Paper 496-2013

Predicting Health Care Expenditures with the MCMC Procedure

Greg Watson, UCLA Center for Health Policy Research, Los Angeles, CAS

ABSTRACT

Substantial variation, excess zeros, skew and extreme outliers make fitting and predicting health care expenditures rather difficult. This paper presents a Bayesian model that uses the first year of the fourteenth panel (2009-2010) of the nationally representative Medical Expenditures Panel Survey (MEPS) to predict health care expenditures for individuals in the second year. The merits of a Bayesian approach are examined and compared to classical alternatives. Implementation in the MCMC procedure is presented in detail, and model diagnostics and validation are discussed.

INTRODUCTION

The MCMC procedure is a powerful platform for Bayesian analysis that fits a wide variety of Bayesian models. This paper gives a brief introduction to Bayesian methods, introduces the basic syntax of the MCMC procedure, and illustrates its use in fitting a zero-inflated gamma regression model to health care expenditure data from the Medical Expenditure Panel Survey (MEPS).

BAYESIAN ANALYSIS

Classical, i.e., frequentist, statistics assigns probability distributions to observed data but not parameters which are regarded as fixed, unknown constants. Within this framework, maximum likelihood estimation of the parameters of a model is used to find the parameter values that maximize the likelihood function $L(Y|\theta)$, which specifies the probability of observed data Y given model parameters θ, i.e., $\hat{\theta}_{mle} = \max(L(Y|\theta))$.

Bayesian statistics also employs the likelihood function (or rather the sampling density, which is proportional to the likelihood) to estimate model parameters from the data, but it also incorporates prior information on the parameters using probability distributions. When the prior information is negligible, Bayesian and maximum likelihood estimates will be very similar. From a Bayesian perspective, probability is subjective and used to represent uncertainty. Consequently it is licit to place a probability distribution on parameters, which are still acknowledged to be fixed but unknown. The prior uncertainty about model parameters is described with a prior distribution $f(\theta)$ that is updated by the data into a posterior distribution $f(\theta|Y)$ based on an application of Bayes' rule:

$$f(\theta|Y) = \frac{f(Y|\theta)f(\theta)}{f(Y)}.$$

Since Bayesian statistics treats observed data as fixed, $f(Y)$ amounts to a normalizing constant that multiplies $f(Y|\theta)f(\theta)$.

The posterior distribution $f(\theta|Y)$ represents the uncertainty about parameter(s) θ given the data Y. In simple cases this posterior distribution may be derived analytically, but typically Bayesian inference proceeds through numerical methods. Markov chain Monte Carlo (MCMC) allows samples to be drawn directly from this posterior distribution. Each iteration of the chain produces values for the parameters of the model, denoted $\theta^{(l)}$ at the l^{th} iteration. These values are used to estimate posterior summaries. For example, their mean converges to the posterior mean as n gets large, so that

$$\frac{1}{n}\sum_{l=1}^{n} \theta^{(l)} \approx \mathrm{E}(\theta|Y).$$

Posterior quantiles such as the 95% posterior interval formed by the 2.5th and 97.5th quantiles, may be obtained similarly and are useful in drawing inference without relying on asymptotic approximations. In addition, these interval estimates have more intuitive interpretations than their frequentist counterpart, the confidence interval. For example, there is probability 0.95 that the true value of a parameter θ falls within its 95% posterior interval.

THE MCMC PROCEDURE

The MCMC Procedure employs Markov chain Monte Carlo to fit a wide variety of Bayesian models. This section illustrates the basic syntax of the procedure on a simple linear regression. The example uses the “cars” data set that ships with SAS® and is a linear regression of highway miles per gallon (MPG) on vehicle weight:

$$MPG_i = \beta_0 + \beta_1 weight + \epsilon_i,$$
$$\epsilon_i \text{ iid } \mathrm{N}(0, \sigma^2).$$

Maximum likelihood estimates for this model are easily produced using the MIXED procedure:

```
proc mixed data=sashelp.cars method=ml;
    model mpg_highway = weight / solution;
quit;
```

Parameter Estimates					
Variable	**Estimate**	**Standard Error**	**DF**	**t Value**	**Pr > |t|**
Intercept	48.2515	0.8182	426	58.97	<.0001
Weight	-0.00598	0.000224	426	-26.75	<.0001

Syntax

The MCMC procedure may also be used to fit this model within a Bayesian framework. In addition to the likelihood derived from the linear function defined above, a Bayesian model requires a prior distribution be specified for the model parameters. In this case there are three parameters: β_0, β_1 and σ^2. Choosing uninformative priors will yield results very similar to the PROC MIXED output above. Consider independent normal priors for β_0 and β_1 centered at zero with a very large variance, and an inverse gamma prior for σ^2 with a large variance. The inverse gamma distribution is well suited as a prior for variance parameters, since it is nonnegative. This yields the following Bayesian model,

$$\begin{aligned} MPG_i =&\ \beta_0 + \beta_1 weight + \epsilon_i, \\ \epsilon_i &\sim \mathrm{N}(0,\ \sigma^2), \\ \beta_0 &\sim \mathrm{N}(0,\ 10^6), \\ \beta_1 &\sim \mathrm{N}(0,\ 10^6), \\ \sigma^2 &\sim \mathrm{InvGamma}(0.001, 0.001). \end{aligned}$$

The following call to the MCMC procedure fits the model.

```
proc mcmc data=sashelp.cars;
    /* declare parameters */
    parms beta0 beta1 sigma2;
    /* prior */
    prior beta0 ~ normal(0, var=10**6);
    prior beta1 ~ normal(0, var=10**6);
    prior sigma2 ~ igamma(.001, scale=.001);
    /* likelihood */
    model mpg_highway ~ normal(beta0 + beta1*weight, var=sigma2);
run;
```

The resulting posterior summaries are very similar to the parameter estimates produced by PROC MIXED above. Since the posterior is (in this case) a weighted average of the prior and the data, with an uninformative prior the posterior is almost entirely determined by the data.

