Bayesian Regressions in Experimental Design

Stanley Sawyer — Washington University — Vs. April 9, 2008

1. Introduction. The purpose here is to derive a formula for the posterior probabilities (given observed data) that one-, two-, or three-factor submodels have all of the active factors in an experimental design (Box and Meyer 1993, Box, Hunter, and Hunter 2005, Appendix 7A). Box *et al.* (2005) (Chapter 7) has references to FORTRAN and R programs to evaluate the posterior likelihoods. This note is a companion to a similar program written in C.

In general in experimental design models, one considers observations of a response variable Y_i $(1 \le i \le n)$ for n different settings of f binary (High,Low) factors $A, B, C, D, E, \ldots = A_1, A_2, \ldots, A_f$ with $f < n$. A regression model can be written

$$\begin{aligned} Y_i &= \beta_0 + \sum_{j=1}^{f} X_{ij}\beta_j + e_i, \qquad e_i \sim N(0,\sigma^2), \quad 1 \le i \le n \qquad (1.1) \\ &= \sum_{j=0}^{m-1} X_{ij}\beta_j + e_i, \quad m = f+1, \quad m \le n \end{aligned}$$

where the $n \times m$ matrix X satisfies $X_{ij} = \pm 1$ and $X_{i0} \equiv 1$ (the intercept term). Thus the first column $X_{i0} \equiv 1$ and the remaining $f = m-1$ columns correspond to the factors $A_1, \ldots, A_f$.

The columns of the matrix X are assumed to be orthogonal. In general, an $n \times m$ matrix with ± 1 entries and orthogonal columns is called a *Hadamard* matrix. This is equivalent to $X_{ij} = \pm 1$ and $X'X = nI_n$. It follows that the least squares and maximum likelihood estimator of the coefficients β_j in (1.1) is

$$\widehat{\beta} = (X'X)^{-1}X'Y = (1/n)X'Y \qquad (1.2)$$

Thus $\widehat{\beta}_i$ depends only on X only through the i^{th} row of X. This implies that one can fill out the matrix X with additional orthogonal columns j with $m \le j \le n-1$ in (1.1) with $X_{ij} = \pm 1$ and estimated $n-m$ additional parameters $\beta_m, \ldots, \beta_{n-1}$ without affecting the estimates of $\beta_0, \ldots, \beta_{m-1}$. In *fractional factorial* designs (see below), the additional parameters give estimates of or information about two-way and higher interactions of the factors $A_1, \ldots, A_f$. In other Hadamard designs, the additional parameters are not easily identifiable as interactions but provide information about the error variance σ^2.

2. More Properties of Experimental Designs. For $1 \le j \le f$ in (1.1), each of the factors A_j is assumed to be either *active* or *inert*. A factor A is *active* if the factor is significant either as a main effect or else as a significant interaction with other factors. Otherwise, A is inert, which means that it has no significant main effect or interactions, and its (High,Low) settings in (1.1) and (1.2) do not have a significant effect on the response variables Y_i. Typically, a minority of the factors A_j are active and the rest are inert. A first major aim of an analysis of an experimental design is to identify the active factors with a minimum number of experimental runs. A second main aim is to identify the significant main effects and interactions among the active factors.

Since $X_{ij} = \pm 1$ with $\sum_{i=1}^{n} X_{ij} = 0$ for $j > 0$, interactions of main effects or interactions of factors can be represented in the regression (1.1) as columns that are component-wise products of the corresponding component columns. The orthogonality property (1.2) means that estimates of interaction effects can be written down independently of estimating the main effects of the factors in (1.1). In fractional factorial designs, these product columns are always other columns in the $n \times n$ matrix X. In other Hadamard designs, such product columns are generally are not among the other columns, and may not be orthogonal to the first m columns.

