A Bayesian Approach to Regression Models

1 References

Once again it seems to make sense to refer you to the sources for this material before presenting the material. Sections 3, 4, and 5 below follow in spirit the presentation in portions of Chapter 8 of

- Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. B. (1995). *Bayesian data analysis.* New York: Chapman-Hall.

A much briefer account, that emphasizes the "shrinkage" aspect of Bayesian estimation, may be found on pp. 38–50 of

- Carlin, B. P. and Louis, T. A. (1996). *Bayes and empirical Bayes methods for data analysis*. New York: Chapman-Hall.

The Carlin-Louis material follows closely that of the classic (but not classical!) reference

- Lindley, D. V. and Smith, A. F. M. (1972). Bayes estimates for the linear model (with discussion). *Journal of the Royal Statistical Society, Series B, 34,* 1–41.

Finally, much more algebraic detail may be found in Chapters 8 and 9 of

- Box, G. E. P. and Tiao, G. C. (1973). *Bayesian inference in statistical analysis*. Reading MA: Addison-Wesley.

In this handout I will give a thumbnail sketch of the "essential" difference between classical and Bayesian inference, and show you the ingredients for the ordinary linear regression model. Then I will describe Bayesian inference with non-informative priors for the ordinary linear regression model, and show how this can be immediately extended to analyses with informative conjugate priors. I will close with an example.

2 Introduction

It is sometimes useful to lay out the computational ingredients of Bayesian and classical inference side-by-side to see what differences to expect in practical data analysis problems. Here is my thumbnail sketch of the difference.

Let us suppose that we are interested in the usual linear regression model

$$\underline{y} = X\underline{\beta} + \underline{\epsilon}, \quad \underline{\epsilon} \sim N(\underline{0}, \sigma^2 I)$$

where X is fixed an known in advance. To make the notation more compact I will collect together all of the parameters of this model into one parameter vector $\underline{\theta}$:

$$\underline{\theta}^T \;=\; (\underline{\beta}^T, \sigma^2)^T \;=\; (\beta_0, \beta_1, \ldots, \beta_k, \sigma^2)^T$$

So far I have specified enough for you to write down the (or "a", if you are fussy about it) *likelihood* of the data

$$\begin{aligned} p(\underline{y}|X,\underline{\theta}) &= \prod_1^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2}(y_i - \underline{x}_i^T\underline{\beta})^2/\sigma^2\right] \\ &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left[-\frac{1}{2}\sum_1^n (y_i - \underline{x}_i^T\underline{\beta})^2/\sigma^2\right] \\ &= \frac{1}{(2\pi)^{n/2}|\sigma^2 I|^{1/2}} \exp\left[-\frac{1}{2}(\underline{y} - X\underline{\beta})^T(\sigma^2 I)^{-1}(\underline{y} - X\underline{\beta})\right], \end{aligned}$$

where as usual $\underline{x}_i^T$ is the i^{th} *row* of the matrix X.

Classical statistical analysis is driven by the interpretation of $p(\underline{y}|X,\underline{\theta})$ as a probability density for the *random quantities* $\underline{y}$, given fixed and known X *and* fixed and *unknown* $\underline{\theta}$. Inferences for $\underline{\theta}$ are driven by "what if" style questions and answers:

- *Point estimation:* What value of $\underline{\theta}$ maximizes the probability of observing the data $\underline{y}$ that we saw, rather than some other data? (MLE's)
- *Interval estimation:* Can we develop recipies $A(\underline{y}, X)$, $B(\underline{y}, X)$ such that $P[\, A(\underline{y}, X) \le \theta_j \le B(\underline{y}, X) \mid \underline{\theta}\,] > 0.95$, in repeated samples of $\underline{y}$? Then apply the recipies to the $\underline{y}$ actually observed. (Confidence intervals; the probability statement is about the recipies, not the particular numbers calculated)
- *Hypothesis testing:*
  - Can we develop test functions $T(\underline{y}, X)$ such that the distribution $q_0[T(\underline{y}, X)|\underline{\theta} = \underline{\theta}_0]$ is well-understood? If we calculate $T(\underline{y}, X)$ for the data $\underline{y}$ actually observed and it is an outlier for the q_0 distribution this is evidence against $\underline{\theta} = \underline{\theta}_0$ (general hypothesis testing)
  - Is the likelihood of the data under the "small model" much smaller than the likelihood of the data under a "big model" in which it is nested? (Likelihood ratio testing)

Note that all of the reasoning about distributions is done prospectively, as if predicting $\underline{y}$ on the basis of hypothetical values of $\underline{\theta}$ before $\underline{y}$ is actually observed. Then $\underline{y}$ is observed and compared with the predictions made to "learn" which hypothetical values of $\underline{\theta}$ make sense for the observed $\underline{y}$.

Bayesian statistical analysis supplies an additional set of modeling assumptions, that $\underline{\theta} \sim p(\underline{\theta}|X)$ is also random, with marginal, or "prior", distribution[1] $p(\underline{\theta}|X)$; and it is driven by the interpretation of

$$p(\underline{\theta}|\underline{y}, X) = \frac{p(\underline{y}|X,\underline{\theta})p(\underline{\theta}|X)}{p(\underline{y}|X)} \propto p(\underline{y}|X,\underline{\theta})p(\underline{\theta}|X) \tag{1}$$

as a probability density for the random quantities $\underline{\theta}$, *after* conditioning on observing $\underline{y}$, the data actually observed. Inferences for $\underline{\theta}$ are driven by direct calculation on the "posterior" distribution $p(\underline{\theta}|\underline{y}, X)$:

[1] Since the particular values of the entries in X are not usually thought of as constraining the "true value" of $\underline{\theta}$ in a regression problem, it usually makes sense to make $\underline{\theta}$ independent of X: $p(\underline{\theta}|X) = p(\underline{\theta})$.