Posterior Summaries						
			Standard	**Percentiles**		
Parameter	**N**	**Mean**	**Deviation**	**25%**	**50%**	**75%**
beta0	1000	48.1918	0.8508	47.7238	48.2513	48.7198
beta1	1000	-0.006	0.00023	-0.0061	-0.006	-0.0058
sigma2	1000	12.421	0.8597	11.8368	12.3965	12.9912

Diagnostic Plots

By default, PROC MCMC displays three diagnostic plots for each parameter that are helpful in assessing MCMC convergence. The Markov chain converges to the posterior distribution in the limit, but with a finite sample there is no guarantee that the chain has converged sufficiently to the posterior distribution. Figure 1 depicts the diagnostic plots that PROC MCMC produced for β_0.

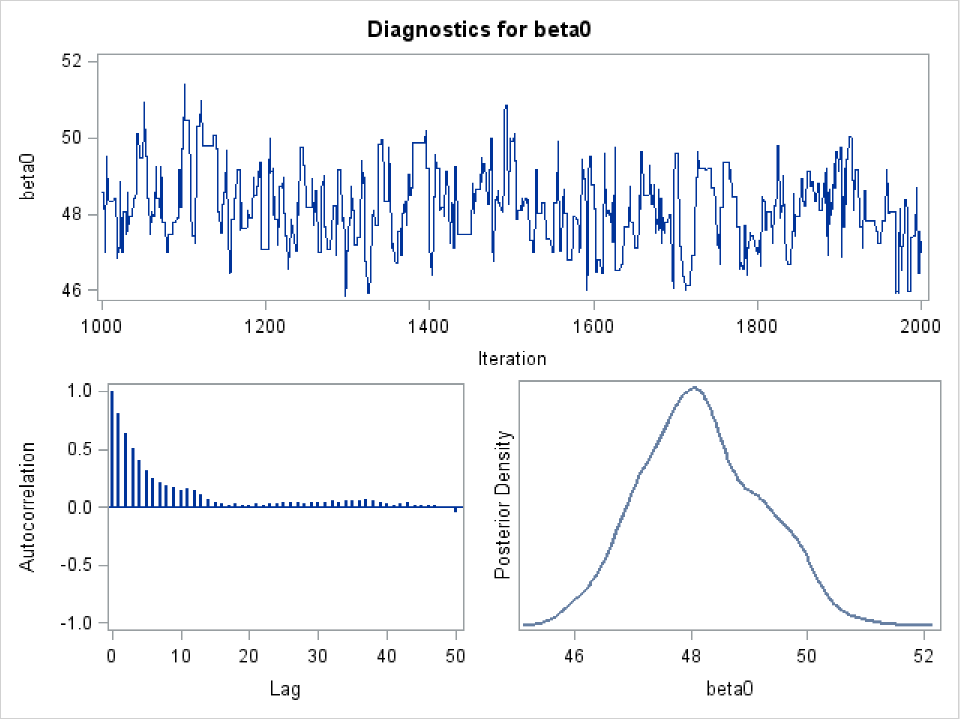


Figure 1. Diagnostic Plots for β_0

The first diagnostic plot is commonly referred to as a trace plot and shows the values of the parameter β_0 at each iteration of the Markov chain. The plot displays iterations 1,001 through 2,000. By default, PROC MCMC keeps 1,000 iterations from the Markov chain after discarding an initial 1,000 iterations. It is standard to discard a number of initial iterations of the Markov chain in MCMC, as the chain moves from its initial value to the region containing the probability mass of the posterior, the chain's stationary distribution. The discarded steps are known as the burn-in.

The flat portions of the plot indicate iterations during which the algorithm rejected proposed values. An ideal trace plot shows no trend over time, an indication that it has reached its stationary distribution, and rapidly traversesup and down to the upper and lower values of the posterior, indicating that the chain is mixing well (see Figures 2a and 2b). In Figure 1, the chain appears to have reached its stationary distribution since it has no trend, but the distinct zigs and zags indicates that the chain may not have mixed quickly enough to have sufficiently converged to the posterior distribution within this number of iterations.

