BAYESIAN MODELING USING SAS/STAT®

CUSTOMER LOYALTY TEAM • Support You Can Count On

AGENDA

- What is Bayesian Analysis?
- Options in SAS/STAT
- Example using Proc FMM (Zero-Inflated Poisson model)
- Examples using Proc MCMC

WHAT IS BAYESIAN ANALYSIS?

- *Bayesian analysis* is a field of statistics that is based on the notion of conditional probability.
- It can be viewed as the formalization of the process of incorporating scientific knowledge using probabilistic tools.
- It provides uncertainty quantification of parameters by its conditional distribution in the light of available data.

BAYES' THEOREM

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

- P(A) is the prior probability of event A. It is called the *prior* because it does not take into account any information about event B.
- P(B|A) is the conditional probability of event B given event A.
- P(B) is the prior or marginal probability of event B.
- P(A|B) is the conditional probability of event A given event B. It is called the posterior probability because it is derived from the specified value of event B.

BAYESIAN ANALYSIS

- The Bayesian approach to statistical inference treats parameters as random variables.

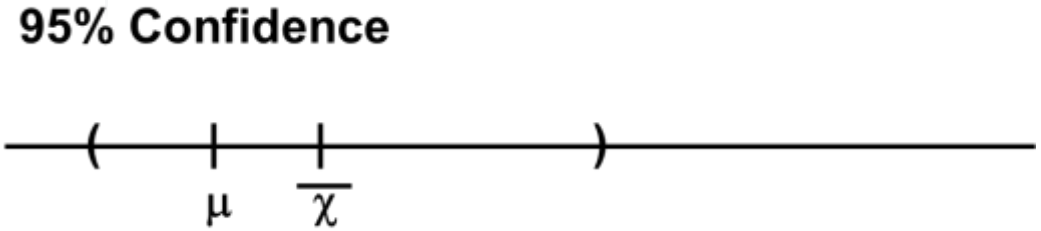
- It includes the incorporation of prior knowledge and its uncertainty in making inferences on unknown quantities (model parameters, missing data, and so on).

- It expresses the uncertainty concerning the parameter through probability statements and distributions.

FREQUENTIST APPROACH TO STATISTICS

- *Classical methods* consider the parameters to be fixed but unknown.
- They do not enable you to make probability statements about parameters because they are fixed.
- They are based on probabilities that are only for observations given the unknown parameters.
- They are judged by how they perform in an infinite number of hypothetical repetitions of the experiments.

CONFIDENCE INTERVALS – CLASSICAL APPROACH

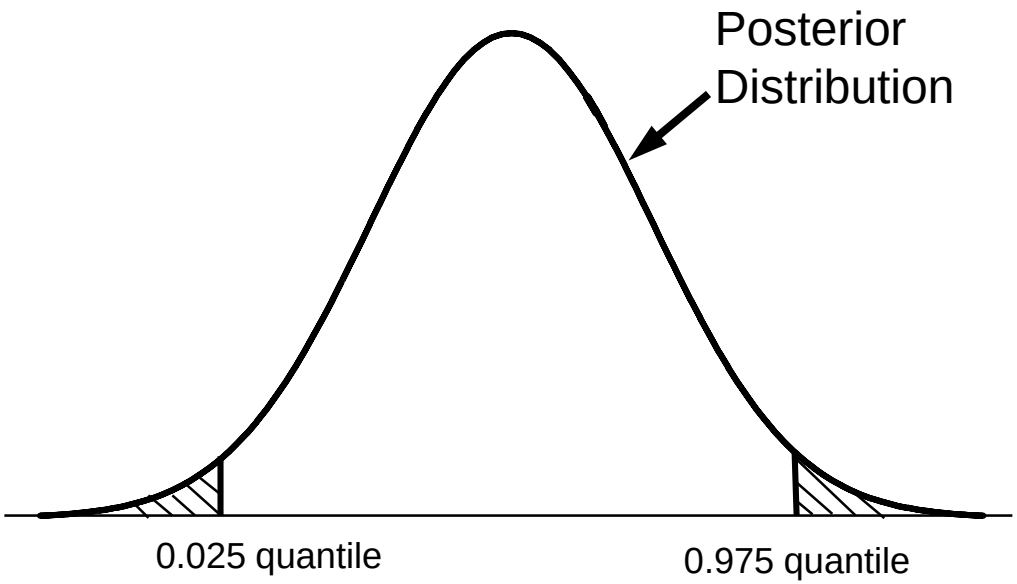


- A 95% confidence interval states that you are 95% confident that random interval contains the true mean.
  - In other words, if 100 different samples were drawn from the same population and 100 intervals were calculated, approximately 95 of them would contain the population mean.