- *Point estimation:* What value of $\underline{\theta}$ does $p(\underline{\theta}|\underline{y}, X)$ assign the highest probability to? (posterior mode) What is the average $\underline{\theta}$, given $\underline{y}$ and X? (posterior mean)
- *Interval estimation:* Can we develop recipies $A(\underline{y}, X)$, $B(\underline{y}, X)$ such that in the posterior distribution $P[A(\underline{y}, X) \leq \theta_j \leq B(\underline{y}, X) \mid \underline{y}, X] > 0.95$? (credible intervals's; the probability statement is about the actual numbers calculated)
- *Hypothesis testing:*
  - Can we calculate the distribution of additional data $\underline{y}^*$ that follows the same model we are fitting? If so, does the test function $T(\underline{y}, X)$ evaluated on the data actually observed look like an outlier in this "predictive" distribution? (Posterior predictive checks)
  - Is the probability of the data $P(\underline{y}|X, M_1)$ under model 1 greater than that under model 2, $P(\underline{y}|X, M_2)$? (Bayes factors)

As indicated above, a common method of point estimation for Bayesian inference is to guess that $\underline{\theta}$ is in the region of the parameter space where the posterior $p(\underline{\theta}|\underline{y}, X) \propto p(\underline{y}|X, \underline{\theta})p(\underline{\theta}|X)$ is highest. The likelihood imposes one set of constraints on where $\underline{\theta}$ should be (near the MLE, where the likelihood $p(\underline{y}|X, \underline{\theta})$ is highest) and the prior imposes another set of constraints (near the prior mode, where the prior distribution $p(\underline{\theta}|X)$ is highest). The Bayesian point estimate is a compromise[2] between these two sets of constraints.

For this reason, especially strong features of the likelihood $p(\underline{y}|X, \underline{\theta})$ or of the prior $p(\underline{\theta}|X)$ will dominate the Bayesian analysis. For example a common situation for iid or exchangeable data is "likelihood dominated" or "data dominated" inference, in which there are so many observations that the likelihood $p(\underline{y}|X, \underline{\theta})$ has a single very sharp, very narrow peak, and the prior looks flat or featureless by contrast. In this case, the prior will exert very little influence in a Bayesian analysis (since the shape of the posterior will be essentially the same as the shape of the likelihood), and the Bayesian results will be very similar to the classical results. One way to ensure likelihood dominance is to choose "flat" or "noninformative" priors. Then we expect the Bayesian answers to be numerically and computationally parallel the classical ones (though the probability interpretations of the answers will be different, in terms of $p(\underline{\theta}|\underline{y}, X)$ rather than $p(\underline{y}|X, \underline{\theta})$). As we move away from flat priors, the special features of the prior will pull the inferences away from the classical results.

3 Analysis of the Ordinary Linear Regression Model with Non-informative Priors

A central difficulty of a purist approach to Bayesian statistics is the careful specification of prior distributions for the parameters. As indicated above, in likelihood-dominated problems, careful specification of the prior is not very important unless there is tension between the prior information and the data. An alternative is to use a *noninformative prior distribution*[3] which is usually so featureless that every problem turns into a likelihood-dominated problem.

[2]This compromise is also called "shrinkage": the Bayesian answer "shrinks" the classical answer toward the prior mean or mode. In typical location models, when the prior mean is 0, the effect is to reduce, or shrink, the unsigned magnitude of the classical answer.

[3]Noninformative priors are usually flat (i.e. constant) or nearly flat, and if the parameter space is unbounded this leads to "improper" prior distributions, that is, a prior distribution that assigns infinite probability (instead of probability 1) to the parameter space. Often the posterior distribution is still "proper" so that formally we can work with the posterior as indicating probable locations for the paramter, as in any other Bayesian problem. However, a careful interpretation of the posterior as a conditional density derived from the joint density $p(\underline{y}, \underline{\theta}|X)$, when the prior is improper, is not possible.

In regression analysis, starting with a non-informative prior has two additional advantages: (1) it shows clearly the parallels between classical and Bayesian analysis; (2) it is a good foundation on which to build Bayesian analyses when we have prior information that can be specified through a conjugate prior distribution.

The standard, or default, non-informative prior for regression analysis is

$$p(\underline{\beta}, \log \sigma) \propto \text{constant}$$

or, after a calculus change-of-variables

$$p(\underline{\beta}, \sigma^2) \propto \sigma^{-2}$$

Note that this is flat (i.e. constant, or uniform) in the regression coefficients $\underline{\beta}$, so no constraints beyond those imposed by the likelihood are placed on them. It is not flat in σ^2, however: values of σ^2 near zero are preferred over values near infinity. It is also easy to verify that this prior is improper in each coordinate of the parameter vector $\underline{\theta} = (\underline{\beta}^T, \sigma^2)^T$.

3.1 Setting up the joint posterior distribution

Following Equation (1), and plugging in the likelihood and default prior for linear regression, we then have

$$\begin{aligned} \log p(\underline{\theta}|\underline{y}, X) &= (\text{Constant}) + \log p(\underline{y}|X, \underline{\theta}) + \log p(\underline{\theta}|X) \qquad (2) \\ &= (\text{Constant}) + \left[-n \log \sigma - \frac{1}{2}(\underline{y} - X\underline{\beta})^T(\underline{y} - X\underline{\beta})/\sigma^2 \right] + \left[-2 \log \sigma \right] \\ &= (\text{Constant}) - (n+2) \log \sigma - \frac{1}{2}(\underline{y} - X\underline{\beta})^T(\underline{y} - X\underline{\beta})/\sigma^2, \end{aligned}$$

where the "constant" terms in each line are not necessarily the same, but none depend on $\underline{\theta}$. One can readily verify[4] (multiply it out, and use the definition of $\hat{\underline{\beta}}$!) that

$$(\underline{y} - X\underline{\beta})^T(\underline{y} - X\underline{\beta})/\sigma^2 = (\underline{y} - X\hat{\underline{\beta}})^T(\underline{y} - X\hat{\underline{\beta}})/\sigma^2 + (\underline{\beta} - \hat{\underline{\beta}})^T X^T X(\underline{\beta} - \hat{\underline{\beta}})/\sigma^2$$

where

$$\hat{\underline{\beta}} = (X^T X)^{-1} X^T \underline{y}$$

is the usual OLS estimator of $\underline{\beta}$, and hence

$$\begin{aligned} \log p(\underline{\theta}|\underline{y}, X) &= (\text{Constant}) - (n+2) \log \sigma - \frac{1}{2}(\underline{y} - X\hat{\underline{\beta}})^T(\underline{y} - X\hat{\underline{\beta}})/\sigma^2 \\ &\quad - \frac{1}{2}(\underline{\beta} - \hat{\underline{\beta}})^T X^T X(\underline{\beta} - \hat{\underline{\beta}})/\sigma^2 \qquad (3) \end{aligned}$$

3.2 Identifying the pieces

The part of Equation (3) that depends on $\underline{\beta}$ gives the shape[5] of $\log p(\underline{\beta}|\sigma^2, \underline{y}, X)$; since this is a quadratic form in $\underline{\beta}$ it follows that $p(\underline{\theta}|\sigma^2, \underline{y}, X)$ is multivariate normal:

$$\underline{\beta} \mid \sigma^2, \underline{y}, X \;\sim\; N(\hat{\underline{\beta}}, \sigma^2(X^T X)^{-1}) \qquad (Bayes)$$

[4]cf. Box and Tiao, p. 439.