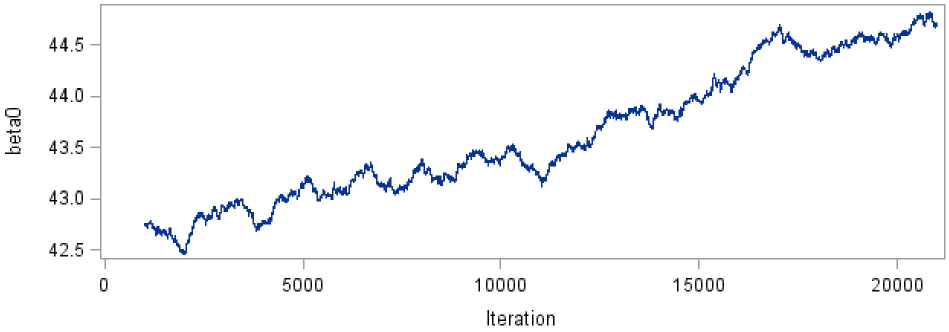


Figure 2a. Markov Chain That Has Not Converged to Its Stationary Distribution

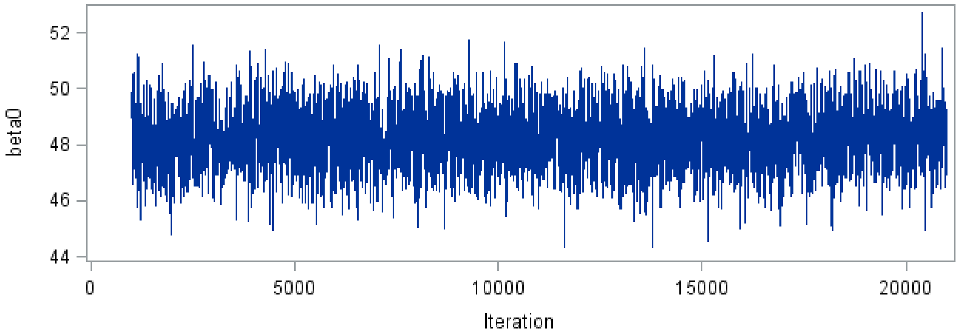


Figure 2b. Markov Chain That Has Converged to Its Stationary Distribution and Mixed Well

The autocorrelation plot displays the correlation between each step of the Markov chain, and the preceding steps. The samples from the posterior distribution produced by Markov chain Monte Carlo are not independent. This is not a theoretical problem, but higher correlation between iterations will cause the chain to mix more slowly and will require more iterations to ensure the chain has converged to the posterior.

Model Tuning

There are a variety of strategies that can help secure MCMC convergence. Table 1 lists a few of the most useful and the corresponding PROC MCMC syntax. Using the PROPCOV option to numerically approximate the parameter covariance matrix for the proposal distribution (e.g., with QUANEW) rather than using the identity matrix (the default) often improves model convergence. Parameter blocking can also have a substantial impact on MCMC convergence. SAS allocates all the parameters specified within a PARMS statement to the same block. One joint proposal distribution is used for all parameters in a block so that acceptance or rejection occurs for the entire block. With many parameters in the same block, the acceptance rate may be too low slowing convergence. With many blocks of one or a few parameters, the chain may mix too slowly. It is important to find a balance conducive to chain convergence in particular applications.

Strategy	PROC MCMC Syntax
More Iterations	NMC
Thinning	THIN
Numerically Approximate Covariance Matrix for Proposal Distribution	PROPCOV
More Informative Priors	PRIOR
Group or Ungroup Parameters in Blocks	PARMS
Multiple Chains	
Better Initial Values	
Put Variables on Same Scale	

Table 1. Strategies to Encourage Markov Chain Convergence

Prior Specification

In the case of highway MPG and vehicle weight, basic familiarity with automobiles and their gas mileage offers prior information that can be used to devise more informative priors than those used above. A little research or consultation with an expert would undoubtedly yield highly detailed information. Based on a general knowledge of vehicles, it is reasonable to suppose that MPG and weight would be negatively correlated, and so the prior distribution of β_1 should be negative with high probability. It is also known that β_1 (the change in MPG for a one pound increase in weight) will be very small, since a small change in weight (in pounds) should have a minimal impact on vehicle MPG. A rough guess that a 1,000 pound increase in weight might be associated with a reduction of five MPG, yields -0.005 as a reasonable candidate for prior mean. Centering a normal distribution at this value with a standard deviation of 0.0025 gives β_1 a prior distribution that is negative with probability approximately 0.98.

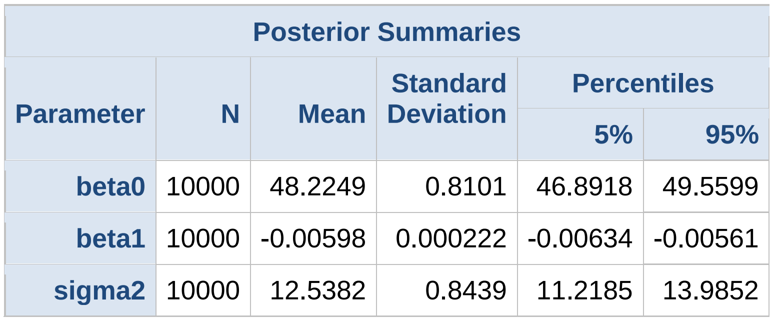
Specifying a prior for the intercept, β_0, may be slightly less intuitive, since there are no vehicles with zero weight. However, since MPG is nonnegative, and the slope β_1 is expected to be negative, an informative prior for β_0 should give it a positive value with very high probability. A normal distribution centered at 60 with a standard deviation of 20 has roughly 95% of its probability density between 20 and 100, which is still rather vague for this example, but is certainly more informative than the example above.