BAYESIAN APPROACH TO STATISTICS

- Bayesian methods treat the unknown parameters as random variables.
- They enable you to make probability statements about parameters and observations.
- They interpret probabilities for parameters as “degree of belief” and can be subjective.
- They use the rules of probability to revise “degree of beliefs” about the parameters given the observed data.
- They base the inferences about the parameters on the probability distribution for the parameter.

95% CREDIBLE INTERVAL



There is a 95% chance that the parameter is in the credible interval.

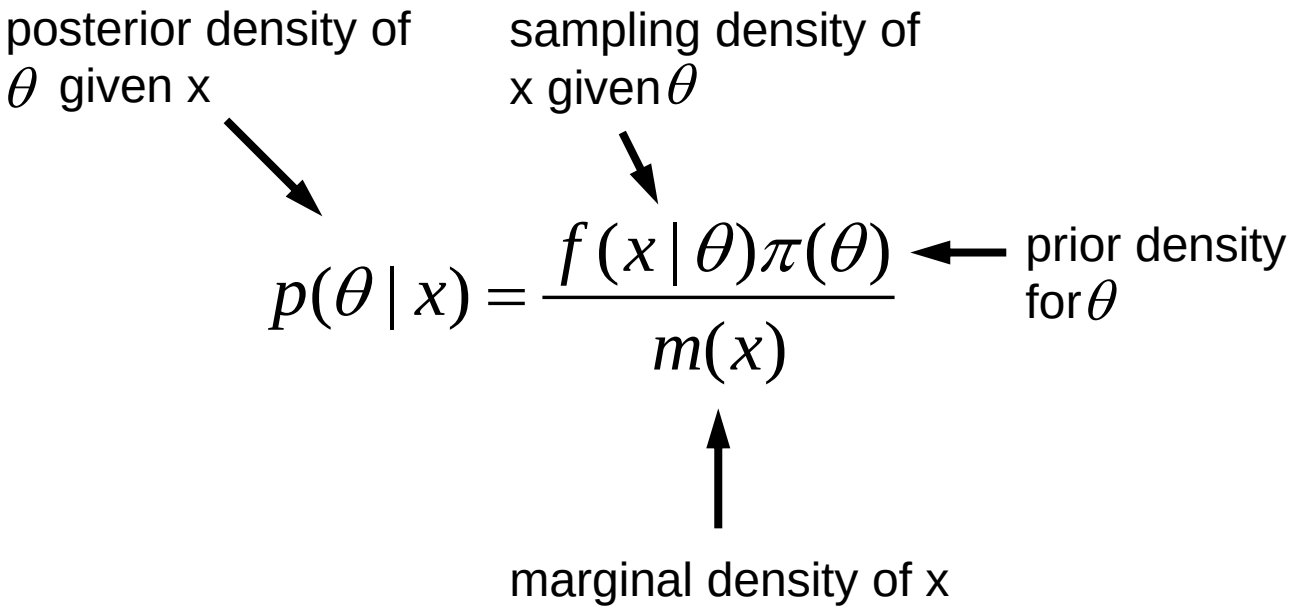
STEPS INVOLVED IN BAYESIAN INFERENCE

1. The probability distribution of the parameter, known as the *prior distribution*, is formulated.

2. Given the observed data, you choose a statistical model that describes the distribution of the data given the parameters.

3. You update your beliefs about the parameter by combining information from the prior distribution and the data through the calculation of the posterior distribution. This is carried out by using Bayes' theorem; hence the term Bayesian analysis.