[5]This is an *extremely* useful trick in general for determining the so-called "complete conditinals" $p(\theta_1|\theta_2, \ldots, \theta_p, data)$ in setting up Markov Chain Monte Carlo (MCMC) calculations.

Note how this corresponds to the classical result

$$\underline{\hat{\beta}} \mid \underline{\beta}, \sigma^2, X \;\sim\; N(\underline{\beta}, \sigma^2(X^TX)^{-1}) \qquad (Classical)$$

Thinking now of

$$p(\underline{\beta}, \sigma^2|\underline{y}, X) \;=\; p(\underline{\beta}|\sigma^2, \underline{y}, X)p(\sigma^2|\underline{y}, X),$$

the marginal density for $p(\sigma^2|\underline{y}, X)$ could be obtained by integrating $\underline{\beta}$ out of the joint density $p(\underline{\beta}|\sigma^2, \underline{y}, X)p(\sigma^2|\underline{y}, X)$, but a cleverer thing to do is to realize that this also says that

$$p(\sigma^2|\underline{y}, X) = \frac{p(\underline{\beta}, \sigma^2|\underline{y}, X)}{p(\underline{\beta}|\sigma^2, \underline{y}, X)}$$

for *any* value of $\underline{\beta}$. After some algebra (plugging in $\underline{\beta} \equiv \underline{\hat{\beta}}$ simplifies things I think) one obtains that

$$\sigma^2 \mid \underline{y}, X \;\sim\; \text{Inv-}\chi^2\,(n-k-1, s^2)$$

a scaled, inverse chi-squared[6] random variable, where s^2 is the usual squared residual standard error,

$$s^2 = \frac{1}{n-k-1}(\underline{y} - X\underline{\hat{\beta}})^T(\underline{y} - X\underline{\hat{\beta}}) \equiv \frac{RSS}{n-k-1}.$$

This is equivalent to saying that

$$\left.\frac{(n-k-1)s^2}{\sigma^2}\,\right|\, \underline{y}, X \;\sim\; \chi^2_{n-k-1}, \qquad (Bayes)$$

which corresponds to the classical result

$$\left.\frac{(n-k-1)s^2}{\sigma^2}\,\right|\, \underline{\beta}, \sigma^2, X \;\sim\; \chi^2_{n-k-1} \qquad (Classical)$$

To summarize, if we use the default noninformative prior $p(\underline{\beta}, \sigma^2) \propto \sigma^{-2}$, we obtain a posterior for $\underline{\beta}$ and σ^2 that can be broken down as follows:

- $p(\underline{\beta}|\sigma^2, \underline{y}, X)$ is $N(\underline{\hat{\beta}}, \sigma^2(X^TX)^{-1})$;
- $p(\sigma^2|\underline{y}, X)$ is Inv-$\chi^2\,(n-k-1, s^2)$.

As with the classical results, the Bayesian results only make sense (and the posterior will only be proper) if X is of full rank $k+1$ and $n > k+1$.

3.3 Further inferences

It is also possible to show that the usual t_{n-k-1} distributions for each $\hat{\beta}_j$ in classical OLS regression carry over to the same t_{n-k-1} posterior distributions for each $\beta_j|\underline{y}, X$ in this Bayesian framework, etc., but it is often simpler to obtain approximations to this distribution by Monte Carlo simulation from the joint posterior for $(\underline{\beta}, \sigma^2)$ and then "throw away" the σ^2's simulated:

[6] Definition: If $V \sim \chi^2_d$ then $ds^2/V \sim$ Inv-$\chi^2\,(d, s^2)$. See Appendix A of Gelman et al. for formulas.

1. Simulate σ^2 from $p(\sigma^2|\underline{y}, X)$, i.e. from Inv-$\chi^2\,(n-k-1, s^2)$;
2. Then simulate $\underline{\beta}$ from $p(\underline{\beta}|\sigma^2, \underline{y}, X)$, i.e. from $N(\underline{\hat{\beta}}, \sigma^2(X^TX)^{-1})$.

This simulation is based on Monte Carlo estimation of the integral

$$\int\!\!\int T(\underline{\beta}, \sigma^2)\, p(\underline{\beta}, \sigma^2|\underline{y}, X)\, d(\sigma^2)\, d(\underline{\beta}) \;=\; \int\!\!\int T(\underline{\beta}, \sigma^2)\, p(\underline{\beta}|\sigma^2, \underline{y}, X)\, p(\sigma^2|\underline{y}, X)\, d(\sigma^2)\, d(\underline{\beta}), \qquad (4)$$

where $T(\underline{\beta}, \sigma^2)$ is any test function of interest.

Finally we turn to the problem of predicting new y values. Suppose

$$\tilde{X} = \begin{bmatrix} 1 & \tilde{x}_{11} & \cdots & \tilde{x}_{1k} \\ 1 & \tilde{x}_{11} & \cdots & \tilde{x}_{1k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \tilde{x}_{m1} & \cdots & \tilde{x}_{mk} \end{bmatrix}$$

are new values of the regressors, and

$$\underline{\tilde{y}} = \begin{bmatrix} \tilde{y}_1 & \tilde{y}_2 & \cdots & \tilde{y}_m \end{bmatrix}^T$$

are the corresponding new y values. The linear regression model says as usual that $p(\underline{\tilde{y}}|\underline{\beta}, \sigma^2, \tilde{X})$ will be $N(\tilde{X}\underline{\beta}, \sigma^2 I_{m\times m})$. We can then use Equation (4) with $T(\underline{\beta}, \sigma^2) = p(\underline{\tilde{y}}|\underline{\beta}, \sigma^2, \tilde{X})$,

$$p(\underline{\tilde{y}}|\underline{y}, X, \tilde{X}) = \int\!\!\int p(\underline{\tilde{y}}|\underline{\beta}, \sigma^2, \tilde{X}, \underline{y}, X)p(\underline{\beta}|\sigma^2, \underline{y}, X)p(\sigma^2|\underline{y}, X)\, d(\sigma^2)d(\underline{\beta})$$

in two ways:

- To derive the exact "predictive" distribution of $\underline{\tilde{y}}$. It turns out for example that (omitting X and $\tilde{X}$ from the conditioning to save space)

$$\begin{aligned} E[\underline{\tilde{y}} \mid \sigma^2, \underline{y}] &= E[E(\underline{\tilde{y}} \mid \underline{\beta}, \sigma^2, \underline{y}) \mid \sigma^2, \underline{y}] \\ &= E[X\underline{\beta} \mid \sigma^2, \underline{y}] \\ &= \tilde{X}\underline{\hat{\beta}} \end{aligned}$$

  and similarly

$$\begin{aligned} \text{Var}\,(\underline{\tilde{y}}|\sigma^2, \underline{y}) &= E[\text{Var}\,(\underline{\tilde{y}} \mid \underline{\beta}, \sigma^2, \underline{y}) \mid \sigma^2, \underline{y}] + \text{Var}\,(E[\underline{\tilde{y}} \mid \underline{\beta}, \sigma^2, \underline{y}] \mid \sigma^2, \underline{y}) \\ &= E[\sigma^2 I_{m\times m} \mid \sigma^2, \underline{y}] + \text{Var}\,(\tilde{X}\underline{\beta} \mid \sigma^2, \underline{y}) \\ &= \sigma^2(I_{m\times m} + \tilde{X}(X^TX)^{-1}\tilde{X}^T) \end{aligned}$$

  Note again how these results correspond with the classical ones: if σ^2 is known, $\underline{\tilde{y}}$ will follow a multivariate normal distribution with this mean and variance; if σ^2 is not known, then the corresponding t distribution results hold.

- To obtain simulated values of $\underline{\tilde{y}}$ by simulating from the joint posterior density of $(\underline{\tilde{y}}, \underline{\beta}, \sigma^2)$ and throwing away the simulated $\underline{\beta}$ and σ^2 values:

  1. Simulate σ^2 from $p(\sigma^2|\underline{y}, X)$, i.e. from Inv-$\chi^2\,(n-k-1, s^2)$;
  2. Then simulate $\underline{\beta}$ from $p(\underline{\beta}|\sigma^2, \underline{y}, X)$, i.e. from $N(\underline{\hat{\beta}}, \sigma^2(X^TX)^{-1})$;
  3. Finally simulate $\underline{\tilde{y}}$ from $p(\underline{\tilde{y}}|\underline{\beta}, \sigma^2, \tilde{X})$, i.e. from $N(\tilde{X}\underline{\beta}, \sigma^2 I_{m\times m})$.

3.4 Example: Radon measurements

Gelman et al. (1995, pp. 189 and 263) consider the data in Table 1 to learn about the effects of geographic location in Minnesota and floor of house on short term radon measurements in residential dwellings. An OLS regression analysis produces the following results

```
> radon _ read.table("radon.txt",header=T)
> radon _ cbind(LogRadon = log(radon$Radon),radon)
> names(radon)
[1] "LogRadon" "Radon"    "County"   "Floor"
> radon$Floor _ ifelse(radon$Floor=="B",0,1)
> radon _ cbind(radon,BE=ifelse(radon$County=="BlueEarth",1,0))
> radon _ cbind(radon,Clay=ifelse(radon$County=="Clay",1,0))
> radon _ cbind(radon,Goodhue=ifelse(radon$County=="Goodhue",1,0))
> names(radon)
[1] "LogRadon"  "Radon"  "County"  "Floor"  "BE"  "Clay"  "Goodhue"
> summary(olsfit _ lm(LogRadon ~ Floor + Clay + Goodhue,data=radon))

Call: lm(formula = LogRadon ~ Floor + Clay + Goodhue, data = radon)
Residuals:
    Min      1Q   Median     3Q   Max
 -1.653 -0.5572 -0.01608 0.5933 1.855

Coefficients:
              Value Std. Error t value Pr(>|t|)
(Intercept)  1.9561  0.2158     9.0663  0.0000
      Floor -0.3283  0.3155    -1.0409  0.3047
       Clay -0.0798  0.3018    -0.2644  0.7929
    Goodhue -0.0379  0.3041    -0.1248  0.9014

Residual standard error: 0.7895 on 37 degrees of freedom
Multiple R-Squared: 0.0332
F-statistic: 0.4235 on 3 and 37 degrees of freedom, the p-value is 0.7373

Correlation of Coefficients:
        (Intercept)   Floor    Clay
  Floor -0.2089
   Clay -0.6526     -0.1493
Goodhue -0.6762     -0.0114  0.4868
```

The predictors do not look like they matter much, but we will forge ahead. We first obtain the basic ingredients needed to specify the posterior distributions of σ^2 and $\underline{\beta}$

```
> X _ model.matrix(olsfit)
> print(r.df _ 41 - 3 - 1)
[1] 37
> R _ qr.R(qr(X))      # look at help(qr), help(qr.R), etc.
```


Radon	County	Floor	Radon	County	Floor
5.0	BlueEarth	B	8.8	Clay	B
13.0	BlueEarth	B	19.5	Clay	B
7.2	BlueEarth	B	2.5	Clay	F
6.8	BlueEarth	B	9.0	Clay	B
12.8	BlueEarth	B	13.1	Clay	B
5.8	BlueEarth	F	3.6	Clay	B
9.5	BlueEarth	B	6.9	Clay	F
6.0	BlueEarth	B	14.3	Goodhue	B
3.8	BlueEarth	B	6.9	Goodhue	F
14.3	BlueEarth	F	7.6	Goodhue	B
1.8	BlueEarth	B	9.8	Goodhue	F
6.9	BlueEarth	B	2.6	Goodhue	B
4.7	BlueEarth	B	43.5	Goodhue	B
9.5	BlueEarth	B	4.9	Goodhue	B
0.9	Clay	F	3.5	Goodhue	B
12.9	Clay	B	4.8	Goodhue	B
2.6	Clay	B	5.6	Goodhue	B
3.5	Clay	F	3.5	Goodhue	B
26.6	Clay	B	3.9	Goodhue	B
1.5	Clay	B	6.7	Goodhue	B
13.0	Clay	B			