Prior specification for variance parameters is often less intuitive, beyond the certainty that they must be nonnegative. In this case, it is useful to note that while vehicle mileage certainly varies among vehicles, this variation is not extreme. An inverse gamma distribution with shape parameter 6 and scale parameter 100 has the bulk of its probability density below 100. This provides a more informative prior for σ^2 without imposing severe restrictions on the support of the posterior. The (somewhat) informative prior for our model parameters is:

$$\beta_0 \sim \mathrm{N}(60,\ 20^2),$$
$$\beta_1 \sim \mathrm{N}(-0.005,\ 0.0025^2),$$
$$\sigma^2 \sim \mathrm{InvGamma}(6, 100).$$

The following call to the MCMC procedure employs these informative priors and illustrates the syntax for specifying burn-in, thin and number of iterations in the Markov chain. In addition it requests the 5th and 95th posterior quantiles for each parameter.

```
/* 100,000 iterations with a burn-in of 5,000 keeping every 10th iteration */
/* display 95% posterior credible intervals */
proc mcmc data=sashelp.cars nmc=100000 nbi=5000 thin=10 stats(percentage=(5 95));
    /* declare parameters */
    parms beta0 beta1 sigma2;
    /* informative prior */
    prior beta0 ~ normal(60, sd=20);
    prior beta1 ~ normal(-0.005, sd=.0025);
    prior sigma2 ~ igamma(6, scale=100);
    /* likelihood */
    model mpg_highway ~ normal(beta0 + beta1*weight, var=sigma2);
run;
```



Posterior Summaries

Parameter	N	Mean	Standard Deviation	Percentiles	
				5%	95%
beta0	10000	48.2249	0.8101	46.8918	49.5599
beta1	10000	-0.00598	0.000222	-0.00634	-0.00561
sigma2	10000	12.5382	0.8439	11.2185	13.9852

The diagnostic plots for β_0 (Figure 3) and β_1 and σ^2 (not shown) seem to indicate that the chain has reached its stationary distribution and mixed well, and we may reasonably conclude that it has converged to the posterior distribution. The posterior summary statistics for the parameters have not changed substantially from the previous model, indicating that even the more informative prior is vague relative to the information in the data.

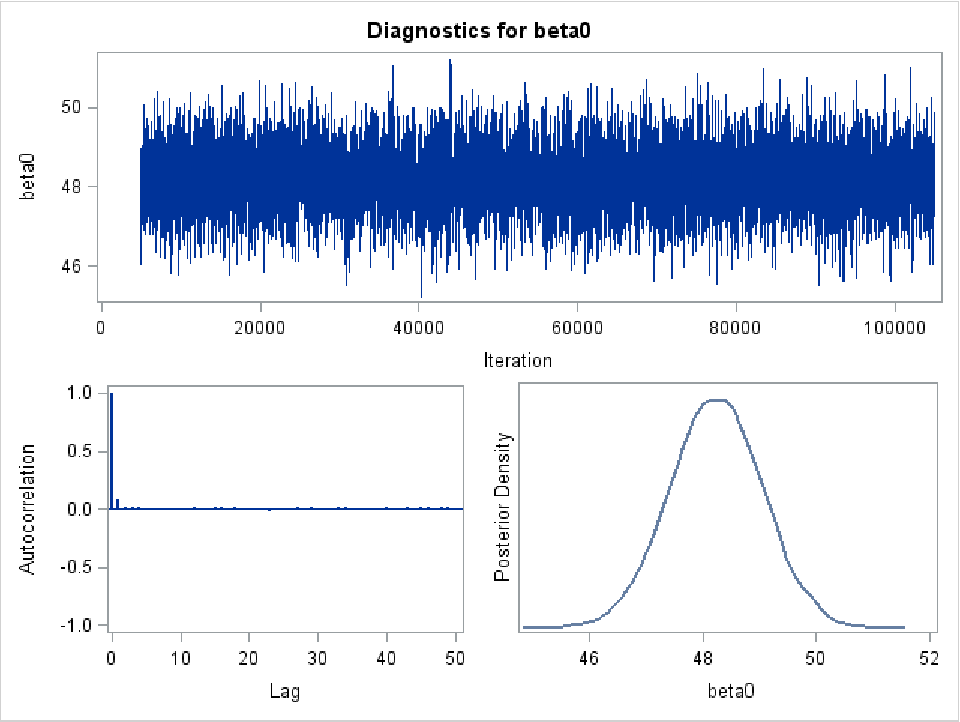


Figure 3. Diagnostic Plots for β_0

BAYESIAN ANALYSIS OF MEPS HEALTH CARE EXPENDITURES

The Medical Expenditure Panel Survey (MEPS) is a nationally-representative survey conducted by the Agency for Healthcare Research and Quality that follows individuals over a two-year panel, recording detailed information on health care utilization and expenditures. These expenditures are not limited to out-of-pocket payments, but include expenditures made on behalf of an individual by payers through whom the individual has health insurance coverage. In this analysis, the nationally-representative MEPS survey weights are ignored, focusing on the relationships between variables without respect for the survey design.

Annual health care expenditures are non-negative and are typically characterized by a large number of null observations, respondents for whom no expenditures were made in a particular year. The tremendous cost of treating certain conditions yields very large observations, resulting in a large variance and a highly skewed distribution for those with non-zero expenditures. During Panel 14 of MEPS, over 14,000 individuals were surveyed in 2009. The mean health care expenditures was $3,484 with 16.8% of individuals having no expenditures. As is typical of skewed data, the mean was considerably larger than the median ($691). The standard deviation of $9,596 was large compared to the mean, which is also typical of non-negative data with a long right tail.

