for design and randomization in crossover studies.

We round out the issue with regular features that include the R Help Desk, in which Duncan Murdoch joins Uwe Ligges to explain the arcane art of building R packages under Windows, descriptions of changes in the 2.2.0 release of R and recent changes on CRAN and an announcement of an upcoming event, useR!2006. Be sure to read all the way through to the useR! announcement. If this meeting is anything like the previous useR! meetings – and it will be – it will be great! Please do consider attending.

This is my last issue on the *R News* editorial board and I would like to take this opportunity to extend my heartfelt thanks to Paul Murrell and Torsten Hothorn, my associate editors. They have again stepped up and shouldered the majority of the work of preparing this issue while I was distracted by other concerns. Their attitude exemplifies the spirit of volunteerism and helpfulness that makes it so rewarding (and so much fun) to work on the R Project.

Douglas Bates
University of Wisconsin – Madison, U.S.A.
bates@R-project.org

BMA: An R package for Bayesian Model Averaging

by Adrian E. Raftery, Ian S. Painter and Christopher T. Volinsky

Bayesian model averaging (BMA) is a way of taking account of uncertainty about model form or assumptions and propagating it through to inferences about an unknown quantity of interest such as a population parameter, a future observation, or the future payoff or cost of a course of action. The BMA posterior distribution of the quantity of interest is a weighted average of its posterior distributions under each of the models considered, where a model's weight is equal to the posterior probability that it is correct, given that one of the models considered is correct.

Model uncertainty can be large when observational data are modeled using regression, or its extensions such as generalized linear models or survival (or event history) analysis. There are often many modeling choices that are secondary to the main questions of interest but can still have an important effect on conclusions. These can include which potential confounding variables to control for, how to transform or recode variables whose effects may be nonlinear, and which data points to identify as outliers and exclude. Each combination of choices represents a statistical model, and the number of possible models can be enormous.

The R package `BMA` provides ways of carrying out BMA for linear regression, generalized linear models, and survival or event history analysis using Cox proportional hazards models. The functions `bicreg`, `bic.glm` and `bic.surv`, account for uncertainty about the variables to be included in the model, using the simple BIC (Bayesian Information Criterion) approximation to the posterior model probabilities. They do an exhaustive search over the model space using the fast leaps and bounds algorithm. The function `glib` allows one to specify one's own prior distribution. The function `MC3.REG` does BMA for linear regression using Markov chain Monte Carlo model composition (MC^3), and allows one to specify a prior distribution and to make inference about the variables to be included and about possible outliers at the same time.

Basic ideas

Suppose we want to make inference about an unknown quantity of interest Δ, and we have data D. We are considering several possible statistical models for doing this, $M_1, \ldots, M_K$. The number of models could be quite large. For example, if we consider only regression models but are unsure about which of p possible predictors to include, there could be as many as 2^p models considered. In sociology and epidemiology, for example, values of p on the order of 30 are not uncommon, and this could correspond to around 2^{30}, or one billion models.

Bayesian statistics expresses all uncertainties in terms of probability, and make all inferences by applying the basic rules of probability calculus. BMA is no more than basic Bayesian statistics in the presence of model uncertainty. By a simple application of the law of total probability, the BMA posterior distribution of Δ is

$$p(\Delta|D) = \sum_{k=1}^{K} p(\Delta|D, M_k)p(M_k|D), \qquad (1)$$

where $p(\Delta|D, M_k)$ is the posterior distribution of Δ given the model M_k, and $p(M_k|D)$ is the posterior probability that M_k is the correct model, given that one of the models considered is correct. Thus the BMA posterior distribution of Δ is a weighted average of the posterior distributions of Δ under each of the models, weighted by their posterior model probabilities. In (1), $p(\Delta|D)$ and $p(\Delta|D, M_k)$ can be prob-

ability density functions, probability mass functions, or cumulative distribution functions.