Table 1: Short-term measurements of radon concentration (in picocuries/liter) in a sample of houses in three Minnesota counties. Floor=B for basement measurements, and Floor=F for first-floor measurements.

```
> Q _ qr.Q(qr(X))
> RI _ solve(R)        # easy since R is upper-triangular
> XTXI _ RI %*% t(RI)  # cheaper than XTXI _ solve(t(X)%*%X) !!!
> print(bhat _ solve(R,t(Q) %*% radon$LogRadon))
                       # solves R %*% bhat = t(Q) %*% y for bhat ...
                       # cheaper than XTXI %*% t(X) %*% radon$LogRadon
                   [,1]
(Intercept)  1.95613199
      Floor -0.32833983
       Clay -0.07980519
    Goodhue -0.03794035
> print(sqrt(s2 _ sum(residuals(olsfit)^2)/r.df))
[1] 0.7894929
```

Now we could give analytic results here but it is more informative to simulate:

```
> V _ rchisq(1000,r.df)
> sig2 _ r.df*s2/V        #  1000 draws from scaled inv-chi-squared
> beta _ matrix(rnorm(4000,0,sqrt(sig2)),nrow=4,ncol=1000,byrow=T)
> beta _ RI %*% beta + matrix(bhat,nrow=4,ncol=1000)
                          #  1000 corresponding draws from post for beta
```

Plotting the results takes more work than the simulations:

```
> smoothplot _ function(x,xlab="",ylab="") {
+ plot(density(x,width=sqrt(var(x))),type="l",xlab=xlab,ylab=ylab)
+ }
> par(mfrow=c(2,3))
> smoothplot(sig2,"sig2")
> abline(v=quantile(sig2,c(0.025,0.5,0.975)))
> abline(v=s2,lty=2)
> olsdata _ summary(olsfit)$coefficients
> olsintervals _ olsdata[,1] + cbind(olsdata[,2]) %*% c(-2,0,2)
> smoothplot(beta[1,],"Intercept")
> abline(v=quantile(beta[1,],c(0.025,0.5,0.975)))
> abline(v=olsintervals[1,c(1,3)],lty=2)
> smoothplot(beta[2,],"Floor")
> abline(v=quantile(beta[2,],c(0.025,0.5,0.975)))
> abline(v=olsintervals[2,c(1,3)],lty=2)
> smoothplot(beta[3,],"Clay")
> abline(v=quantile(beta[3,],c(0.025,0.5,0.975)))
> abline(v=olsintervals[3,c(1,3)],lty=2)
> smoothplot(beta[4,],"Goodhue")
> abline(v=quantile(beta[4,],c(0.025,0.5,0.975)))
> abline(v=olsintervals[4,c(1,3)],lty=2)

> c(sig2=s2,quantile(sig2,c(0.025,0.5,0.975)))
```


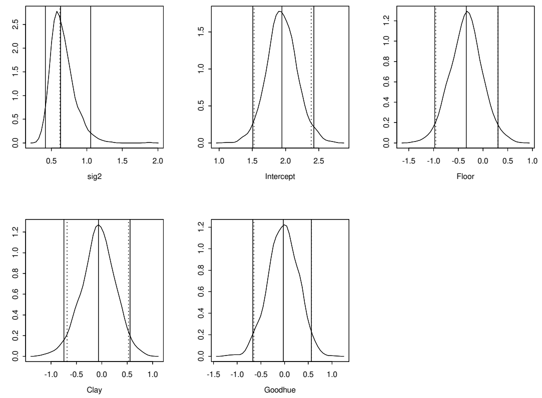

Figure 1: Simulated posteriors for the parameters of the radon regression problem, with noninformative priors. Solid lines indicate equal tail area 95% posterior credible intervals and posterior medians; dotted lines indicate approximate 95% confidence intervals (for β's) or point estimate (for σ^2) from OLS analysis.

```
     sig2      2.5%     50.0%    97.5%
 0.623299 0.4170907 0.6317195 1.053781
> dimnames(olsintervals)[[2]] _ c("bhat-2SE","bhat","bhat+2SE")
> cbind(post025 = apply(beta,1,quantile,0.025),olsintervals,
+ post975=apply(beta,1,quantile,0.975))
               post025   bhat-2SE        bhat  bhat+2SE   post975
(Intercept)  1.5070074  1.5246130  1.95613199 2.3876510 2.4275849
      Floor -0.9767665 -0.9592412 -0.32833983 0.3025615 0.3038866
       Clay -0.7473966 -0.6833730 -0.07980519 0.5237626 0.5542808
    Goodhue -0.6708612 -0.6461490 -0.03794035 0.5702683 0.5657630
```

The results are shown in Figure 1.

Now suppose we wish to predict the radon measurement of another house in Blue Earth county. If the radon concentration is measured in the basement, we can simulate the distribution of the new measurement $\tilde{y}$ as follows

```
> X[1,]
 (Intercept) Floor Clay Goodhue
           1     0    0       0
> errors _ rnorm(1000,0,sqrt(sig2))
> ynew _ c(c(1,0,0,0) %*% beta + errors)
> smoothplot(ynew,"Predicted radon, Blue Earth Basement Meas")
> abline(v=quantile(ynew,c(0.025,0.5,0.975)))
```

```
> abline(v=c(1,0,0,0) %*% bhat +
+ c(-2,2)*sqrt(s2*(1+c(1,0,0,0)%*%XTXI%*%c(1,0,0,0)))
+  , lty=2)
> rbind(Bayes=quantile(ynew,c(0.025,0.975)),OLS= c(1,0,0,0) %*% bhat +
+ c(-2,2)*sqrt(s2*(1+c(1,0,0,0)%*%XTXI%*%c(1,0,0,0)))
+ )
           2.5%    97.5%
Bayes 0.2933040 3.635836
  OLS 0.3192433 3.593021
```

If the measurement is taken on the first floor, the simulated distribution can be obtained as follows

```
> ynew _ c(c(1,1,0,0) %*% beta + errors)
> smoothplot(ynew,"Predicted radon, Blue Earth First Fl. Meas")
> abline(v=quantile(ynew,c(0.025,0.5,0.975)))
> abline(v=c(1,1,0,0) %*% bhat +
+ c(-2,2)*sqrt(s2*(1+c(1,1,0,0)%*%XTXI%*%c(1,1,0,0)))
+  , lty=2)
> rbind(Bayes=quantile(ynew,c(0.025,0.975)),OLS= c(1,1,0,0) %*% bhat +
+ c(-2,2)*sqrt(s2*(1+c(1,1,0,0)%*%XTXI%*%c(1,1,0,0)))
+ )
             2.5%    97.5%
Bayes -0.13944779 3.313950
  OLS -0.09375259 3.349337
```

The results are shown in Figure 2.

