PhUSE 2013

Paper SP02

Bayesian statistics: concept and Bayesian capabilities in SAS

Mark Janssens, I-BioStat, Hasselt University, Belgium

ABSTRACT

The use of Bayesian statistics has risen rapidly in the industry, and software for Bayesian analysis has become widely available.

This paper outlines what Bayesian statistics is about, and shows how SAS implemented Bayesian capabilities into some of the procedures.

We will walk through an example using the procedures GENMOD (normal and binomial model) and GLIMMIX (binomial model with random effects).

We will also cover the procedure MCMC (for the same models) since the MCMC procedure provides additional functionality and improved performance in SAS/STAT 12.1 (SAS 9.3).

The aim of the paper is to ease the way from a “classical” procedure towards the use of the “Bayesian” procedure MCMC.

THE CONCEPT OF BAYESIAN STATISTICS

Bayesian methods are used to compute a probability distribution of parameters in a statistical model, using observed data as well as existing knowledge about these parameters.

Bayesian methods differ from classical statistics when it comes to the meaning of probability and the use of prior evidence. The difference can be understood through a simple example:

Meaning of probability

In a clinical setting, let the parameter θ be the parameter of interest, θ=0 our null hypothesis, and X the data which was collected to draw a conclusion about θ. The parameter θ could for example be the effect of an experimental treatment versus control.

In a classical analysis, θ is considered fixed, and X random. More formally, we estimate P(X | θ=0). The interpretation of the p-value illustrates the idea that X is random, since the interpretation goes as follows: the p-value tells us how likely the observed data are if I *repeat* my experiment (i.e. X) many times under a *fixed* null hypothesis (i.e. θ=0). So in classical inference, X is the element which varies.

The classical paradigm, however, is not suited to answer *all* questions at hand, such as: what is the probability that θ is bigger that some clinically relevant value C? This cannot be answered when θ is fixed. Instead, we would need a distribution for θ, for we now make a claim regarding P(θ | X).

This being said, the two paradigms are not totally distinct. The method of maximum likelihood is close to Bayesian estimation with noninformative priors. The maximum likelihood procedures in SAS make use of this connection and are able to provide a posterior sample for θ, as shown by example further below.

Prior evidence

The maximum likelihood procedures in SAS, when providing such a posterior sample for θ, by default use noninformative priors. This means that posterior sampling gives a distribution of θ, conditioning on the observed data and nothing more. A next step could be to take prior evidence regarding θ into account, at least if some prior knowledge is available. In the area of clinical development, previous trials are often a source of such prior knowledge.

Conjugate prior

When the posterior distribution $f(\theta \mid X, \theta_0)$ is in the same family as the prior distribution $f(\theta_0)$, then the prior and posterior are called conjugate distributions. In the absence of numerical methods such as MCMC, conjugacy is essential because it allows to analytically compute the parameters of the posterior distribution. In practice, most Bayesian models are not solved analytically. Instead, numerical methods are used.

This being said, within the numerical SAS procedure PROC MCMC, conjugacy is still important: when PROC MCMC detects conjugacy, efficient conjugate sampling methods are used to draw conditional posterior samples (Chen 2011). Following combinations lead to conjugate sampling in PROC MCMC (Table 1):

Table 1: Model distributions leading to conjugate sampling in PROC MCMC

Model distribution	Parameter	Prior Distribution
Normal with known μ	σ^2	Inverse gamma
Normal with known μ	τ	Gamma
Normal with known scale (σ^2 or $\tau = 1/\sigma^2$)	μ	Normal
Multivariate normal with known Σ	**μ**	Multivariate normal
Multivariate normal with known **μ**	Σ	Inverse Wishart
Multinomial	**p**	Drichelet

Binomial/Binary	p	Beta
Possion	λ	Gamma

Strong/Weak prior

Data and prior evidence can be regarded as the result of two stochastic processes, and the posterior distribution as the combination of these two competing processes.

If the prior evidence is strong, then the prior distribution will clearly affect the posterior distribution (Figure 1a). For example, when doing a limited experimental trial using a well-studied control drug, the established effect of the well-studied drug could serve as a strong prior in the statistical model. In that case, the posterior distribution involving the control drug will heavily rely on the established prior evidence.

On the other hand, if the prior evidence is weak, then the prior distribution will make little difference, and the posterior distribution will almost coincide with the data (Figure 1b). The posterior estimates will be close to the maximum likelihood estimates.