THE BAYES' RULE



PRIOR DISTRIBUTIONS

- You cannot carry out any Bayesian inference or perform any modeling without using a prior distribution.
- It is not necessarily specified beforehand because prior does not refer to time.
- It is not necessarily unique, as the prior distribution could be a combination of prior distributions expressing a range of reasonable opinions.
- It is not necessarily completely specified, as it might be possible to have unknown parameters in the prior, which are then estimated.
- It is not necessarily important, as it could have a negligible influence on the conclusions, especially when the sample size is large.

ADVANTAGES OF BAYESIAN ANALYSIS

- Bayesian analysis is useful when you have prior information, either expert opinion or historical knowledge, that you want to incorporate into the analysis.
- It is useful if you want to communicate your findings in terms of probability notions that can be more easily understood by non-statisticians.
- It provides inferences that are conditional on the data and are exact, without reliance on asymptotic approximation.
- It provides the full uncertainty of parameters via the posterior distribution in contrast to point estimates and standard errors only.
- The simulations make the computations tractable even for complex hierarchical models.

DISADVANTAGES OF BAYESIAN ANALYSIS

- It does not tell you how to select a prior and there is no one correct way to choose a prior…
  - **Bayesian inferences require skills to translate subjective prior beliefs into a mathematically formulated prior**. If you do not proceed with caution, you can generate misleading results.
- It can produce posterior distributions that are heavily influenced by the priors.
- It often comes with a high computational cost, especially in models with a large number of parameters.

HOW DO WE IMPLEMENT IN SAS/STAT®?

BAYESIAN ANALYSIS IN SAS

- Bayesian methods in SAS 9.4 (& 9.3) are found in the following procedures:

  - the **FMM** procedure, which fits finite mixture models
  - the **GENMOD** procedure, which fits generalized linear models
  - the **PHREG** procedure, which performs regression analysis of survival data based on the Cox proportional hazards model
  - the **LIFEREG** procedure, which fits parametric models to survival data
  - the **MCMC** procedure, which is a general purpose Markov Chain Monte Carlo simulation procedure that is designed to fit Bayesian models.

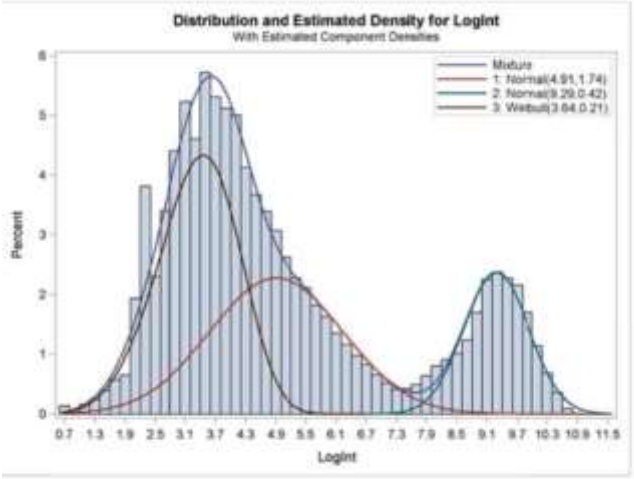
TWO AVENUES TO BAYESIAN COMPUTING

- Support for Bayesian analysis in four existing procedures
  - **GENMOD, LIFEREG, PHREG, and FMM**
  - You use options to specify prior distributions, generate posterior samples, and request convergence diagnostics and posterior summaries.

- **MCMC procedure**
  - General-purpose simulation procedure
  - You specify prior distributions and likelihood functions with programming statements.
  - Experimental in SAS 9.2, production in SAS 9.22

FMM PROCEDURE

- PROC FMM fits statistical models to data where the distribution of the response is a finite mixture of univariate distributions.
- Performs maximum likelihood estimation for all models
- Provides Bayesian analysis for several models.
- Useful for applications such as
  - estimating multimodal or heavy-tailed densities
  - modeling over dispersed data.



GENMOD PROCEDURE

- PROC GENMOD provides Bayesian analysis for generalized linear models.
- Sampling methods include adaptive rejection Metropolis sampling (ARMS), Gamerman sampling, and independent Metropolis sampling. When there is a normal distribution with a conjugate prior, Gibbs sampling is performed.
- Diagnostic tests include Gelman and Rubin, Geweke, Heidelberger and Welch, and Raftery and Lewis.
- Prior distributions for the regression coefficients include uniform, normal, and Jeffrey's prior.