4 Bayesian GLS

A Bayesian analysis corresponding to GLS (generalized least squares) proceeds as follows[7]. We will assume that

$$\begin{aligned} \underline{y} \mid \underline{\beta}, \Sigma_y, X &\sim N(X\underline{\beta}, \Sigma_y) \\ \text{and } p(\underline{\beta}|\Sigma_y, X) &\propto \text{constant}. \end{aligned}$$

Let us assume Σ_y is known; this lets us calculate $p(\underline{\beta}|\Sigma_y, \underline{y}, X)$. Let $\Sigma_y^{1/2}$ be a Cholesky square root of Σ_y (i.e., a triangular matrix such that $[\Sigma_y^{1/2}]^T \Sigma_y^{1/2} = \Sigma_y$). Then as with our discussion of WLS, we can premultiply by $\Sigma_y^{-1/2}$ and obtain

$$\Sigma_y^{-1/2}\underline{y} \mid \beta, \Sigma_y \sim N(\Sigma_y^{-1/2} X\underline{\beta}, I)$$

Applying the ordinary regression results above to $\underline{z} = \Sigma_y^{-1/2}\underline{y}$ we immediately obtain

$$\underline{\beta} \mid \Sigma_y, \underline{y}, X \sim N(\hat{\underline{\beta}}, V_\beta)$$

[7]This follows very closely Section 8.7 of Gelman et al.

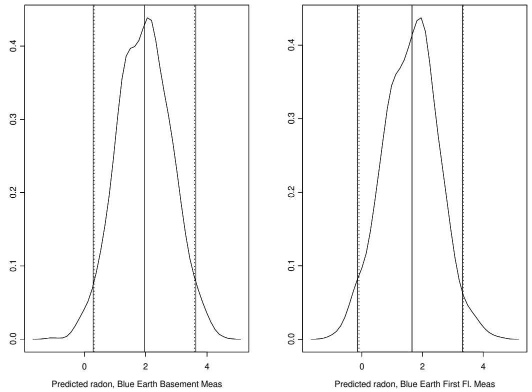

Figure 2: Simulated predictive distributions for a new radon measurement in Blue Earth County: basement measurement (left) and first floor measurement (right). Solid lines indicate equal tail area 95% posterior predictive credible intervals and medians; dotted lines indicate approximate 95% predictive confidence intervals from OLS analysis.

where

$$\begin{aligned} \hat{\underline{\beta}} &= (X^T \Sigma_y^{-1} X)^{-1} X^T \Sigma_y^{-1} \underline{y} \\ V_\beta &= (X^T \Sigma_y^{-1} X)^{-1} \end{aligned}$$

These calculations turn out to be very useful in Section 5 below, where we are considering Bayesian linear regression with informative, conjugate prior information.

For completness we also indicate the form of $p(\Sigma_y|\underline{y}, X)$, but we don't intend to use it in the rest of these notes. As before, we use

$$p(\underline{\beta}, \Sigma_y \mid \underline{y}, X) = p(\underline{\beta} \mid \Sigma_y, \underline{y}, X)\, p(\Sigma_y \mid \underline{y}, X)$$

to see that, for *any* value of $\underline{\beta}$,

$$p(\Sigma_y \mid \underline{y}, X) = \frac{p(\underline{\beta}, \Sigma_y \mid \underline{y}, X)}{p(\underline{\beta} \mid \Sigma_y, \underline{y}, X)} \propto \frac{p(\Sigma_y) N(\underline{y} \mid X\underline{\beta}, \Sigma_y)}{N(\underline{\beta} \mid X\hat{\underline{\beta}}, V_\beta)}$$

where $p(\Sigma_y)$ is now the prior on Σ_y (the prior on $\underline{\beta}$ is still flat). Substituting $\underline{\beta} \equiv \hat{\underline{\beta}}$ again and simplifying, we see that

$$p(\Sigma_y|\underline{y}, X) \propto p(\Sigma_y)|V_\beta|^{1/2} \exp\left[-\frac{1}{2}(\underline{y} - X\hat{\underline{\beta}})^T \Sigma_y^{-1} (\underline{y} - X\hat{\underline{\beta}})\right]. \tag{5}$$

This is relatively easy to write down but difficult to work with—both because $p(\Sigma_y)$ is difficult to specify (or "elicit") and because Σ_y is wrapped up in the computation of both $\hat{\underline{\beta}}$ and V_β—except in very special circumstances, such as WLS (weighted least squares) where at least $\Sigma_y = \sigma^2 \text{diag}(1/w_1, \ldots, 1/w_n)$ only depends on a single parameter.

5 Prior Information

5.1 Prior information for $\underline{\beta}$; Σ_y known

We start by assuming that the variance-covariance matrix Σ_y is known, and that we have multivariate normal prior information on $\underline{\beta}$:

$$\begin{aligned} \underline{y} \mid \underline{\beta}, \Sigma_y &\sim N(X\underline{\beta}, \Sigma_y) \\ \text{and } \underline{\beta} &\sim N(\underline{\beta}_0, \Sigma_\beta) \end{aligned}$$

where Σ_y, $\underline{\beta}_0$, and Σ_β are known and fixed in advance.

Aside: Another way to say the same thing is that

$$\left.\begin{aligned} \underline{y} &= X\underline{\beta} + \underline{\epsilon}, \quad \underline{\epsilon} \sim N(\underline{0}, \Sigma_y) \\ \underline{\beta} &= \underline{\beta}_0 + \underline{\delta}, \quad \underline{\delta} \sim N(\underline{0}, \Sigma_\beta) \end{aligned}\right\} \tag{6}$$

from which it immediately follows that

$$\begin{aligned} \underline{y} &= X\underline{\beta}_0 + X\underline{\delta} + \underline{\epsilon} \\ \text{i.e. } \underline{y} &\sim N(X\underline{\beta}_0, X\Sigma_\beta X^T + \Sigma_y) \end{aligned}$$

- If $\underline{\beta}_0$ were *not* known, it could now be estimated by (non-Bayesian) WLS. If we then proceed with a Bayesian analysis for $\underline{\beta}$ pretending that $\underline{\beta}_0$ was fixed in advance to be $\hat{\underline{\beta}}_0$ from the WLS analysis, then the resulting analysis would be an example of an *empirical Bayes* analysis.
- If $\underline{\beta}_0$ had the further linear structure $\underline{\beta}_0 = Z\underline{\gamma}$, the resulting model (6) would be called a (two-level) *hierarchical linear model* (HLM).

End of vocabulary lesson.