Figure 1a: Bayesian model with strong prior

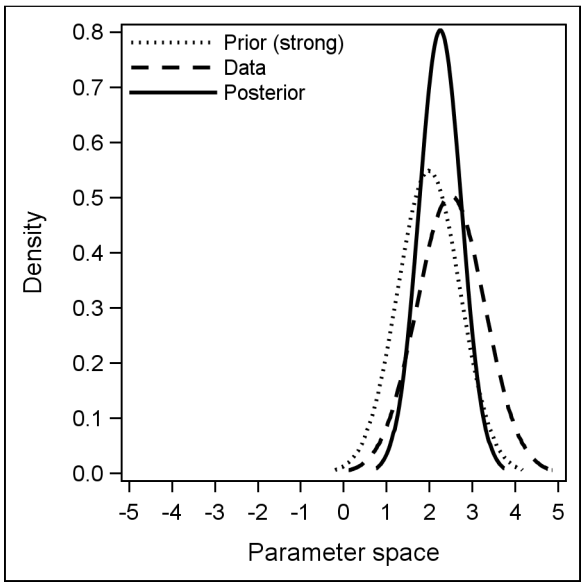
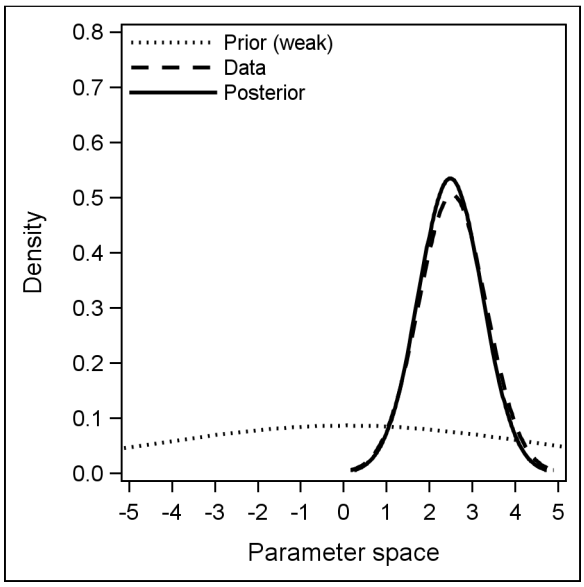


Figure 1b: Bayesian model with weak prior



Different wording is used for “very weak” priors, such as: noninformative, vague, or diffuse priors. The classification and use of priors is a relevant body of work in itself, but beyond the scope of this paper. The choice of –even very weak– priors is . especially important in generalized linear and nonlinear models, because the prior distribution is typically not invariant to transformations (a notable exception is Jeffreys prior).

THE DATA

For the sake of ease, a simple instructive data set is used: the growth data, introduced by Potthoff & Roy in 1964, and used by several textbook authors thereafter (e.g. Little & Rubin 1987, Verbeke & Molenberghs 2000, SAS/STAT 9.22 User's Guide).

The growth data contain dental growth measures for 11 girls and 16 boys. The dental measurements are taken at age 8, 10, 12, and 14.

Figure 1 shows the individual and mean profiles, for girls and boys respectively.

The individual profiles are gray, the mean profiles are colored bold-black. Some profiles look unusual. One the most unusual profiles is plotted in black.

One can easily discern the hierarchical structure of the data: individuals starting low tend to stay low, whereas individuals starting high tend to maintain high values.

On average, both girls and boys grow as time goes by. For girls, the average growth is very linear. For boys, the growth curve seems to take off at the age of 10.

Let us define “normal” minimal growth between the age of 8 and 14 as a growth increase of at least 10%. This is an arbitrary choice. This extra definition will yield a binary response measure, and enable us to fit a binomial model later on.