The posterior model probability of M_k is given by

$$p(M_k|D) = \frac{p(D|M_k)p(M_k)}{\sum_{\ell=1}^{K} p(D|M_\ell)p(M_\ell)}. \quad (2)$$

In equation (2), $p(D|M_k)$ is the *integrated likelihood* of model M_k, obtained by integrating (not maximizing) over the unknown parameters:

$$\begin{aligned} p(D|M_k) &= \int p(D|\theta_k, M_k)p(\theta_k|M_k)d\theta_k \quad (3) \\ &= \int (\text{ likelihood } \times \text{ prior })d\theta_k, \end{aligned}$$

where θ_k is the parameter of model M_k and $p(D|\theta_k, M_k)$ is the likelihood of θ_k under model M_k. The prior model probabilities are often taken to be equal, so they cancel in (2), but the `BMA` package allows other formulations also.

The integrated likelihood $p(D|M_k)$ is a high dimensional integral that can be hard to calculate analytically, and several of the functions in the `BMA` package use the simple and surprisingly accurate BIC approximation

$$2\log p(D|M_k) \approx 2\log p(D|\hat{\theta}_k) - d_k \log(n) = -\text{BIC}_k, \quad (4)$$

where $d_k = \dim(\theta_k)$ is the number of independent parameters in M_k, and $\hat{\theta}_k$ is the maximum likelihood estimator (equal to the ordinary least squares estimator for linear regression coefficients). For linear regression, BIC has the simple form

$$\text{BIC}_k = n\log(1 - R_k^2) + p_k \log n \quad (5)$$

up to an additive constant, where R_k^2 is the value of R^2 and p_k is the number of regressors for the kth regression model. By (5), $\text{BIC}_k = 0$ for the null model with no regressors.

When interest focuses on a model parameter, say a regression parameter such as β_1, (1) can be applied with $\Delta = \beta_1$. The BMA posterior mean of β_1 is just a weighted average of the posterior means of β_1 under each of the models:

$$E[\beta_1|D] = \sum \tilde{\beta}_1^{(k)} p(M_k|D), \quad (6)$$

which can be viewed as a model-averaged Bayesian point estimator. In (6), $\tilde{\beta}_1^{(k)}$ is the posterior mean of β_1 under model M_k, and this can be approximated by the corresponding maximum likelihood estimator, $\hat{\beta}_1^{(k)}$ (Raftery, 1995). A similar expression is available for the BMA posterior standard deviation, which can be viewed as a model-averaged Bayesian standard error. For a survey and literature review of Bayesian model averaging, see Hoeting et al. (1999).

Computation

Implementing BMA involves two hard computational problems: computing the integrated likelihoods for all the models (3), and averaging over all the models, whose number can be huge, as in (1) and (6). In the functions `bicreg` (for variable selection in linear regression), `bic.glm` (for variable selection in generalized linear models), and `bic.surv` (for variable selection in survival or event history analysis), the integrated likelihood is approximated by the BIC approximation (4). The sum over all the models is approximated by finding the best models using the fast leaps and bounds algorithm. The leaps and bounds algorithm was introduced for all subsets regression by Furnival and Wilson (1974), and extended to BMA for linear regression and generalized linear models by Raftery (1995), and to BMA for survival analysis by Volinsky et al. (1997). Finally, models that are much less likely *a posteriori* than the best model are excluded. This is an exhaustive search and finds the globally optimal model.

If the number of variables is large, however, the leap and bounds algorithm can be substantially slowed down; we have found that it can slow down substantially once the number of variables goes beyond about 40–45. One way around this is via specification of the argument `maxCol`. If the number of variables is greater than `maxCol` (whose current default value is 31), the number of variables is reduced to `maxCol` by backwards stepwise elimination before applying the leaps and bounds algorithm.

In the function `glib` (model selection for generalized linear models with prior information), the integrated likelihood is approximated by the Laplace method (Raftery, 1996). An example of the use of `glib` to analyze epidemiological case-control studies is given by Viallefont et al. (2001). In the function `MC3.REG` (variable selection and outlier detection for linear regression), conjugate priors and exact integrated likelihoods are used, and the sum is approximated using Markov chain Monte Carlo (Hoeting et al., 1996; Raftery et al., 1997).