In *(full or fractional) factorial* models (with one observation per cell), $m = n = 2^d$ for $d \le f$ with columns $f+1 \le j \le m-1$ in (1.1) corresponding to second-order and higher interactions among the f factors. The columns of X are assumed to be closed under component-wise multiplication, so that there are columns for any set of interactions of the other columns. The model is *full factorial* if $f = d$, which means that the 2^f columns of X correspond to all main and interaction effects among the f factors along with the intercept. The model is *fractional factorial* if $d < f \le n-1$. This is called a 2^{f-a} design for $a = f - d$, or equivalently a 2^{-a}-*fraction* of a full factorial 2^d design. In full or fractional factorial designs, the columns of the matrix X in (1.1) form a cyclic Abelian group of order two with $n = 2^d$ elements. In fractional factorial designs, the factors A_j for $d < j \le f$ are *confounded* with interactions of the first d factors. More generally, each parameter in (1.1) is a linear combination with ± 1 coefficients of 2^a parameters corresponding to some subset of size 2^a of the 2^f possible main and interaction effects among the f factors. For either a full factorial or fractional factorial model, the goal is to quickly identify the active factors $A_{a_1}, \ldots, A_{a_b}$ for $1 \le a_k \le f$ and then re-interpret (1.1) as 2^b full factorial design with $n/2^b = 2^{f-a-b}$ observations per cell.

The *resolution* (R) of a factorial design (1.1) is the size of the lower-order interaction among the factors that is confounded with the intercept.

That means that at least one main effect is confounded with an $(R-1)^{\text{th}}$-order interaction. Thus X is a Resolution 3 design if and only if some main factor is confounded with some two-way interaction, and is a Resolution 4 (or higher) design if and only if this does not happen.

The *projectivity* P of a design is that largest value of P such that, for any subset $A_{c_1},\ldots,A_{c_P}$ of the f factors, the columns of X in (1.1) are sufficient to distinguish all of the main effects and interactions among $A_{c_1},\ldots,A_{c_P}$. For a fractional factorial design, that is equivalent to saying that the columns of X contain a full factorial 2^P design for any set of P factors. It is not difficult to show that $P = R-1$ for any Hadamard design (1.1). In particular, if a design is of resolution 3, then main effects are confounded with two-way interactions. If a design is of resolution 4, then main effects are not confounded with any two-way interactions, but two-way interactions can be confounded with one another.

3. Submodels of an Experimental Design. In general, for any set of factors $M = \{\, A_{c_1},\ldots,A_{c_r}\,\}$ in (1.1), one can consider the *submodel*

$$Y_i \;=\; \sum_{j=0}^{m_M-1} X_{ij}^M \beta_j^M + e_i^M, \qquad e_i^N \sim N\bigl(0,(\sigma^M)^2\bigr) \tag{3.1}$$

with $m_M = 2^r$ in which (3.1) (or a subset of the observations in (3.1)) is a full factorial design with r factors. If X corresponds to a factorial or fractional factorial design, the columns of X^M are a subset of the columns in the original model (1.1). In other cases, for example *Plackett-Burman* or *Marshall Hall* designs, this is not the case and X^M need not be Hadamard for $r \ge 3$. Note that the same response values Y_i are used in (1.1) and (3.1), so that no new observations must be made.

With fractional factorial designs, the either full or nonexistent confounding between different main effects and interactions often means that it is often possible to identify the active factors from a normal plot of the estimated parameters in (1.1) if there are no significant two-way interactions. It is generally assumed that 3-way or higher interactions are not significant. Otherwise, different methods may have to be used.

A more computationally intensive, but still practical, method for identifying the active factors is to use model selection techniques for linear regressions. Specifically, one looks for the best-fitting model of type (3.1) with the number of factors satisfying e.g. $r \le 3$, with appropriate penalties for higher r. Among classical model selection methods, finding the model with the largest model F-statistic seems to work better than maximum R^2 or adjusted R^2. However, a Bayesian regression model due to Box and Meyer (1993) appears to work better yet, at least in particular examples.