Figure 2: growth data, profiles

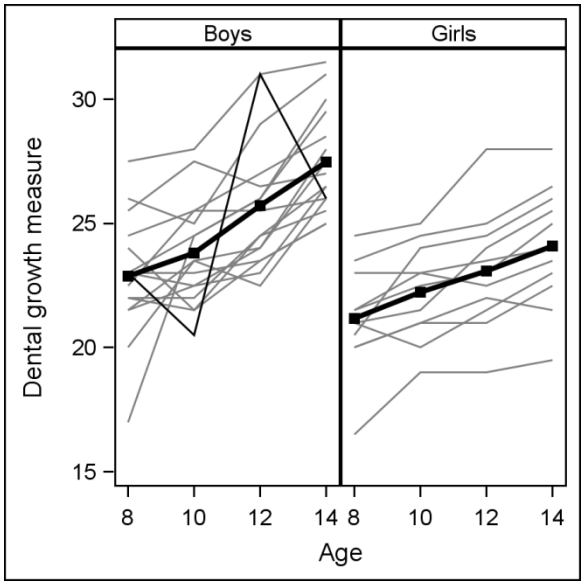


Table 2: growth data, response

Response = growth increase of at least 10%

	Girls	Boys
Response=Yes	7 (64%)	13 (81%)
Response=No	4 (36%)	3 (19%)
	11 (100%)	16 (100%)

The following two questions will be addressed in the light of these data:

1. Is the *change from baseline* different between boys and girls?
2. Is the *response* different between boys and girls?

STATISTICAL MODEL 1 – LINEAR REGRESSION

The first statistical model addresses research question n° 1: Is the change from baseline different between boys and girls? Formally:

$Y \sim \text{normal}(\mu; \sigma^2)$

where Y is the continuous growth measure at age 14, μ the linear predictor, and σ^2 the residual variance. Y and μ are on the same linear scale with

$\mu = \beta_0 + \beta_1 \text{ YBASE} + \beta_2 \text{ BOY}.$

Direct likelihood

The above model can be estimated with PROC GLM, PROC MIXED, or with PROC GENMOD as we will do here:

```
proc genmod data=PERM.ANALYSIS_SET;
 where AGE=14;
 model Y = YBASE BOY / dist=normal;
run;
```

Later in PROC MCMC, we cannot use class level variables, and we will need to use dummy variables instead. We therefore used the dummy variable BOY (values: 1, 0) in the coding solutions of both PROC GENMOD and PROC MCMC.

The estimate of regression parameter β_2 ("BOY" in Output 1) is 2.53 and highly significant.

Output 1: PROC GENMOD – linear regression model – direct likelihood

Analysis Of Maximum Likelihood Parameter Estimates

Parameter	*DF*	*Estimate*	*Standard Error*	*Wald 95% Confidence Limits*		*Wald Chi-Square*	*Pr > ChiSq*
Intercept	1	13.4902	3.3826	6.8604	20.1201	15.90	<.0001
YBASE	1	0.5005	0.1575	0.1917	0.8093	10.09	0.0015
BOY	1	2.5305	0.7660	1.0292	4.0317	10.91	0.0010
Scale	1	1.8332	0.2495	1.4040	2.3935		

Hence, based on these data, boys and girls are different, and the estimated difference is about 2.5.

Suppose that the gender difference was a known effect, and that the size of the effect was equal to 2 (not 2.5). How probable is the situation that β_2 actually equals 2,

and that the current finding of 2.5 results from sampling error or poor data quality? This question is a “Bayesian” question and will now be addressed.

Bayesian likelihood

We now step to the Bayesian likelihood, since we like to obtain a sampling distribution for the parameter of interest, i.e. β_2.

There are two ways to get there. The first option is to make use of the BAYES statement within PROC GENMOD. The second option is to use PROC MCMC.

The PROC GENMOD option is straightforward: the BAYES statement is added, the remaining syntax is left unchanged.

```
proc genmod data=PERM.ANALYSIS_SET;
 where AGE=14;
 model Y = YBASE BOY / dist=normal;
 bayes nbi=1000 nmc=10000 thin=2 seed=159 cprior=jeffreys out=posterior;
run;
```

We split the Bayesian output –conceptually– in two blocks: diagnostic information and posterior information. We will briefly discuss the diagnostic information for the β_1 parameter of this model. In practice, the diagnostics of all model parameters need to be inspected.