Example 1: Linear Regression

To illustrate how BMA takes account of model uncertainty about the variables to be included in linear regression, we use the `UScrime` dataset on crime rates in 47 U.S. states in 1960 (Ehrlich, 1973); this is available in the MASS library. There are 15 potential independent variables, all associated with crime rates in the literature. The last two, probability of imprisonment and average time spent in state prisons, were the predictor variables of interest in the original study, while the other 13 were control variables. All variables for which it makes sense were logarithmically transformed. The commands are:

```
library(BMA)
library(MASS)
data(UScrime)
x.crime<- UScrime[,-16]
y.crime<- log(UScrime[,16])
x.crime[,-2]<- log(x.crime[,-2])
crime.bicreg <- bicreg(x.crime, y.crime)
summary (crime.bicreg, digits=2)
```

This summary yields Figure 1. (The default number of digits is 4, but this may be more than needed.) The column headed "p!=0" shows the posterior probability that the variable is in the model (in %). The column headed "EV" shows the BMA posterior mean, and the column headed "SD" shows the BMA posterior standard deviation for each variable. The following five columns show the parameter estimates for the best five models found, together with the numbers of variables they include, their R^2 values, their BIC values and their posterior model probabilities.

The `plot` command shows the BMA posterior distribution of each of the regression parameters, given by (1) with $\Delta = \beta_k$. For example, the BMA posterior distribution of the coefficient of the average time in prison variable is shown in Figure 2. The spike at 0 shows the probability that the variable is not in the model, in this case 0.555. The curve shows the posterior density given that the variable is in the model, which is approximated by a finite mixture of normal densities and scaled so that the height of the density curve is equal to the posterior probability that the variable is in the model. The BMA posterior distributions of all the parameters are produced by the command

```
> plot (crime.bicreg,mfrow=c(4,4))
```

and shown in Figure 3.

A visual summary of the BMA output is produced by the `imageplot.bma` command, as shown in Figure 4. Each row corresponds to a variable, and each column corresponds to a model; the corresponding rectangle is red if the variable is in the model and white otherwise. The width of the column is proportional to the model's posterior probability. The basic idea of the image plot was proposed by Clyde (1999); in her version all the columns had the same width, while in the `imageplot.bma` output they depend on the models' posterior probabilities.

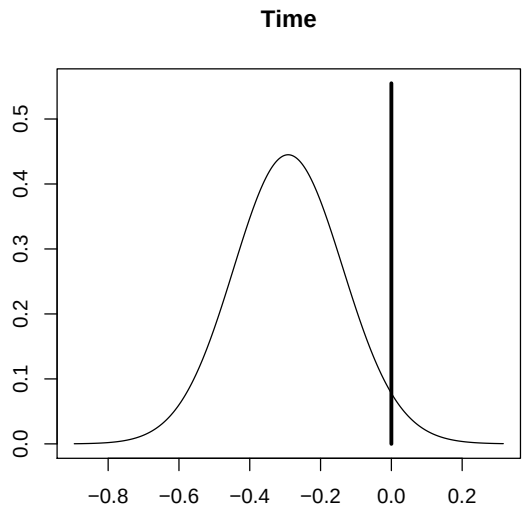

Figure 2: BMA posterior distribution of the coefficient of the average time in prison variable in the crime dataset. The spike at 0 shows the posterior probability that the variable is not in the model. The curve is the model-averaged posterior density of the coefficient given that the variable is in the model, approximated by a finite mixture of normal distributions, one for each model that includes the variable. The density is scaled so that its maximum height is equal to the probability of the variable being in the model.