THE MODEL

In this study, a zero-inflated gamma regression is employed, combining a logistic regression for the zero observations with a gamma distribution for non-zero values, which is strictly positive and allows for skew and over-dispersion,.

Let y_i denote the health care expenditures of the i^{th} individual, and let y_i conditional on parameters p_i, a_i and b_i follow a zero-inflated gamma distribution with density function,

$$f(y_i|p_i, a_i, b_i) = \begin{cases} p_i & , \quad y_i = 0 \\ (1-p_i)\dfrac{b_i^{a_i}}{\Gamma(a_i)} y_i^{a_i-1} e^{-b_i y_i}, & \quad y_i > 0. \end{cases}$$

The zero-inflated gamma distribution is a mixture of a Bernoulli distribution and a gamma distribution. The Bernoulli component has parameter p_i, which denotes the probability that y_i is zero (i.e., the probability that person i has no health care expenditures in 2009). This probability, as well as the mean and variance of the gamma distribution are modeled as functions of linear combinations of predictors x_i,

$$\text{logit}(p_i) = x_i'\gamma,$$

$$\log(\mu_i) = x_i'\beta,$$
$$\log(\sigma_i) = x_i'\alpha.$$

Allowing the variance of the gamma distribution to vary for different predictor values is particularly important for health care expenditures, which are highly variable. In this case the same set of predictors x_i is used for all three expressions, but this need not be the case. The shape parameter a_i and rate parameter b_i of the gamma distribution can be expressed in terms of μ_i and σ_i^2:

$$a_i = \frac{\mu_i^2}{\sigma^2}\, ,\; b_i = \frac{\mu_i}{\sigma_i^2}\, .$$

The regression coefficients γ, β and α are given reasonably uninformative normal priors.

PROC MCMC

While the normal distribution is a standard distribution available in the MCMC procedure, the zero-inflated gamma must be manually specified using the general function. The general function takes as its argument the log likelihood of the desired distribution. If y_i is distributed zero-inflated gamma then its contribution to the log likelihood is

$$\ell_i = \begin{cases} \log(p_i) & , \quad \mathrm{y_i} = 0 \\ \log(1-p_i) +\ a_i\log(b_i) -\ \log(\Gamma(a_i)) + (a_i - 1)\log(y_i) - b_i y_i, & \quad \mathrm{y_i} > 0. \end{cases}$$

The PROC MCMC statement below fits this model, keeping every fiftieth iteration from a single Markov chain of 200,000 iterations after an initial burn-in of 10,000 iterations.

```
  proc mcmc data=meps14 (where = (year = 1))
          outpost=mcmcout nmc=200000 nbi=10000 thin=50 propcov=quanew init=random;

      array alpha[37] alpha1-alpha37;
      array beta[37] beta1-beta37;
      array gamma[37] gamma1-gamma37;
      array x[37] female age0_5 age6_30 age61plus latino black amerind asian
             private_ins medicaid medicare uninsured collegeplus highschool
             underweight overweight obese morbidlyobese actlim smoke cancer
             diabetes asthma usualsource stroke heart otherheart employ povertylev
             midwest south west married widowed divorced goodhealth goodmental;

      /* one block of parameters */
      parms (alpha: beta: gamma: alpha0 beta0 gamma0) 1;

      /* uninformative prior distributions */
      prior alpha: ~ normal(1,sd=1);
      prior gamma: ~ normal(0,sd=1);
      prior beta: ~ normal(0,sd=1);

      /* compute linear predictors */
      xg = gamma0;
      xb = beta0;
      xa = alpha0;
      do i = 1 to 37;
             xg = xg + x[i]*gamma[i];
             xb = xb + x[i]*beta[i];
             xa = xa + x[i]*alpha[i];
      end;
      p0 = logistic(xg);
      mu = exp(xb);
      sd = exp(xa);
      a = (mu**2) / (sd**2);
      b = mu / (sd**2);

      /* compute contribution to the log likelihood */
      if y = 0 then ll = log(p0);
      else if y > 0 then ll = log(1-p0) + a*log(b) - log(gamma(a))
                                + (a-1)*log(y) - b*y;
      model y ~ general(ll);
  run;
```

Specifying PROPCOV=QUANEW employs a quasi-Newtonian numerical approximation for the parameter covariance matrix in the proposal distribution. In this case the chain does not converge even after very many iterations if the default PROPCOV=IND is used. Note also that since all 114 parameters are listed in the same PARMS statement, they are grouped in one block. In this case that proved to be efficient due to the relatively high correlations between parameters, however, in many cases it is useful to split model parameters into several blocks.

RESULTS

Parameter posterior distributions are summarized in Tables 2, 3 and 4 for the logistic and gamma regressions. The parameter estimates in Table 2 denote the log odds ratio of having no expenditures for the specified value relative to that variable's reference value, holding all other predictors constant. Many of the predictors have posterior intervals that do not overlap zero, indicating a significant increase or decrease in the odds of zero expenditures. Education, illness, having been married, being female, and income all have negative coefficients, indicating a reduction in the odds of zero expenditures.