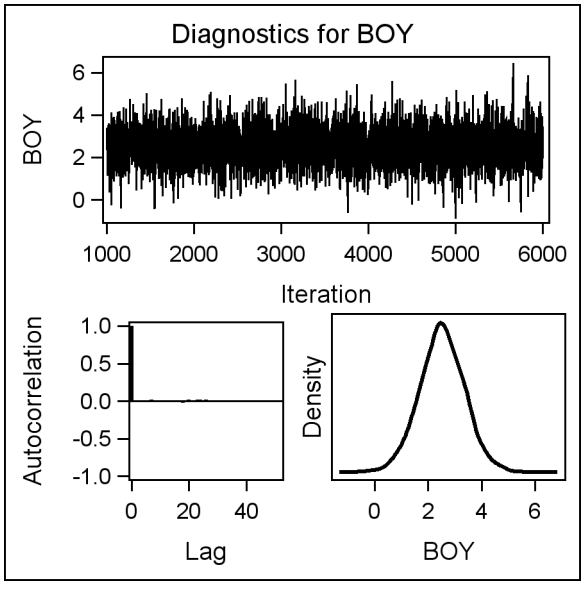
Diagnostic information:

In line with the Monte Carlo Standard Errors and the Geweke Diagnostics (not shown), the trace plot and the autocorrelation plot (Figure 3) show that the Markov chain is stationary and efficiently explores all areas of the posterior distribution.

In other words, subsequent MCMC iterations produce parameter estimates which quickly jump from the mode to the tail of the –stable– posterior distribution.

In case of high autocorrelation (poor mixing), the posterior distribution is not adequately explored, and inference from that posterior distribution becomes problematic.

Figure 3: Trace plot (top), autocorrelation plot (left), and posterior distribution (right) for β_2



Autocorrelation is not problematic in the statistical model at hand. In case of an autocorrelation issue, it could be handled in several ways, such as: center the data variables, thin the chain, block the model parameters, and/or reparameterize the model.

Posterior information:

The posterior sample for β_2 is nicely centered around the maximum likelihood value of 2.53 (Figure 3 and Output 2). The posterior intervals (or credible intervals) provided by SAS are "Equal-Tail" and "HPD" (Highest Posterior Density). Equal-tail intervals are easy to construct: if alpha is set to 5%, then the equal-tail limits are simply the 2.5% and 97.5% percentiles of the estimated posterior distribution. To understand HPD, one should think of several candidate sets {lower limit, upper limit} which all represent 95% of the posterior density. From these candidate sets, the one with the smallest range is chosen. HPD intervals are always smaller (or equal to) equal-tail intervals. For a symmetric posterior, such as β_2, the HPD interval will not deviate much from the equal-tail interval.

Output 2: PROC GENMOD – linear regression model – Bayesian likelihood

Posterior Summaries

Parameter	N	Mean	Standard Deviation	Percentiles 25%	50%	75%
INTERCEPT	5000	13.5349	3.8101	11.0107	13.5875	16.0522
YBASE	5000	0.4985	0.1776	0.3807	0.4975	0.6140
BOY	5000	2.5258	0.8506	1.9806	2.5178	3.0903
Dispersion	5000	4.1222	1.3183	3.1888	3.8622	4.7274

Posterior Intervals

Parameter	Alpha	Equal-Tail Interval		HPD Interval	
INTERCEPT	0.050	5.9257	21.0660	6.3736	21.3969
YBASE	0.050	0.1612	0.8544	0.1565	0.8471
BOY	0.050	0.8391	4.2064	0.8702	4.2139
Dispersion	0.050	2.2947	7.4673	2.0738	6.8210

The second option to fit the Bayesian likelihood, was to use PROC MCMC, requiring the following syntax:

```
proc mcmc data=PERM.ANALYSIS_SET nbi=1000 nmc=10000 thin=2 seed=159
          monitor=(beta0-beta2 sigma2 beta2_gt_2);
 where AGE=14;

 1  parms beta0 13.49 beta1 0.50 beta2 2.53;
    parms sigma2 3.36;

 2  prior beta0-beta2 ~ normal(mean = 0, var = 1000);
    prior sigma2 ~ igamma(shape = 0.001, scale = 0.001);
```

```
3   mu = beta0 + beta1*YBASE + beta2*BOY;
    model Y ~ normal(mean = mu, var = sigma2);

4   beta2_gt_2 = beta2 > 2;

run;
```

1 The PARM statements contain the same model parameters as in PROC GENMOD, apart from the residual variance parameter (σ^2 in PROC MCMC, "dispersion" in PROC GENMOD) which is implicit in PROC GENMOD. The initial values for the parameters are the maximum likelihood estimates (see Output 1). For a simple model, the choice of initial values usually is not a prime concern, but is good practice to think about initial values carefully. We specify two PARM statements to control the "blocking" in PROC MCMC: the regression coefficients **β** are estimated in one block (given the estimate for σ^2), and in the same way σ^2 is updated (i.e. given the values for **β**).