Continuing under the assumption that Σ_y, $\underline{\beta}_0$, and Σ_β are known and fixed in advance, we know that

$$p(\underline{\beta}|\Sigma_y, \underline{y}, X) \propto N(y|X\underline{\beta}, \Sigma_y) \cdot N(\underline{\beta}|\underline{\beta}_0, \Sigma_\beta)$$

and so

$$\begin{aligned} \log p(\underline{\beta}|\Sigma_y, \underline{y}, X) &= (\text{Constant}) - \frac{1}{2}(y - X\underline{\beta})^T \Sigma_y^{-1}(y - X\underline{\beta}) - \frac{1}{2}(\underline{\beta} - \underline{\beta}_0)^T \Sigma_\beta^{-1}(\underline{\beta} - \underline{\beta}_0) \\ &= (\text{Constant}) - \frac{1}{2}\left(\begin{bmatrix} \underline{y} \\ \underline{\beta}_0 \end{bmatrix} - \begin{bmatrix} X\underline{\beta} \\ I\underline{\beta} \end{bmatrix}\right)^T \begin{pmatrix} \Sigma_y^{-1} & 0 \\ 0 & \Sigma_\beta^{-1} \end{pmatrix} \left(\begin{bmatrix} \underline{y} \\ \underline{\beta}_0 \end{bmatrix} - \begin{bmatrix} X\underline{\beta} \\ I\underline{\beta} \end{bmatrix}\right) \\ &= (\text{Constant}) - \frac{1}{2}(\underline{y}_* - X_*\underline{\beta})^T \Sigma_*^{-1}(\underline{y}_* - X_*\underline{\beta}) \end{aligned}$$

where the "constant's" do not depend on $\underline{\beta}$, where I is $(k+1) \times (k+1)$ identity matrix, and where

$$\underline{y}_* = \begin{bmatrix} \underline{y} \\ \underline{\beta}_0 \end{bmatrix} \qquad X_* = \begin{bmatrix} X \\ I \end{bmatrix} \qquad \Sigma_* = \begin{pmatrix} \Sigma_y & 0 \\ 0 & \Sigma_\beta \end{pmatrix}$$

are "augmented" data and design matrices. In other words *regression with prior information on $\underline{\beta}$ is like regression with the "augmented" data $\underline{y}_*$ and design matrix X_* and noninformative prior on $\underline{\beta}$, and known variance-covariance matrix* Σ_*. In other words, the "prior" information can be interpreted as additional "observations" in a noninformative-prior Bayesian analysis of the linear regression model

$$\underline{y}_* = X_*\underline{\beta} + \underline{\epsilon}_*, \quad \underline{\epsilon}_* \sim N(\underline{0}, \Sigma_*)$$

We can immediately apply the results on "Bayesian GLS" analysis (Section 4) to conclude that

$$\underline{\beta}|\Sigma_y, \underline{y}, X \sim N(\hat{\underline{\beta}}, V_\beta)$$

with

$$\begin{aligned} \hat{\underline{\beta}} &= (X_*^T \Sigma_*^{-1} X_*)^{-1} X_*^T \Sigma_*^{-1} y_* \\ &= \left(\begin{bmatrix} X^T & I^T \end{bmatrix} \begin{bmatrix} \Sigma_y^{-1} & 0 \\ 0 & \Sigma_\beta^{-1} \end{bmatrix} \begin{bmatrix} X \\ I \end{bmatrix}\right)^{-1} \begin{bmatrix} X^T & I^T \end{bmatrix} \begin{bmatrix} \Sigma_y^{-1} & 0 \\ 0 & \Sigma_\beta^{-1} \end{bmatrix} \begin{bmatrix} \underline{y} \\ \underline{\beta}_0 \end{bmatrix} \\ &= \left(X^T \Sigma_y^{-1} X + \Sigma_\beta^{-1}\right)^{-1} \left[X^T \Sigma_y^{-1} \underline{y} + \Sigma_\beta^{-1} \underline{\beta}_0\right] \end{aligned}$$

$$\begin{aligned} \text{and } V_\beta &= (X_*^T \Sigma_*^{-1} X_*)^{-1} \\ &= \left(X^T \Sigma_y^{-1} X + \Sigma_\beta^{-1}\right)^{-1} \end{aligned}$$

If some of the β_j's should have noninformative prior distributions, we set the corresponding rows and columns of Σ_*^{-1} (or equivalently, of Σ_β^{-1}) equal to all zeros.

The case of ordinary regression with iid priors on $\underline{\beta}$ gives some insight into the "shrinkage" interpretation of Bayesian estimates. This case occurs when

$$\Sigma_y = \sigma^2 I_{n \times n} \qquad \Sigma_\beta = \tau^2 I_{(k+1)\times(k+1)}$$

so that the posterior mean and variance reduce to

$$\begin{aligned} \hat{\underline{\beta}} &= \left(X^T X + \frac{\sigma^2}{\tau^2} I\right)^{-1} \left[X^T \underline{y} + \frac{\sigma^2}{\tau^2} \underline{\beta}_0\right] \\ \text{and } V_\beta &= \sigma^2 \left(X^T X + \frac{\sigma^2}{\tau^2} I\right)^{-1} \end{aligned}$$

There are two interesting things to note about this formulation:

- As the prior information gets weaker ($\tau^2 \to \infty$), $\hat{\underline{\beta}} \to (X^T X)^{-1} X^T \underline{y}$, the classical estimate, as in Section 3; and as the prior information gets stronger ($\tau^2 \to 0$), $\hat{\underline{\beta}} \to \underline{\beta}_0$, the prior mean. This is the usual "shrinkage" behavior.
- When $\underline{\beta}_0 = \underline{0}$, the estimate is of the form

  $$\hat{\underline{\beta}} = \left(X^T X + \lambda I\right)^{-1} X^T \underline{y}$$

  (with $\lambda = \sigma^2/\tau^2$), i.e. it is the *ridge regression* estimate.