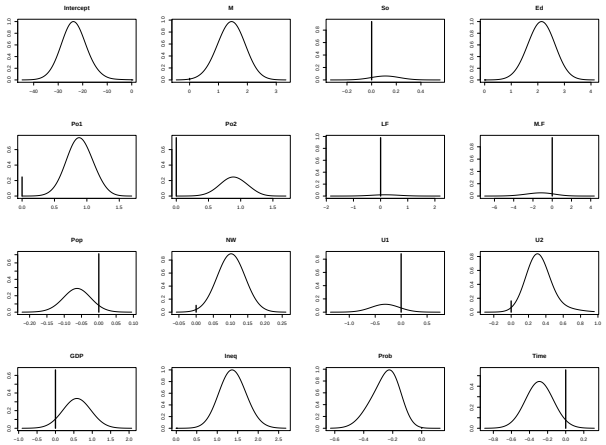

Figure 3: BMA posterior distribution of all the coefficients for the crime dataset, produced by the command `plot(crime.bicreg,mfrow=c(4,4))`

It is clear that M (percent male), Education, NW (percent nonwhite), Inequality and Prob (probability of imprisonment) have high posterior probabilities of being in the model, while most other variables, such as So (whether the state is in the South), have low posterior probabilities. The Time variable (average time spent in state prisons) does appear in the best

```
Call: bicreg(x = x.crime, y = y.crime)
  51  models were selected
 Best  5  models (cumulative posterior probability =  0.29 ):

           p!=0    EV       SD     model 1  model 2  model 3  model 4  model 5
Intercept  100.0  -23.4778  5.463  -22.637  -24.384  -25.946  -22.806  -24.505
M           97.5    1.4017  0.531    1.478    1.514    1.605    1.268    1.461
So           6.3    0.0069  0.039      .        .        .        .        .
Ed         100.0    2.1282  0.513    2.221    2.389    2.000    2.178    2.399
Po1         75.4    0.6707  0.423    0.852    0.910    0.736    0.986      .
Po2         24.6    0.2186  0.396      .        .        .        .      0.907
LF           2.1    0.0037  0.080      .        .        .        .        .
M.F          5.2   -0.0691  0.446      .        .        .        .        .
Pop         28.9   -0.0183  0.036      .        .        .     -0.057      .
NW          89.6    0.0911  0.050    0.109    0.085    0.112    0.097    0.085
U1          11.9   -0.0364  0.136      .        .        .        .        .
U2          83.9    0.2740  0.191    0.289    0.322    0.274    0.281    0.330
GDP         33.8    0.1971  0.354      .        .      0.541      .        .
Ineq       100.0    1.3810  0.333    1.238    1.231    1.419    1.322    1.294
Prob        98.8   -0.2483  0.100   -0.310   -0.191   -0.300   -0.216   -0.206
Time        44.5   -0.1289  0.178   -0.287      .     -0.297      .        .

nVar                                    8        7        9        8        7
r2                                    0.842    0.826    0.851    0.838    0.823
BIC                                 -55.912  -55.365  -54.692  -54.604  -54.408
post prob                             0.089    0.067    0.048    0.046    0.042
```

Figure 1: Summary of the output of `bicreg` for the crime data

model, but there is nevertheless a great deal of uncertainty about whether it should be included. The two variables Po1 (police spending in 1959) and Po2 (police spending in 1960) are highly correlated, and the models favored by BMA include one or the other, but not both.

Example 2: Logistic Regression

We illustrate BMA for logistic regression using the low birthweight data set of Hosmer and Lemeshow (1989), available in the MASS library. The dataset consists of 189 babies, and the dependent variable measures whether their weight was low at birth. There are 8 potential independent variables of which two are categorical with more than two categories: race and the number of previous premature labors (`ptl`). Figure 5 shows the output of the commands

```
library(MASS)
data(birthwt)
birthwt$race <- as.factor (birthwt$race)
birthwt$ptl <- as.factor (birthwt$ptl)
bwt.bic.glm <- bic.glm (low ~ age + lwt
 + race + smoke + ptl + ht + ui + ftv,
 data=birthwt, glm.family="binomial")
summary (bwt.bic.glm,conditional=T,digits=2)
```

The function `bic.glm` can accept a formula, as here, instead of a design matrix and dependent variable (the same is true of `bic.surv` but not of `bicreg`). By default, the levels of a categorical factor such as race are constrained to be either all in the model or all out of it. However, by specifying `factor.type=F` in `bic.glm`, one can allow the individual dummy variables specifying a factor to be in or out of the model.