2 In the PRIOR statements, a diffuse normal distribution is chosen for the regression coefficients, and a diffuse inverse gamma for the residual variance. As a matter of fact, note that the choice of prior distributions is different between the PROC MCMC and the (default) PROC GENMOD implementation. In PROC GENMOD, Jeffreys priors were used. In PROC MCMC, we have used a different set of diffuse priors. For this linear model, there is no difference in results.

There is no PARM or PRIOR statement in the coding solution of PROC GENMOD. Numerical procedures force us to be explicit about the parameters in the model and their initial values (PARM statement). Additionally, in case of Bayesian model fitting, all these model parameters need to have a prior distribution (PRIOR statement). The PARM and PRIOR statement need to be fully in sync. If a parameter shows up in the PARM statement but not in the PRIOR statement, SAS will issue an error ("The symbol needs to be specified in PRIOR statement"). In the reverse case, SAS will not proceed either ("The symbol is not declared in a PARMS statement").

3 The MODEL statement together with the specification of the linear predictor (μ) are in essence no different from the model specification in PROC GENMOD. A coding difference is that class level variables cannot be used in PROC MCMC directly, and dummy variables need to be specified instead.

In practice, a Bayesian model is often coded from step 3 up to step 1. First, the statistical model is spelled out. Then, the model parameters are given sensible prior distributions. And finally, initial values are set.

4 The last statement in the above PROC MCMC is optional but convenient. It was added to easily answer the question about the probability of β_2 being greater than 2.

The PROC MCMC results are –as expected– almost identical to the Bayesian likelihood of PROC GENMOD. The probability that the gender difference is at least 2, based on the current data alone, is 73% (Output 3).

Output 3: PROC MCMC – linear regression model – Bayesian likelihood

Posterior Summaries

Parameter	N	Mean	Standard Deviation	Percentiles 25%	50%	75%
BETA0	5000	13.3587	3.5826	10.8873	13.3665	15.8217
BETA1	5000	0.5074	0.1666	0.3965	0.5112	0.6205
BETA2	5000	2.5018	0.8250	1.9534	2.4941	3.0815
SIGMA2	5000	4.0658	1.2536	3.2020	3.8352	4.6738
beta2_gt_2	5000	0.7326	0.4426	0	1.0000	1.0000

Bayesian likelihood incorporating prior evidence

Until now, no prior evidence regarding the gender difference is included in the statistical model. At this point, we use the knowledge about the size of the gender effect. The gender effect was known to lie around 2.

The PROC MCMC syntax changes only slightly:

```
proc mcmc data=PERM.ANALYSIS_SET nbi=1000 nmc=10000 thin=2 seed=159
          monitor=(beta0-beta2 sigma2 beta2_gt_2);
 where AGE=14;
 parms beta0 13.49 beta1 0.50 beta2 2.53;
 parms sigma2 3.36;
 prior beta0-beta1 ~ normal(mean = 0, var = 1000);
 prior beta2 ~ normal (mean = 2, var = 0.5);
 prior sigma2 ~ igamma(shape = 0.001, scale = 0.001);
 mu = beta0 + beta1*YBASE + beta2*BOY;
 model Y ~ normal(mean = mu, var = sigma2);
 beta2_gt_2 = beta2 > 2;
run;
```

The highlighted line of code shows that the prior evidence is centered around 2 (“mean = 2”). This prior distribution, as well as the data likelihood and resulting posterior, are plotted in Figure 1a (see above). The posterior estimate for β_2 being greater than 2 should be smaller than 73%, given the conservative prior. The posterior estimate turns out to be 67% (Output 4).

Output 4: PROC MCMC – linear regression model – Bayesian likelihood with strong prior

Posterior Summaries

Parameter	N	Mean	Standard Deviation	Percentiles 25%	50%	75%
BETA0	5000	12.9850	3.6821	10.4937	12.9318	15.4242

BETA1	5000	0.5306	0.1677	0.4204	0.5312	0.6435
BETA2	5000	2.2422	0.5540	1.8716	2.2367	2.6207
SIGMA2	5000	4.0383	1.2820	3.1455	3.8103	4.6395
beta2_gt_2	5000	0.6740	0.4688	0	1.0000	1.0000

STATISTICAL MODEL 2 – LOGISTIC REGRESSION

The second statistical model addresses research question n° 2: Is the response different between boys and girls? Formally:

Y_{BIN} ~ bernoulli(π)

where Y_{BIN} equals 1 (response=Yes) or 0 (response=No), and π is the response probability.