Conversely, positive parameter estimates indicate values associated with an increased odds of zero expenditures compared to the reference value. Black and Asian race increase the odds of zero expenditures compared to the White reference category. Latino ethnicity, smoking, good health status, and being uninsured are also associated with higher odds of no expenditures relative to non-Latino, not smoking, fair or poor health status and being insured.

The regression coefficients for the mean of the gamma distribution summarized in Table 3 may be interpreted as the log of the multiplicative effect on the mean, i.e., the log of the ratio of means, associated with a one unit increase in the value of that variable holding all other predictors constant. For dichotomous variables this is simply the log of the ratio of the category means. As with the logistic regression coefficients, many predictors have posterior intervals that do not include zero. Increased age and income indicate increased mean expenditures, and private group health insurance has increased mean expenditures compared to all other health insurance categories. Illness, fair or poor health status, non-white race, Latino ethnicity, and never married are associated with decreased mean expenditures.

The coefficients summarized in Table 4 are the multiplicative effect on the standard deviation of the gamma distribution corresponding to a one unit increase in the value of that predictor. In general these coefficients are similar in sign and significance to those in Table 3. Those predictors associated with increased expenditures tend to also be associated with increased variance of those expenditures. Interestingly this does not hold true for increased income, which is associated with a significant increase in average expenditures but no corresponding increase in expenditure variance.

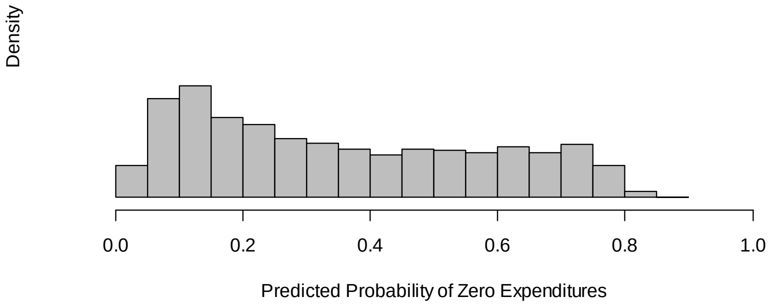


Figure 4. Predicted Probability of Zero Expenditures for Observations with Zero Expenditures

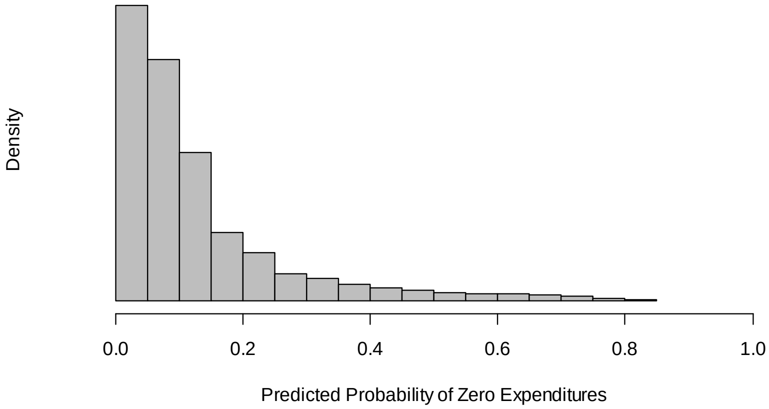


Figure 5. Predicted Probability of Zero Expenditures for Observations with Non-Zero Expenditures

Coefficient (reference)	Posterior Mean	Standard Deviation	2.5th Quantile	97.5th Quantile	Interval Does Not Overlap Zero
Intercept	-1.16	0.19	-1.52	-0.78	x
Female	-0.50	0.05	-0.60	-0.40	x
Age (31–60)					
0–5	-0.56	0.13	-0.81	-0.30	x
6–30	-0.04	0.08	-0.19	0.13	
Over 60	-0.22	0.14	-0.49	0.06	
Latino	0.57	0.07	0.43	0.69	x
Race (White)					
Black	0.45	0.08	0.29	0.59	x
American Indian	-0.34	0.24	-0.80	0.12	
Asian	0.45	0.11	0.25	0.66	x
Health Insurance (Private Group)					
Private Non-Group	< 0.01	0.19	-0.38	0.37	
Medicaid/Other Public	-0.08	0.08	-0.25	0.07	
Medicare	0.26	0.19	-0.14	0.63	
Uninsured	0.93	0.07	0.77	1.05	x
Education (Lower than High School)					
College or Higher Degree	-0.53	0.09	-0.70	-0.36	x
High School/GED	-0.11	0.07	-0.24	0.03	
Body Mass Index (Normal)					
Underweight	-0.42	0.13	-0.64	-0.15	x
Overweight	-0.08	0.06	-0.21	0.04	
Obese	-0.08	0.08	-0.25	0.08	
Morbidly Obese	-0.04	0.17	-0.40	0.26	
Activity Limitations	-0.71	0.17	-1.06	-0.37	x
Smoking	0.25	0.08	0.09	0.40	x
Cancer	-0.80	0.20	-1.21	-0.41	x
Diabetes	-1.16	0.20	-1.54	-0.75	x
Asthma	-0.98	0.14	-1.25	-0.70	x
Usual Source of Care	-1.33	0.06	-1.44	-1.22	x
Stroke	-1.15	0.41	-2.06	-0.38	x
Heart	-0.78	0.25	-1.31	-0.31	x
Other Heart	-0.88	0.19	-1.26	-0.53	x
Employed	0.27	0.07	0.13	0.40	x
Income	-0.77	0.13	-1.03	-0.51	x
Region (Northeast)					
Midwest	-0.18	0.10	-0.38	-0.01	x
South	-0.17	0.08	-0.33	-0.02	x
West	< 0.01	0.09	-0.19	0.16	
Marital Status (Never Married)					
Married	-0.27	0.07	-0.41	-0.13	x
Widowed	-0.56	0.23	-1.02	-0.13	x
Divorced	-0.31	0.11	-0.52	-0.10	x
Good/Very Good/Excellent Health	0.57	0.11	0.34	0.77	x
Good/Very Good/Excellent Mental Health	0.35	0.13	0.10	0.60	x