The posterior distributions of the model parameters are shown in Figure 6, and the image plot is shown in Figure 7. Note that each dummy variable making up the race factor has its own plot in Figure 6, but the race factor has only one row in Figure 7.

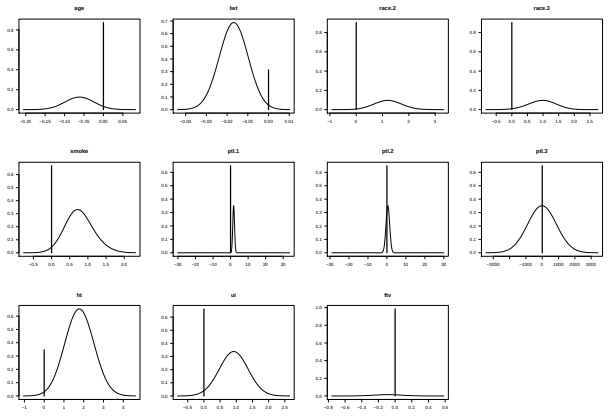

Figure 6: BMA posterior distributions of the parameters of the logistic regression model for the low birthweight data. Note that there is a separate plot for each of the dummy variables for the two factors (race and `ptl`)

Figure 4: Image plot for the crime data produced by the command `imageplot.bma(crime.bicreg)`

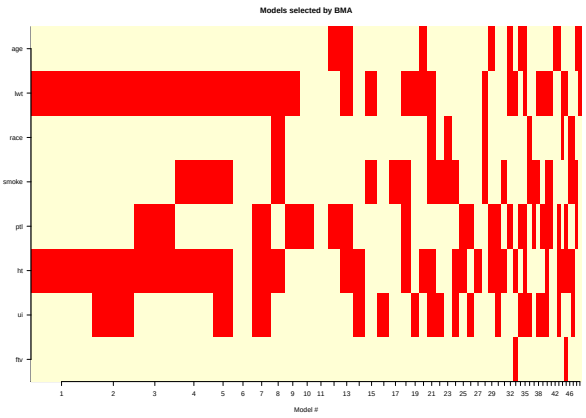

Figure 7: Image plot summarizing the BMA analysis for logistic regression of the low birthweight data

Example 3: Survival Analysis

BMA can account for model uncertainty in survival analysis using the Cox proportional hazards model. We illustrate this using the well-known Primary Biliary Cirrhosis (PBC) data set of Fleming and Harrington (1991). This consists of 312 randomized patients with the disease, and the dependent variable was survival time; 187 of the records were censored. There were 15 potential predictors. The data set is available in the `survival` library, which is loaded when `BMA` is loaded.

The analysis proceeded as follows:

```
data(pbc)
x.pbc<- pbc[1:312,]
surv.t<- x.pbc$time
cens<- x.pbc$status
x.pbc<- x.pbc[,-c(6,7,10,17,19)]
x.pbc$bili<- log(x.pbc$bili)
x.pbc$alb<- log(x.pbc$alb)
x.pbc$protime<- log(x.pbc$protime)
x.pbc$copper<- log(x.pbc$copper)
x.pbc$sgot<- log(x.pbc$sgot)
pbc.bic.surv <- bic.surv(x.pbc,surv.t,cens)
summary(pbc.bic.surv,digits=2)
plot(pbc.bic.surv,mfrow=c(4,4))
imageplot.bma(pbc.bic.surv)
```