5.2 Prior information for Σ_y

Now that we have beaten $p(\underline{\beta}|\Sigma_y, \underline{y}, X) = N(\underline{\beta}|\hat{\underline{\beta}}, V_\beta)$ into the ground, let's complete the formulation of a joint posterior by specifying $p(\Sigma_y|\underline{y}, X)$. We will only do this in the case of ordinary regression, $\Sigma_y = \sigma^2 I$, with a general normal prior $\underline{\beta} \sim N(\underline{\beta}_0, \Sigma_\beta)$. For this case, we saw in Section 3 that the natural posterior with a non-informative prior was an Inv-χ^2. Therefore it will not be a surprise that the conjugate prior family for σ^2 in Bayesian ordinary regression is

$$\text{Inv-}\chi^2\,(\sigma^2|n_0, \sigma_0^2).$$

Following the same calculations as before [see especially Equation (5)], using the joint prior

$$p(\underline{\beta}, \sigma^2|X) \;=\; N(\underline{\beta}|\underline{\beta}_0, \Sigma_\beta)\cdot \text{Inv-}\chi^2\,(\sigma^2|n_0, \sigma_0^2)$$

it follows that

- $p(\underline{\beta} \mid \sigma^2, \underline{y}, X) = N(\underline{\beta}|\hat{\underline{\beta}}, V_\beta)$ where

$$\hat{\underline{\beta}} \;=\; (X_*^T\Sigma_*^{-1}X_*)^{-1}X_*^T\Sigma_*^{-1}y_* \;=\; \left(X^TX/\sigma^2+\Sigma_\beta^{-1}\right)^{-1}\left[X^T\underline{y}/\sigma^2+\Sigma_\beta\underline{\beta}_0\right] \tag{7}$$

$$V_\beta \;=\; (X_*^T\Sigma_*^{-1}X_*)^{-1} \;=\; \left(X^TX/\sigma^2+\Sigma_\beta^{-1}\right)^{-1} \tag{8}$$

  and

- $p(\sigma^2|\underline{y}, X)$ is given by

$$p(\sigma^2|\underline{y}, X) \;=\; \text{Inv-}\chi^2\left(\sigma^2 \,\middle|\, n_0+n', \frac{n_0\sigma_0^2+n's^2}{n_0+n'}\right) \tag{9}$$

  where $s^2 = (\underline{y}_* - X_*\hat{\underline{\beta}})^T\Sigma_*^{-1}(\underline{y}_* - X_*\hat{\underline{\beta}})/n'$; $n' = n_* + k - 1$; and $n_* = rank(\Sigma_*) = n + rank(\Sigma_\beta^{-1})$ is typically $n + k + 1$. Once again, the effect of the informative conjugate prior is similar to adding n_0 additional observations' worth of data to the non-informative analysis.

5.3 Continuing the radon example

Suppose our radon measurement data were a replication of an earlier study of 41 houses in the same three counties. Suppose moreover that an OLS regression had been performed on the earlier data set, with the results

```
> summary(oldfit)

Call: lm(formula = LogRadon ~ Floor + Clay + Goodhue, data = oldrad)
Residuals:
    Min      1Q   Median   3Q   Max
 -1.853 -0.6117 -0.08181 0.48 1.592

Coefficients:
              Value Std. Error t value Pr(>|t|)
```



```
(Intercept)  1.9987  0.2218     9.0128  0.0000
      Floor -0.3447  0.3161    -1.0905  0.2826
       Clay  0.0937  0.3216     0.2915  0.7723
    Goodhue  0.1815  0.3065     0.5924  0.5572

Residual standard error: 0.8232 on 37 degrees of freedom
Multiple R-Squared: 0.04275
F-statistic: 0.5508 on 3 and 37 degrees of freedom, the p-value is 0.6508

Correlation of Coefficients:
        (Intercept)   Floor    Clay
  Floor -0.2851
   Clay -0.5961     -0.1311
Goodhue -0.6816      0.0589  0.4507
```

For our current data we might like to specify prior distributions based on the earlier study:

$$\underline{\beta} \;\sim\; N(\hat{\underline{\beta}}_{old}, s_{old}^2(X_{old}^TX_{old})^{-1})$$
$$\sigma^2 \;\sim\; \text{Inv-}\chi^2\,(41, s_{old}^2)$$

These can be obtained as follows

```
> Xold _ model.matrix(oldfit)
> XTXIO _ solve(t(Xold)%*%Xold)
> s2old _ sum(residuals(oldfit)^2)/37
> beta0 _ summary(oldfit)$coefficients[,1]
> Sigbeta _ XTXIO*s2old
> SigbetaI _ solve(Sigbeta)
> n0 _ 41
```

Following Equations (7), (8) and (9), and noting that $n' = n^* - (k+1) = n + (k+1) - (k+1) = n$, we may set up a simulation for the posterior distribution of the parameters again as follows (though this time, not thinking much about computational efficiency):

```
> V _ rchisq(1000,41+41)
> sig2 _ (41*s2old + 41*s2)/V # inv-chisq(41+41,(41*s2old + 41*s2)/(41+41))
> X _ model.matrix(olsfit)
> XTX _ t(X)%*%X
> XTy _ t(X)%*%radon$LogRadon
> SbIbeta0 _ SigbetaI%*%beta0
> beta _ matrix(0,ncol=1000,nrow=4)
> for (i in 1:1000) {
+ Vbeta _ solve(XTX/sig2[i] + SigbetaI)
+ Vbeta _ (Vbeta + t(Vbeta))/2      # guarantee symmetry (numerical instabilities)
+ Ch _ t(chol(Vbeta))               # look at help(chol)
+ bmean _ Vbeta %*% (XTy/sig2[i] + SbIbeta0)
+ beta[,i] _ bmean + Ch%*%rnorm(4)
+ }
```

The equal-tail 95% posterior credible intervals here can now be compared with the corresponding intervals for the non-informative prior analysis. For σ^2 we obtain

```
NONINFORMATIVE:
> quantile(sig2,c(0.025,0.5,0.975))
      2.5%     50.0%    97.5%
 0.4170907 0.6317195 1.053781

INFORMATIVE:
> quantile(sig2,c(0.025,0.5,0.975))
      2.5%     50.0%     97.5%
 0.4879442 0.6552671 0.9014055
```

and for the regression coefficients $\underline{\beta}$ the equal-tail 95% credible intervals look like this:

```
NONINFORMATIVE:
> t(apply(beta,1,quantile,c(0.025,0.975)))
                  2.5%     97.5%
(Intercept)  1.5070074 2.4275849
      Floor -0.9767665 0.3038866
       Clay -0.7473966 0.5542808
    Goodhue -0.6708612 0.5657630

INFORMATIVE:
> t(apply(beta,1,quantile,c(0.025,0.975)))
                  2.5%      97.5%
(Intercept)  1.6968706 2.28816000
      Floor -0.7624844 0.09801322
       Clay -0.4491324 0.43696194
    Goodhue -0.3421766 0.53141934
```