bad mixing.

Output 2: Diagnostic Plots for the Coefficient of White Race



Though fitting a Bayesian GLM model is relatively easier by using PROC GENMOD, the BAYES statement provides limited Bayesian capability—it can only request the regression coefficients be estimated by Bayesian method. When fitting a Bayesian logistic model, directly presenting the estimated coefficients is usually not clinically meaningful and therefore researchers ordinarily perform exponential transformations to convey the odds ratios. With regard to compute the more intuitive estimates, PROC MCMC is much more flexible than the Bayes statement in PROC GENMOD. The procedure allows users to simulate the point estimates and intervals of odds ratios instead, not just the coefficients. It can also be used to estimate the probability that the coefficients or odds ratios excess certain critical values. The following code illustrates how to fit Bayesian logistic model and calculate the estimates of odds ratios by using PROC MCMC.

```
proc mcmc data=ads seed=14 nbi=1000 nmc=20000 outpost=posterior monitor=(or);
    beginnodata;   /*the statement shown in red here does not imply a mistake*/
    or=exp(beta1);
    endnodata;
    parms beta: 0;  /*declare coefficients parameters and assign initial values*/
```

```
    prior beta: ~ normal(mean=0, var=1e6); /*assign normal prior to coefficients*/
    pi=logistic(beta0+ beta1*white+ beta2*age+ … …);
    model p_utdhpv ~ binary(pi);  /*specify likelihood function*/
run;
```

The PROC MCMC statement invokes the procedure and the options specified is similar to that in the BAYES statement of PROC GENMOD. The MONITOR option outputs analysis for the symbols of interest which is specified in the BEGINNODATA/ENDNODATA statement. In this case, we are interested in the odds ratio of initiating HPV vaccine among the white compared to minorities. The PARMS statement identifies the parameters in the model: regression coefficients of beta0, beta1, beta2…, and assigns their initial values as zeros. The PRIOR statement specifies prior distributions for the parameters. Again, we apply a normal prior with mean 0 and variance 10^6 for each parameter. The following "P_i" assignment statement calculates the logit of expected probability of initiating HPV vaccination as a linear function of a set of covariates. The MODEL statement specifies the likelihood function using the binary distribution to indicate that the response variable, p_utdhpv, follows binary distribution with parameter P_i.

Output 3 reports the estimated odds ratios of initiating HPV vaccine among the white compared to minorities. Again, since we adopted non-informative priors, the mean, standard deviation, and the interval of the posterior distribution for the odds ratio is close to the maximum likelihood estimate and standard error. It reveals that the odds ratios of receiving at least one dose of HPV vaccine is about 50% lower among the white teen girls compared to their counterparts of minorities when controlling for teen's and mom's demographic characteristics and the attributes of health care providers and facilities.

Output 3: Summary Statistics for the Odds Ratio of Initiating HPV Vaccine among the White (Upper Panel: Bayesian model; Button Panel: GLM model)

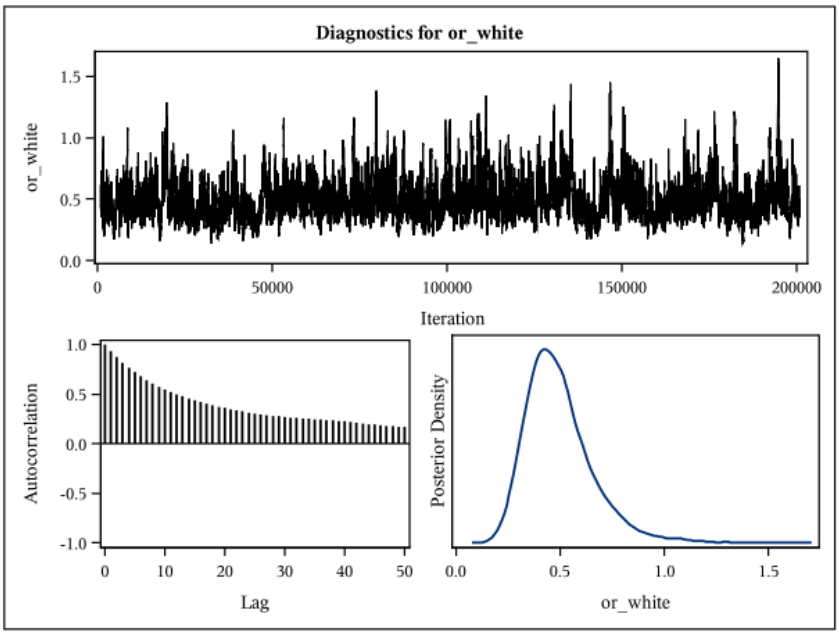
Posterior Summaries								
Parameter	N	Mean	Standard Deviation	Alpha	Equal-Tail Interval		HPD Interval	
or_white	20000	0.4956	0.1627	0.050	0.2528	0.8895	0.2226	0.8217

Contrast Estimate Results								
Label	Mean	Mean Confidence Limits		L'Beta Estimate	Standard Error	Alpha	L'Beta Confidence Limits	
	Estimate							
log O.R. white vs minorities	0.3282	0.2092	0.4743	-0.7164	0.3130	0.05	-1.3299	-0.1029
Exp(log O.R. white vs minorities)				0.4885	0.1529	0.05	0.2645	0.9022