On the linear scale,

logit (π) = log[π / (1 - π)] = $\beta_0 + \beta_1$ BOY

Direct likelihood

The above model can be estimated with PROC LOGISTIC or with PROC GENMOD, with very similar syntax. We stick to PROC GENMOD:

```
proc genmod data=PERM.ANALYSIS_SET descending;
 where AGE=14;
 model YBIN = BOY / dist=bin;
run;
```

The odds ratio boys vs girls is 2.47 ($e^{0.9067}$ = 2.47) yet this gender effect is not significant (Output 5).

Output 5: PROC GENMOD – logistic regression model – direct likelihood

Analysis Of Maximum Likelihood Parameter Estimates

Parameter	*DF*	*Estimate*	*Standard Error*	*Wald 95% Confidence Limits*		*Wald Chi-Square*	*Pr > ChiSq*
Intercept	1	0.5596	0.6268	-0.6689	1.7881	0.80	0.3719
BOY	1	0.9067	0.8962	-0.8497	2.6632	1.02	0.3116
Scale	0	1.0000	0.0000	1.0000	1.0000		

Bayesian likelihood

The nonsignificance in the logistic model may seem surprising because –in Model 1– the average growth differed significantly between boys and girls (p = 0.001), and the effect size (Δ = 2.53) was larger than the previously discovered effect size (Δ = 2). This being said, the power for a binary variable is much lower compared to a continuous outcome, and the data set contains only 27 subjects.

The nonsignificance does not preclude us to calculate the probability that the odds ratio is greater than 1. Given the nonsignificance (at α=0.05), we already know that the probability P[odds ratio > 1] will be lower than 95%, but let us calculate the exact value:

```
proc genmod data=PERM.ANALYSIS_SET descending;
 where AGE=14;
 model YBIN = I BOY / noint dist=bin;
 bayes nbi=1000 nmc=10000 thin=2 seed=159 out=posterior;
run;
proc sql;
 select count(*)/5000 as boy_gt_1 from posterior where exp(boy) > 1;
quit;
```

Despite the nonsignificance, we see that the probability of an odds ratio above 1 is equal to 84% (Output 6).

Output 6: PROC GENMOD – logistic regression model – Bayesian likelihood

Posterior Summaries

Parameter	N	Mean	Standard Deviation	Percentiles 25%	50%	75%
INTERCEPT	5000	0.5541	0.6221	0.1239	0.5423	0.9631
BOY	5000	0.9146	0.8929	0.3041	0.8987	1.4911

boy_gt_1
0.8466

The step from maximum likelihood to Bayesian likelihood is done within PROC GENMOD itself. The Bayesian likelihood could have been fitted with PROC MCMC too. However, to replicate the PROC GENMOD results, the Jeffreys prior should be constructed in PROC MCMC. This involves extra programming steps within PROC MCMC and exceeds the scope of this paper. The construction of the Jeffreys prior for a logistic regression model is a worked example in the in the SAS/STAT 9.3 documentation (SAS/STAT 9.3 User's Guide , PROC MCMC, Example 54.4).

STATISTICAL MODEL 3 – RANDOM EFFECTS LOGISTIC REGRESSION

The third statistical model revisits research question n° 2: Is the response different between boys and girls? Suppose that researchers were surprised to observe a gender difference in change from baseline (Model 1), but not in response (Model 2), and that they therefore decided to collect additional data. Presume that response data were collected in 10 other centers. Center is a contextual variable possibly affecting outcome, and needs to be included in the model. The center variable has 11 levels (1 from the original setup, plus 10 replications) and will be modeled as a random effect rather than as a fixed effect.

The sample size of the 11 experiments varied from 27 to 50, so no single experiment dominates the pooled estimate.