4. A Bayesian Model-selection Method for 2^r Submodels. Given $f \le m-1$ factors being screened in (1.1), there are

$$n_T = f + \binom{f}{2} + \binom{f}{3} \tag{4.1}$$

submodels of the form (3.1) with $r \le 3$. Given the model M, the likelihood of the parameters β^M, σ^M in (3.1) is of the form

$$\begin{aligned} L(\beta, \sigma, Y) &= \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2} \left(Y_i - \sum_{j=0}^{m} X_{ij}\beta_j \right)^2 \right) \\ &= \frac{1}{(2\pi\sigma^2)^{m/2}} \exp\left(-\frac{1}{2\sigma^2} (Y - X\beta)'(Y - X\beta) \right) \end{aligned}$$

corresponding to the assumption $Y_i \sim N\big((X\beta)_i, \sigma^2\big)$ in a classical linear regression, suppressing superscripts M for easy of notation. The next tasks are to prescribe reasonable priors for M, β and σ, write down the full likelihood, and then find the posterior distributions of the models M in (3.1) given the response variables Y_i in (1.1).

The first assumption for models M is essentially to assume that the r factors in the model are *active* (or are potentially active) and the remaining $f-r$ factors are *inert*. The prior distribution $\pi_0(M)$ models M corresponds to the f factors in (1.1) being *active* with probability π for some value π and *inert* with probability $1-\pi$, with independence for the f factors. This means that the prior for models is

$$\pi_0(M) = C_1 \pi^r (1-\pi)^{f-r} = C_1 (1-\pi)^f \left(\frac{\pi}{(1-\pi)} \right)^r \tag{4.2}$$

where C_1 is determined by $\sum_M \pi_0(M) = 1$. This depends only on the size $m_M = 2^r$ of the model in (3.1).

Conditional on the model M, we assume that β^M and σ^M have the prior distributions

$$\begin{aligned} &\beta_0^M \sim N(0, \sigma^2/\epsilon), \qquad \epsilon = 10^{-6} \qquad \text{(Intercept)} \\ &\beta_j^M \sim N(0, \gamma^2\sigma^2), \qquad j \ge 1, \\ &\sigma^M \sim g_\sigma(y), \qquad g_\sigma(y) = C/y^a \end{aligned} \tag{4.3}$$

for some $\gamma > 0$, assuming an infinite improper prior for σ for some $a > 0$. (Box and Meyer essentially assume $a = 0$.) Thus the priors for M, β^M, σ^M

depend on two parameters π, γ. Box, Hunter, and Hunter (2005, Chapter 7) suggest $\pi \approx 0.25$ and γ in the range 2–3.

By (4.3) and (3.1), for $1 \le j \le m-1$ and β_q^M known to be inert,

$$\frac{E\big((\widehat{\beta}_j^M)^2\big)}{E\big((\widehat{\beta}_q^M)^2\big)} \;=\; \frac{\gamma^2\sigma^2 + (1/n)\sigma^2}{(1/n)\sigma^2} \;=\; n\gamma^2 + 1$$

where $\widehat{\beta}_q^M$ represents a parameter estimate of an inert factor. Since

$$\gamma \;=\; \frac{1}{\sqrt{n}}\sqrt{\frac{E\big((\widehat{\beta}_j^M)^2\big)}{E\big((\widehat{\beta}_q^M)^2\big)} \;-\; 1} \tag{4.4}$$

the parameter γ is related to the ratios of expected squares of estimates of active and inert factors.

The full likelihood (with Bayesian priors) can be written

$$\begin{aligned} L(M,\beta,\sigma,Y) \;&=\; C_2\left(\frac{\pi}{1-\pi}\right)^r \frac{1}{\sigma^a} \\ &\times\; \frac{\sqrt{\epsilon}}{\sqrt{2\pi\sigma^2}}\exp\left(-\epsilon\frac{\beta_0^2}{2\sigma^2}\right)\left(\frac{1}{\sqrt{2\pi\gamma^2\sigma^2}}\right)^{\frac{m-1}{2}} \exp\left(-\frac{1}{2\gamma^2\sigma^2}\sum_{j=1}^m \beta_j^2\right) \\ &\times\; \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^{\frac{n}{2}} \exp\left(-\frac{1}{2\sigma^2}(Y-X\beta)'(Y-X\beta)\right) \end{aligned} \tag{4.5}$$