The summary results, BMA posterior distributions, and image plot visualization of the results are shown in Figures 8, 9 and 10.

```
Call: bic.glm.formula(f = low ~ age+lwt+race+smoke+ptl+ht+ui+ftv, data = birthwt, glm.family = "binomial")
  49  models were selected
 Best  5  models (cumulative posterior probability =  0.37 ):

           p!=0    EV       SD       cond EV   cond SD   model 1   model 2   model 3   model 4   model 5
Intercept  100     0.4716   1.3e+00    0.472     1.309     1.451     1.068     1.207     1.084     0.722
age         12.6  -0.0078   2.4e-02   -0.062     0.037       .         .         .         .         .
lwt         68.7  -0.0116   9.7e-03   -0.017     0.007    -0.019    -0.017    -0.019    -0.018    -0.016
race         9.6
    .2             0.1153   3.9e-01    1.201     0.547       .         .         .         .         .
    .3             0.0927   3.2e-01    0.966     0.461       .         .         .         .         .
smoke       33.2    0.2554   4.3e-01    0.768     0.397       .         .         .         0.684     0.653
ptl         35.1
   .1              0.6174   8.9e-01    1.758     8.211       .         .         1.743     .         .
   .2              0.1686   6.1e-01    0.480     7.592       .         .         0.501     .         .
   .3             -4.9110   5.2e+02  -13.988   882.840       .         .       -13.986     .         .
ht          65.4    1.1669   1.0e+00    1.785     0.729     1.856     1.962     1.924     1.822     1.922
ui          33.8    0.3105   5.1e-01    0.918     0.445       .         0.930     .         .         0.896
ftv          1.5   -0.0013   2.4e-02   -0.087     0.173       .         .         .         .         .

nVar                                                           2         3         3         3         4
BIC                                                         -753.823  -753.110  -752.998  -752.865  -751.656
post prob                                                      0.111     0.078     0.073     0.069     0.037
```

Figure 5: Summary of the `bic.glm` output for the low birthweight data. For factors, either all levels are in or all levels are out of the model. With the option `conditional=T`, the posterior means and standard deviations of the parameters conditionally on the variable being in the model are shown in addition to the unconditional ones.

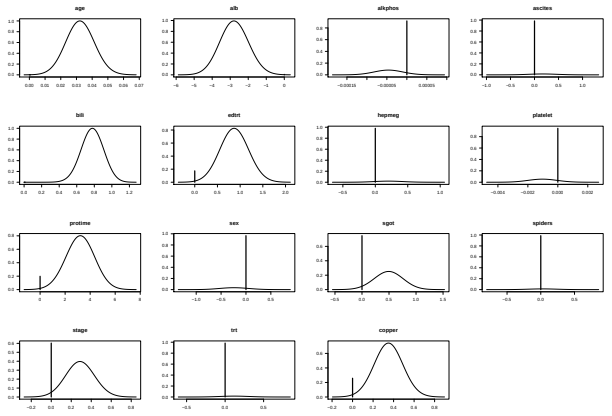

Figure 9: BMA posterior distributions of the parameters of the Cox proportional hazards model model for the PBC data

Summary

The `BMA` package carries out Bayesian model averaging for linear regression, generalized linear models, and survival or event history analysis using Cox proportional hazards models. The library contains functions for plotting the BMA posterior distributions of the model parameters, as well as an image plot function that provides a way of visualizing the BMA output. The functions `bicreg`, `bic.glm` and `bic.surv` provide fast and automatic default ways of doing this for the model classes considered. Prior information can be incorporated explicitly using the `glib` and `MC3.REG` functions, and `MC3.REG` also takes account of uncertainty about outlier removal.