The data over all experiments looked as follows (fragment):

CENTER	BOY	YYES	N
1	1	13	16
1	0	7	11
2	1	11	20
2	0	9	15
3	1	14	22
3	0	12	22
…	…	…	…
…	…	…	…
10	1	17	22
10	0	15	23
11	1	22	27
11	0	18	28

The model specification now is:

$Y_{YES} \sim$ binomial(n, π)

where Y_{YES} represents the number of observations with response=Yes, where n is the total number of observations within each experiment, and π is the response probability.

On the linear scale,

$\text{logit}(\pi) = \log[\pi / (1 - \pi)] = \beta_0 + b_{0i} + \beta_1 \text{ BOY}$

where index i refers to the 11 centers and b_{0i} stems from an underlying normal distribution $b_0 \sim \text{normal}(\mu; \sigma^2)$.

Direct likelihood

The above model can be estimated with PROC NLMIXED, or as follows with PROC GLIMMIX:

```
proc glimmix data=PERM.ANALYSIS_SETS method=quad;
```

```
 class CENTER;
 model YYES/N = BOY / dist=bin solution cl;
 random intercept / subject=CENTER solution cl;
run;
```

The odds ratio boys vs girls is 1.75 ($e^{0.5623}$ = 1.75). The odds ratio is significant (p=0.02), yet it is lower than the odds ratio based on experiment 1 alone (2.47, see Model 2).

Output 7: PROC GLIMMIX – random effects logistic regression – direct likelihood

Solutions for Fixed Effects

Effect	*Estimate*	*Standard Error*	*DF*	*t Value*	*Pr > |t|*	*Alpha*	*Lower*	*Upper*
Intercept	0.03913	0.1428	10	0.27	0.7896	0.05	-0.2790	0.3573
BOY	0.5623	0.1938	10	2.90	0.0158	0.05	0.1304	0.9942

Bayesian likelihood

The RANDOM statement in PROC MCMC is a new feature of SAS/STAT 12.1 (SAS 9.3). Random effect parameters share the same prior distribution and are independent of each other. The syntax of the RANDOM statement in PROC MCMC is similar to the RANDOM statement in PROC NLMIXED. In case of multilevel data, several RANDOM statements can be specified, each with a level specific prior distribution. With the RANDOM statement of MCMC, hierarchical Bayesian models can be now be fitted in a flexible and efficient way.

The Bayesian likelihood for Model 3 can be coded in PROC MCMC in the following way:

```
proc mcmc data=PERM.ANALYSIS_SETS nbi=1000 nmc=10000 thin=2 seed=159
outpost=posterior
          monitor=(beta0-beta1 sigma2 or pooled) statistics=(summary
intervals);
 parms beta0-beta1 0;
 parms sigma2 1;
 prior beta0-beta1 ~ normal(mean = 0, var = 1000);
 prior sigma2 ~ igamma(shape = 0.001, scale = 0.001);
 random b0 ~ normal(mean = 0, var=sigma2) subject=CENTER;
 eta = beta0 + b0 + beta1*BOY;
 pi = logistic(eta);
 model YYES ~ binomial(n = N, p = pi);
 array or[11];
 or[CENTER]=exp(b0 + beta1);
 pooled=exp(beta1);
run;
```

In terms of syntax, the main differences with Model 1 are:

1. The outcome Y_{YES} is binomial and no longer continuous, so the MODEL statement now contains the binomial distribution (~ binomal).
2. In the normal model, the distributional parameter μ is a linear function of covariates. In the binomial model, η is the linear function of covariates, and η is tied to the distributional parameter π through a logit link (π = logistic[η]).
3. The RANDOM statement specifies that b_0 is a random center-specific regression parameter.
4. With the last statements we calculate the center-specific odds ratios ("or[CENTER]" and the overall odds ratio ("pooled"). The odds ratios are listed in Output 8 and visualized in Figure 4.

Output 8: PROC GLIMMIX – random effects logistic regression – Bayesian likelihood