Thus

$$\begin{aligned} L(M,\beta,\sigma,Y) \quad &=\; C\left(\frac{\pi}{1-\pi}\right)^r \left(\frac{1}{\gamma}\right)^{m-1} \\ &\qquad \times\; \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^{\frac{n+m+a}{2}} \exp\left(-\frac{1}{2\sigma^2}Q(\beta)\right) \end{aligned} \tag{4.6}$$

where $m = 2^r$ and

$$Q(\beta) \;=\; (Y-X\beta)'(Y-X\beta) \;+\; \beta'\Gamma\beta \tag{4.7}$$

for the $m \times m$ matrix

$$\Gamma \;=\; \begin{pmatrix} \epsilon & 0 \\ 0 & \frac{1}{\gamma^2} I_{m-1} \end{pmatrix} \tag{4.8}$$

5. Simplifying the Bayesian Likelihod. The purpose here is to show that, for the likelihood $L(M,\beta,\sigma,Y)$ in (4.5), the marginal likelihood

$$\begin{aligned} L(M,Y) \;&=\; \int_0^\infty \int_{R^m} L(M,\sigma,Y)\,d\beta\,d\sigma & (5.1)\\ &=\; C_1\left(\frac{\pi}{1-\pi}\right)^r \left(\frac{1}{\gamma}\right)^{m-1} \frac{\det\Big(X'X+\Gamma\Big)^{-1/2}}{Q(c_0)^{(n-1+a)/2}} \end{aligned}$$

for Γ in (4.8), $Q(\beta)$ in (4.7), and $c_0 = (X'X+\Gamma)^{-1}X'Y$. In particular, the posterior probabilities of the models M in (3.1) given the data Y_i in (1.1) is

$$L(M\mid Y) \;=\; \frac{L(M,Y)}{\sum_W L(W,Y)} \;=\; C_2\,L(M,Y)$$

for $L(M,Y)$ as in (5.1). For $a=0$, this corresponds to equation (A.1) in Box *et al.* (2005, Appendix 7A).

First, expanding the quadratic form (4.7)

$$\begin{aligned} Q(\beta) \;&=\; (Y-X\beta)'(Y-X\beta) \;+\; \beta'\Gamma\beta \\ &=\; (X\beta-Y)'(X\beta-Y) \;+\; \beta'\Gamma\beta \\ &=\; \beta'(X'X+\Gamma)\beta \;-\; 2\beta'X'Y \;+\; Y'Y & (5.2) \end{aligned}$$

For any vector $c\in R^m$

$$\begin{aligned} &(\beta-c)'(X'X+\Gamma)(\beta-c) & (5.3)\\ &\quad=\; \beta'(X'X+\Gamma)\beta \;-\; 2\beta'(X'X+\Gamma)c \;+\; c'(X'X+\Gamma)c \end{aligned}$$

If $c_0 = (X'X+\Gamma)^{-1}X'Y$, the first two terms in (5.2) and (5.3) are the same and

$$\begin{aligned} Q(\beta) \;&=\; (Y-X\beta)'(Y-X\beta) \;+\; \beta'\Gamma\beta & (5.4)\\ &=\; (\beta-c_0)'(X'X+\Gamma)(\beta-c_0) \;+\; Y'Y \;-\; c_0'(X'X+\Gamma)c_0 \\ &=\; (\beta-c_0)'(X'X+\Gamma)(\beta-c_0) \;+\; Q(c_0) \end{aligned}$$

This implies

$$\begin{aligned} &\min_\beta\Big((Y-X\beta)'(Y-X\beta)+\beta'\Gamma\beta\Big) \\ &\quad=\; Q(c_0) \;=\; (Y-Xc_0)'(Y-Xc_0)+c_0'\Gamma c_0 \\ &\quad=\; Y'Y - c_0'(X'X+\Gamma)c_0 \;=\; Y'\Big(I_n - X(X'X+\Gamma)^{-1}X'\Big)Y \end{aligned}$$