PHREG PROCEDURE

- PROC PHREG provides Bayesian analysis for Cox regression models with time-independent and time-dependent predictor variables and accommodates all the methods handling ties.
- PROC PHREG also provides Bayesian analysis for piecewise exponential models where you can divide the time axis into sections having its own hazard rate.
- In SAS 9.4, Bayesian frailty models are supported and you can specify the gamma or lognormal distributions for the shared frailty.
- Sampling algorithms include ARMS, random walk Metropolis sampling, and Gibbs sampling when there is conjugacy.

LIFEREG PROCEDURE

- PROC LIFEREG provides Bayesian analysis for parametric location-scale survival models.

- Supported prior distributions are normal and uniform.

BAYESIAN ANALYSIS WITH GENMOD, LIFEREG, PHREG AND FMM

- The **BAYES** statement requests Bayesian analysis.
- A set of standard prior distributions, posterior summary statistics, and convergence diagnostics are provided.
- You can specify **Adaptive rejection, Gamerman** or **Metropolis** sampling algorithms.

SYNTAX FOR THE BAYES STATEMENT

- **BAYES** < options >;
- Options available in all BAYES statements:

INITIAL=	initial values of the chain
NBI=	number of burn-in iterations
NMC=	number of iterations after burn-in
OUTPOST=	output data set for posterior samples
SEED=	random number generator seed
THINNING=	thinning of the Markov chain
DIAGNOSTICS=	convergence diagnostics
PLOTS=	diagnostic plots
SUMMARY=	summary statistics
COEFFPRIOR=	prior for the regression coefficients

THE MCMC PROCEDURE

- *PROC MCMC* is a general purpose simulation procedure that uses Markov chain Monte Carlo (MCMC) techniques to fit a wide range of Bayesian models.

- It requires the specification of a likelihood function for the data and a prior distribution for the parameters.

- It enables you to analyze data that have any likelihood or prior distribution as long as they are programmable using SAS DATA step functions.

PROC MCMC STATEMENTS

- You declare the parameters in the model and assign the starting values for the Markov chain with PARMS statements.
- You specify prior distributions for the parameters with PRIOR statements.
- You specify the likelihood function for the data with the MODEL statement.
- The model specification is similar to PROC NLIN and shares much of the same syntax as PROC NLMIXED.

PROC MCMC SYNTAX

```
PROC MCMC options;
        PARMS parameters and starting values;
        BEGINCNST;
           Programming Statements;
        ENDCNST;
        BEGINNODATA;
           Programming Statements;
        ENDNODATA;
        PRIOR  parameter ~ distribution;
        MODEL variable ~ distribution;
        RANDOM random effects specification;
        PREDDIST <'label'> OUTPRED=SAS-data-set
                    <options>;
RUN;
```

PARMS STATEMENT

- The PARMS statement lists the names of the parameters and specifies optional initial values.
- PROC MCMC generates values for uninitialized parameters from the corresponding prior distributions.
- If the initial values lead to an invalid prior or likelihood calculation, PROC MCMC prints an error message and stops.
- Every parameter in the PARMS statement must have a corresponding prior distribution in the PRIOR statement.

MULTIPLE PARMS STATEMENTS

- When multiple PARMS statements are used, each statement defines a block of parameters.
- PROC MCMC updates parameters in each block sequentially, conditional on the current values of other parameters in other blocks.
- Forming blocks of parameters has its advantages with regard to achieving good mixing of the chains.
- One recommendation is to form small groups of correlated parameters that belong to the same context in the formulation of the model. For example, regression coefficients are in one block and a scale parameter is in a separate block.

PRIOR STATEMENT

- The PRIOR statement is used to specify the prior distribution of the model parameters.
- You must specify a single parameter or a list of parameters, a tilde, and then a distribution with its parameters.
- Multiple PRIOR statements are allowed and you can have as many hierarchical levels as desired.
- A HYPERPRIOR statement is also available to fit a multilevel hierarchical model.