Posterior Summaries

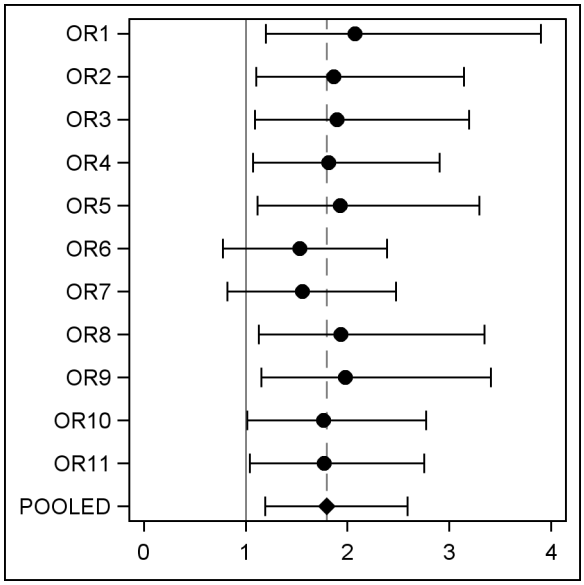
Parameter	N	Mean	Standard Deviation	Percentiles 25%	50%	75%
BETA0	5000	0.0439	0.1493	-0.0492	0.0369	0.1413
BETA1	5000	0.5658	0.1945	0.4352	0.5645	0.6994
SIGMA2	5000	0.0513	0.0787	0.00492	0.0215	0.0660
or1	5000	2.0760	0.6823	1.6346	1.9345	2.3313
or2	5000	1.8679	0.5149	1.5192	1.7859	2.1084
or3	5000	1.8940	0.5246	1.5360	1.8223	2.1513
or4	5000	1.8154	0.4693	1.4950	1.7463	2.0625
or5	5000	1.9261	0.5526	1.5583	1.8318	2.1791
or6	5000	1.5337	0.4196	1.2422	1.5130	1.8124
or7	5000	1.5586	0.4185	1.2695	1.5399	1.8186
or8	5000	1.9347	0.5670	1.5591	1.8413	2.1813
or9	5000	1.9775	0.5666	1.5998	1.8786	2.2344
or10	5000	1.7614	0.4545	1.4628	1.7146	2.0054
or11	5000	1.7696	0.4400	1.4730	1.7190	2.0095
pooled	5000	1.7945	0.3521	1.5453	1.7586	2.0125

The pooled odds ratio estimate equals 1.79 and is close to the direct likelihood estimate (OR=1.75). The individual estimates ("OR1" to "OR11") dither around the overall value of 1.79.

The OR estimate for experiment 1 is equal to 2.07. This is lower compared to Model 2 which was fed by the data from experiment 1 alone (OR=2.47).

This difference is an expected artifact of random effects modeling, known as "shrinkage". Shrinkage means that, for every level of the random effect, the odds ratio is a weighted combination of the level-specific estimate and the overall estimate. Shrinkage implies that the extreme, individual odds ratios are pulled towards the overall estimate.

Figure 4: Odds Ratio (OR) and 95% credible interval for Model 3



CONCLUDING REMARKS

This paper shows that maximum likelihood procedures such as PROC GENMOD provide readily available Bayesian functionality. More advanced statistical models can be fitted with PROC MCMC. With the introduction of the RANDOM statement in PROC MCMC, Bayesian random effect models have become easy to specify & run. For a linear mixed model, Bayesian inference could be obtained using either PROC MIXED or PROC MCMC. There is no BAYES statement in PROC GLIMMIX, so for a Bayesian binomial mixed model, PROC MCMC is the only coding option. Although not overly dealt with in this paper, Bayesian model fitting requires careful inspection of the model diagnostics, and advanced models require in-depth understanding of prior distributions (choice, construction, operational characteristics).

REFERENCES

Adamina M, Tomlinson G, Guller U. "Bayesian Statistics in Oncology: A Guide for the Clinical Investigator". *Cancer*, 2009, Volume 115, Issue 23, 5371-5381. DOI: 10.1002/cncr.24628.

Chen F(SAS Institute). "Bayesian Modeling Using the MCMC Procedure". SAS Global Forum 2009, Paper 257-2009.

Chen F (SAS Institute). "The RANDOM Statement and More: Moving On with PROC MCMC", SAS Global Forum 2011, Paper 334-2011.

Potthoff RF, Roy SN. "A Generalized Multivariate Analysis of Variance Model Useful Especially for Growth Curve Problems". *Biometrika*, 1964, Volume 51, Issue 3/4 (December), 313-326.

SAS Institute. SAS/STAT 9.22 User's Guide. SAS Publishing, 2010.

SAS Institute. SAS/STAT 9.3 User's Guide. SAS Publishing, 2011.

Verbeke G, Molenberghs G. "Linear Mixed Models for Longitudinal Data". Springer, 2000.