Thus $Q(c_0)$ is the Bayesian analog of the error sum of squares in a classical regression. By (4.6), $\beta = c_0$ is both the posterior mean and the posterior mode of the density $L(M, \beta, \sigma, Y)$. However, we are more interested in the marginal likelihood $L(M, Y)$ and the corresponding marginal posterior distribution $L(M \mid Y)$ averaged over all possible values of β and σ. By (4.6) and (5.4)

$$L(M, \beta, \sigma, Y) \;=\; C\left(\frac{\pi}{1-\pi}\right)^r \left(\frac{1}{\gamma}\right)^{m-1} \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^{\frac{n+m+a}{2}} \tag{5.5}$$
$$\times\; \exp\left(-\frac{1}{2\sigma^2}(\beta - c_0)'(X'X + \Gamma)(\beta - c_0)\right) \exp\left(-\frac{1}{2\sigma^2}Q(c_0)\right)$$

In general, if A is any $m \times m$ positive definite matrix

$$\int_{R^m} \exp\left(-\frac{1}{2\sigma^2}(\beta - c)'A(\beta - c)\right) d\beta \;=\; \int_{R^m} \exp\left(-\frac{1}{2\sigma^2}\beta' A \beta\right) d\beta$$
$$=\; \int_{R^m} \exp\left(-\frac{1}{2\sigma^2}\beta' U' D U \beta\right) d\beta$$

where $A = U'DU$ for an orthogonal matrix U and diagonal matrix $D = \text{diag}(\lambda_1, \ldots, \lambda_m)$ with $\lambda_i > 0$. Thus the integral above equals

$$\prod_{i=1}^m \left(\int \exp\left(-\frac{1}{\sigma^2}\beta_i^2 \lambda_i\right) d\beta_i\right) \;=\; \prod_{i=1}^m \sqrt{\frac{2\pi\sigma^2}{\lambda_i}}$$
$$=\; \left(2\pi\sigma^2\right)^{m/2} \det(A)^{-1/2} \tag{5.6}$$

since $\det(A) = \det(UDU') = \det(D)$. Thus by (5.5) and (5.6)

$$L(M, \sigma, Y) \;=\; \int_{R^n} L(M, \beta, \sigma, Y)\, d\beta \;=\; C\left(\frac{\pi}{1-\pi}\right)^r \left(\frac{1}{\gamma}\right)^{m-1} \tag{5.7}$$
$$\times\; \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^{\frac{n+a}{2}} \det\Big(X'X + \Gamma\Big)^{-1/2} \exp\left(-\frac{1}{2\sigma^2}Q(c_0)\right)$$

Finally

$$L(M, Y) \;=\; \int_0^\infty L(M, \sigma, Y)\, d\sigma \;=\; C_1\left(\frac{\pi}{1-\pi}\right)^r \left(\frac{1}{\gamma}\right)^{m-1}$$
$$\times\; \det\Big(X'X + \Gamma\Big)^{-1/2} \;\times\; \frac{1}{Q(c_0)^{(n-1+a)/2}}$$

since

$$\int_0^\infty \left(\frac{1}{\sigma^2}\right)^{\frac{n+a}{2}} \exp\left(-\frac{1}{2\sigma^2}Q(c_0)\right)\, d\sigma$$
$$= \int_0^\infty \left(\frac{1}{y^2}\right)^{\frac{n+a}{2}} \exp\left(-\frac{1}{2y^2}\right)\, dy \left(\frac{1}{Q(c_0)^{\frac{n-1+a}{2}}}\right)$$

where the integral equals

$$\int_0^\infty (2v)^{\frac{n+a-1}{2}-1} e^{-v}\, dv \;=\; 2^{\frac{n+a-3}{2}} \Gamma\left(\frac{n+a-1}{2}\right)$$

References.

1. Box, George E. P., J. Stuart Hunter, and William G. Hunter (2005) Statistics for Experimenters: Design, Innovation, and Discovery, 2nd Edition. John Wiley and Sons. ISBN13 978-0471-71813-0.

2. Box, G. E. P., and Meyer, R. D. (1993) Finding the active factors in fractionated screening experiments. *J. Quality Technol.* **25,** 94–105.