Table 2. Posterior Summary for Zero-Inflation Regression Coefficients

Coefficient (reference)	Posterior Mean	Standard Deviation	2.5th Quantile	97.5th Quantile	Interval Does Not Overlap Zero
Intercept	8.41	0.09	8.25	8.59	x
Female	0.08	0.02	0.03	0.12	x
Age (31–60)					
0–5	-0.68	0.06	-0.80	-0.57	x
6–30	-0.23	0.04	-0.30	-0.15	x
Over 60	0.18	0.04	0.10	0.25	x
Latino	-0.22	0.03	-0.28	-0.16	x
Race (White)					
Black	-0.20	0.03	-0.26	-0.14	x
American Indian	-0.10	0.11	-0.30	0.11	
Asian	-0.30	0.05	-0.39	-0.21	x
Health Insurance (Private Group)					
Private Non-Group	-0.32	0.06	-0.43	-0.19	x
Medicaid/Other Public	-0.11	0.03	-0.18	-0.05	x
Medicare	-0.20	0.05	-0.29	-0.10	x
Uninsured	-0.41	0.04	-0.49	-0.34	x
Education (Lower than High School)					
College or Higher Degree	0.31	0.04	0.23	0.38	x
High School/GED	0.25	0.03	0.19	0.32	x
Body Mass Index (Normal)					
Underweight	0.05	0.05	-0.04	0.16	
Overweight	0.08	0.03	0.03	0.14	x
Obese	0.07	0.03	0.01	0.14	x
Morbidly Obese	0.16	0.07	0.03	0.29	x
Activity Limitations	0.57	0.04	0.49	0.65	x
Smoking	-0.01	0.04	-0.08	0.06	
Cancer	0.46	0.04	0.37	0.55	x
Diabetes	0.41	0.04	0.33	0.49	x
Asthma	0.34	0.04	0.26	0.41	x
Usual Source of Care	0.25	0.04	0.17	0.32	x
Stroke	0.26	0.07	0.13	0.39	x
Heart	0.27	0.05	0.17	0.36	x
Other Heart	0.39	0.04	0.31	0.47	x
Employed	-0.20	0.03	-0.26	-0.15	x
Income	0.16	0.04	0.08	0.25	x
Region (Northeast)					
Midwest	0.06	0.04	-0.02	0.13	
South	-0.05	0.03	-0.11	0.01	
West	0.05	0.03	-0.02	0.11	
Marital Status (Never Married)					
Married	0.09	0.04	0.02	0.16	x
Widowed	0.15	0.06	0.04	0.26	x
Divorced	0.13	0.05	0.03	0.22	x
Good/Very Good/Excellent Health	-0.60	0.04	-0.67	-0.53	x
Good/Very Good/Excellent Mental Health	-0.25	0.05	-0.35	-0.15	x

Table 3. Posterior Summary for Gamma Mean Regression Coefficients

Coefficient (reference)	Posterior Mean	Standard Deviation	2.5th Quantile	97.5th Quantile	Interval Does Not Overlap Zero
Intercept	8.81	0.09	8.62	8.99	x
Female	0.01	0.03	-0.05	0.06	
Age (31–60)					
0–5	-0.77	0.06	-0.89	-0.67	x
6–30	-0.26	0.04	-0.34	-0.18	x
Over 60	0.07	0.04	-0.01	0.16	
Latino	-0.19	0.03	-0.26	-0.13	x
Race (White)					
Black	-0.15	0.04	-0.22	-0.08	x
American Indian	-0.09	0.13	-0.32	0.17	
Asian	-0.27	0.05	-0.38	-0.16	x
Health Insurance (Private Group)					
Private NonGroup	-0.32	0.07	-0.45	-0.18	x
Medicaid/Other Public	-0.07	0.04	-0.15	< 0.01	
Medicare	-0.19	0.05	-0.29	-0.09	x
Uninsured	-0.33	0.04	-0.41	-0.25	x
Education (Lower than High School)					
College or Higher Degree	0.28	0.04	0.20	0.37	x
High School/GED	0.27	0.03	0.21	0.34	x
Body Mass Index (Normal)					
Underweight	0.09	0.06	-0.03	0.20	
Overweight	0.09	0.03	0.02	0.15	x
Obese	0.05	0.04	-0.02	0.13	
Morbidly Obese	0.17	0.07	0.04	0.32	x
Activity Limitation(s)	0.56	0.04	0.47	0.65	x
Smoking	-0.01	0.04	-0.08	0.08	
Cancer	0.50	0.05	0.39	0.59	x
Diabetes	0.34	0.05	0.24	0.43	x
Asthma	0.29	0.04	0.20	0.37	x
Usual Source of Care	0.16	0.04	0.08	0.23	x
Stroke	0.24	0.07	0.11	0.40	x
Heart	0.24	0.06	0.13	0.34	x
Other Heart	0.40	0.05	0.31	0.49	x
Employed	-0.19	0.03	-0.25	-0.13	x
Income	0.03	0.05	-0.06	0.14	
Region (Northeast)					
Midwest	0.09	0.04	< 0.01	0.17	x
South	-0.06	0.04	-0.14	0.01	
West	0.09	0.04	0.01	0.16	x
Marital Status (Never Married)					
Married	0.09	0.04	0.01	0.16	x
Widowed	0.11	0.06	< 0.01	0.23	x
Divorced	0.08	0.05	-0.03	0.18	
Good/Very Good/Excellent Health	-0.62	0.04	-0.70	-0.54	x
Good/Very Good/Excellent Mental Health	-0.25	0.06	-0.35	-0.14	x