STANDARD DISTRIBUTIONS

beta	binary	binomial	cauchy
chisq	exponential	gamma	geometric
inverse chi-square	inverse gamma	laplace	logistic
lognormal	negative binomial	normal	Pareto
Poisson	scaled inverse chi-square	t-distribution	uniform
wald	weibull	general	dgeneral
Dirichlet	inverse Wishart	multivariate normal	multinomial

MODEL STATEMENT

- The MODEL statement is used to specify the conditional distribution of the data given the parameters (the likelihood function).
- You must specify a single dependent variable or a list of dependent variables, a tilde, and a distribution with its arguments.
- The dependent variables can be either variables from the data set or functions of variables in the program.
- Multiple MODEL statements are allowed for defining models with multiple independent components.

SPECIFYING A NEW DISTRIBUTION

- The GENERAL and DGENERAL functions enable you to analyze data that have any distribution function, as long as these functions are programmable with SAS statements.
- The new distributions have to be specified on the logarithm scale (logarithm of the density must be specified).
- PROC MCMC does not verify that the GENERAL function that you specify is a valid distribution, and you can easily construct prior and log-likelihood functions that lead to improper posterior distributions.

BEGINCNST/ENDCNST STATEMENTS

- These statements define a block within which PROC MCMC processes the programming statements only during the setup stage of the simulation.
- You can use them to define constants or import data set variables into arrays, and to assign initial values to the parameters.
- Using these statements can reduce redundant processing.

BEGINNODATA/ENDNODATA STATEMENTS

- These statements define a block within which PROC MCMC executes the programming statements only twice: at the first and last observation of the data set.
- These statements are best used to reduce unnecessary observation-level computations.
- Any computations that are identical to every observation, such as transformation of parameters, should be enclosed in these statements.
- These statements should not contain data set variables.

RANDOM EFFECTS MODELS

The RANDOM statement is similar to the one in the NLMIXED procedure.

RANDOM *random-effect* ~ *distribution* SUBJECT= *options*;

random-effect is either a univariate or an array of random effects

distribution can be beta, normal, binary, inverse gamma, gamma, Laplace, Poisson, multivariate normal with autoregressive structure, or general distribution.

SUBJECT= identifies the subjects in the model. The variable can be numeric or character, and does not need to be sorted.

PREDDIST STATEMENT

- The PREDDIST statement creates a new SAS data set that contains random samples from the posterior predictive distribution of the response variable.

- The posterior predictive distribution can often be used to check whether the model is consistent with the data.

- The PREDDIST statement works only on response variables that have standard distributions, and it does not support either the GENERAL or DGENERAL functions.

DEMONSTRATION

RESOURCES

The one place for all your SAS Training needs.

support.sas.com/training

It's where you'll find the latest information on:

- New training courses and services
- Special offers and discounts
- The latest course schedules
- New training locations
- Events and conferences
- SAS certification news
- And, much more.

Everything you need – in one place.
Visit and bookmark it today.

Classroom and
Live Web Training

[Bayesian Analysis Using SAS]

FOR FURTHER INFORMATION

Introduction to Bayesian Analysis Procedures:
http://support.sas.com/documentation/cdl/en/statug/66859/HTML/default/viewer.htm#statug_introbayes_toc.htm

The Proc FMM example is documented here:
http://support.sas.com/documentation/cdl/en/statug/66859/HTML/default/viewer.htm#statug_fmm_gettingstarted02.htm

The Proc MCMC examples all come from this paper:
http://support.sas.com/resources/papers/proceedings09/257-2009.pdf

Our Bayesian Analysis Using SAS/STAT landing page (and links within) is really helpful:
http://support.sas.com/rnd/app/da/Bayesian/index.html

Bayesian Analysis Using SAS
https://support.sas.com/edu/schedules.html?ctry=us&id=2047

Connect with me:
LinkedIn: https://www.linkedin.com/in/melodierush
Twitter: @Melodie_Rush

QUESTIONS?

Thank you for your time and attention!

sas | THE POWER TO KNOW.

CUSTOMER LOYALTY TEAM • Support You Can Count On

sas.com