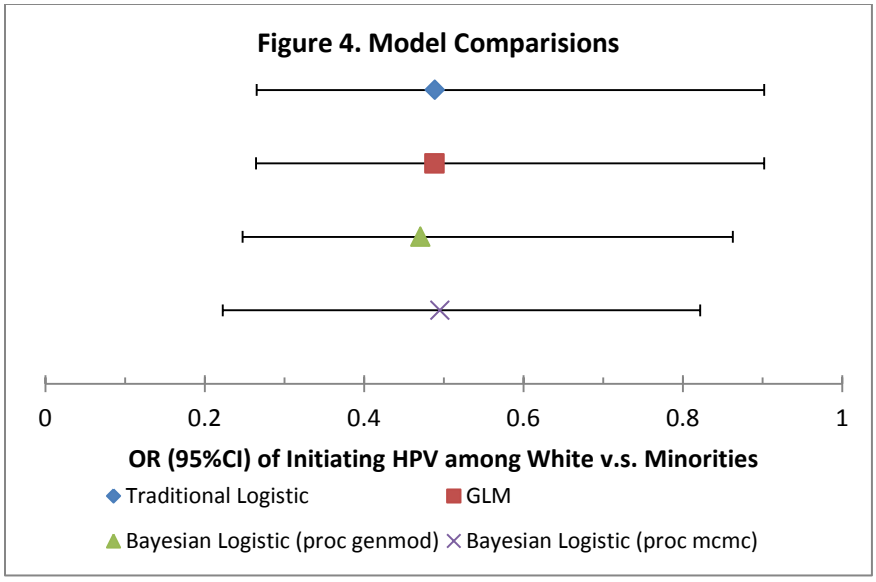
Output 4 displays the diagnostic plots for the estimated odds ratio of initiating HPV vaccine among the white compared to that of minorities. Here, the diagnostic graphs conclude the simulation draws are reasonably converged and therefore we can be more confident about the accuracy of posterior inference.

Additionally, the kernel density plot (bottom right panel) confirms the mean of odds ratio is around 0.5 which is consistent with the summary statistics in Output 3.

Output 4: Diagnostic Plots for the Odds of Initiating HPV Vaccine among the White



Last, this report compares the odds ratios estimated by logistic, GLM and Bayes' models. Bayes' model is fitted using both GENMOD and MCMC procedures. Figure 4 shows the estimated ORs are around 0.5 and consistent across different approaches. The intervals of odds ratios for both logistic and GLM models appear to be less symmetric. It is probably because of that the exponent transformed regression coefficient (OR) does not remain normally distributed even though its original value is assumed following normal distribution. In summary, the results of Bayes' models with non-informative priors are very similar to that of traditional logistic and GLM models—both point estimates and intervals do not seem to vary much.



6. Useful Resources

Tutorial Introduction of Bayesian Statistics

1) http://cocosci.berkeley.edu/tom/bayes.html (Bayesian reading list)

2) Bolstad, William M. Introduction to Bayesian Statistics. 2nd ed. Wiley-Interscience, 2007.

3) Lee, Peter M. Bayesian Statistics: An Introduction. John Wiley & Sons, 2012.

4) "Bayesian Statistic" wrote by Dr. José M. Bernardo. Accessible at http://old.cba.ua.edu/~mhardin/BayesStat.pdf (it includes a comprehensive glossary)

Applications of Bayesian Method in Health Services Research

5) Dixon-Woods, Mary, Shona Agarwal, David Jones, Bridget Young, and Alex Sutton. "Synthesising Qualitative and Quantitative Evidence: a Review of Possible Methods." Journal of Health Services Research & Policy 10, no. 1 (January 2005): 45–53.

6) Harbison, Jean. "Clinical Judgement in the Interpretation of Evidence: A Bayesian Approach." Journal of Clinical Nursing 15, no. 12 (December 2006): 1489–1497.

7) Spiegelhalter, D J, J P Myles, D R Jones, and K R Abrams. "Bayesian Methods in Health Technology Assessment: a Review." Health Technology Assessment (Winchester, England) 4, no. 38 (2000): 1–130.

Prior Selection

8) Berger, James O., José M. Bernardo, and Dongchu Sun. "The Formal Definition of Reference Priors." The Annals of Statistics 37, no. 2 (April 2009): 905–938.

9) Bernardo, Jose M. "Reference Posterior Distributions for Bayesian Inference." Journal of the Royal Statistical Society. Series B (Methodological) 41, no. 2 (January 1, 1979): 113–147.

10) Jaynes, E.T. "Prior Probabilities." IEEE Transactions on Systems Science and Cybernetics 4, no. 3 (Sept.): 227–241.

11) Kass, Robert E., and Larry Wasserman. "The Selection of Prior Distributions by Formal Rules." Journal of the American Statistical Association 91, no. 435 (1996): 1343–1370.

Bayesian Model

12) Christensen, Ronald. Bayesian Ideas and Data Analysis - CRC Press Book. CRC Press, 2010.

13) Kruschke, John K. Doing Bayesian Data Analysis: A Tutorial with R and BUGS. 1st ed. Academic Press, 2010.