Table 4. Posterior Summary for Gamma Variance Regression Coefficients

Figures 4 and 5 depict posterior predicted probabilities of zero expenditures for observations with zero and non-zero observed expenditures respectively. The mean predicted probability was 0.37 for those with zero observed expenditures, and 0.13 for those with non-zero expenditures observed.

Figure 6 plots the predictive posterior median and interval (given non-zero expenditures) against observed 2010 expenditures for all observations with non-zero 2010 expenditures. The observed expenditures exceeded the 97.5th quantile of the predictive posterior 3.9% of the time, indicating that the posterior predictions may slightly underestimate the highest expenditures.

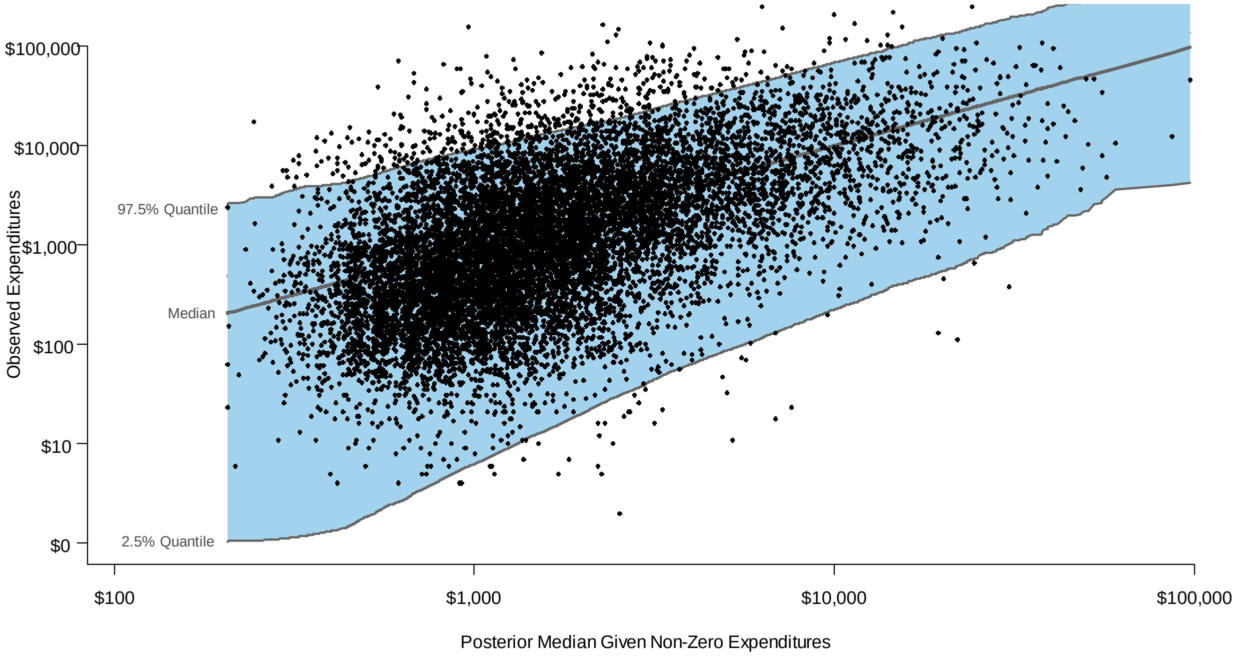


Figure 6. 2010 Observed Expenditures and Posterior Predictive Distribution (Log Scale)

CONCLUSION

Bayesian statistics allows prior information to be incorporated into data analysis and inference, and offers intuitive interpretations of many quantities. PROC MCMC enables the SAS user to fit a wide variety of Bayesian models using Markov chain Monte Carlo. A simple example illustrates the basic syntax of the procedure and the equivalence of Bayesian and maximum likelihood inference when uninformative priors are used. Diagnostic plots are discussed, and suggestions for improving Markov chain are offered. Finally, a more complicated Bayesian model is illustrated in PROC MCMC, fitting a zero-inflated gamma regression to health care expenditures from the Medical Expenditure Panel Survey.

ACKNOWLEDGMENTS

This paper would not have been possible with the encouragement and support of Xiao Chen.

CONTACT INFORMATION

Greg Watson
UCLA Center for Health Policy Research
10960 Wilshire Blvd, Suite 1550
Los Angeles, CA 90024
gwatson@ucla.edu