Figure 10: Image plot summarizing the BMA analysis for Cox proportional hazards modeling of the PBC data

Acknowledgement

The `MC3.REG` function is based on an earlier Splus function written by Jennifer A. Hoeting.

```
Call: bic.surv.data.frame(x = x, surv.t = surv.t, cens = cens)
  39  models were selected
 Best  5  models (cumulative posterior probability =  0.37 ):

           p!=0    EV       SD       model 1   model 2   model 3   model 4   model 5
age        100.0   3.2e-02  9.2e-03    0.032     0.029     0.031     0.031     0.036
alb         99.2  -2.7e+00  8.3e-01   -2.816    -2.446    -2.386    -3.268    -2.780
alkphos      8.1  -3.7e-06  1.6e-05      .         .         .         .         .
ascites      1.6   2.6e-03  4.3e-02      .         .         .         .         .
bili       100.0   7.8e-01  1.3e-01    0.761     0.743     0.786     0.801     0.684
edtrt       82.6   7.2e-01  4.4e-01    0.813     0.821     1.026     .         0.837
hepmeg       2.0   4.1e-03  4.3e-02      .         .         .         .         .
platelet     5.4  -5.5e-05  3.2e-04      .         .         .         .         .
protime     80.2   2.6e+00  1.6e+00    3.107     2.706     .         3.770     3.417
sex          3.4  -8.0e-03  6.8e-02      .         .         .         .         .
sgot        25.2   1.2e-01  2.5e-01      .         .         .         .         0.407
spiders      1.3   3.7e-04  2.5e-02      .         .         .         .         .
stage       39.7   1.1e-01  1.7e-01      .         0.244     0.318     .         .
trt          1.6   1.9e-03  2.8e-02      .         .         .         .         .
copper      74.2   2.6e-01  2.0e-01    0.358     0.347     0.348     0.357     0.311

nVar                                      6         7         6         5         7
BIC                                    -177.952  -176.442  -176.212  -175.850  -175.537
post prob                                 0.147     0.069     0.062     0.051     0.044
```

Figure 8: Summary of the `bic.surv` output for the PBC data.

Bibliography

M. A. Clyde. Bayesian model averaging and model search strategies (with Discussion). In *Bayesian Statistics 6* (edited by J. M. Bernardo et al.), pages 157–185. Oxford University Press, Oxford, U.K., 1999. 4

I. Ehrlich. Participation in illegitimate activities: a theoretical and empirical investigation. *Journal of Political Economy*, 81:521–565, 1973. 3

T. R. Fleming and D. P. Harrington. *Counting processes and survival analysis*. Wiley, New York, 1991. 6

G. M. Furnival and R. W. Wilson. Regression by leaps and bounds. *Technometrics*, 16:499–511, 1974. 3

J. A. Hoeting, A. E. Raftery, and D. Madigan. A method for simultaneous variable selection and outlier identification in linear regression. *Computational Statistics and Data Analysis*, 22:251–270, 1996. 3

J. A. Hoeting, D. Madigan, A. E. Raftery, and C. T. Volinsky. Bayesian model averaging: A tutorial (with discussion). *Statistical Science*, 14:382–417, 1999. 3

D. W. Hosmer and S. Lemeshow. *Applied Logistic Regression*. Wiley, New York, 1989. 5

A. E. Raftery. Bayesian model selection in social research (with Discussion). In *Sociological Methodology 1995* (edited by P. V. Marsden), pages 111–163. Cambridge, Mass. : Blackwell Publishers, 1995. 3

A. E. Raftery. Approximate Bayes factors and accounting for model uncertainty in generalized linear models. *Biometrika*, 83:251–266, 1996. 3

A. E. Raftery, D. Madigan, and J. A. Hoeting. Model selection and accounting for model uncertainty in linear regression models. *Journal of the American Statistical Association*, 92:179–191, 1997. 3

V. Viallefont, A. E. Raftery, and S. Richardson. Variable selection and Bayesian model averaging in case-control studies. *Statistics in Medicine*, 20:3215–3230, 2001. 3

C. T. Volinsky, D. Madigan, A. E. Raftery, and R. A. Kronmal. Bayesian model averaging in proportional hazards models: Assessing the risk of a stroke. *Applied Statistics*, 46:433–448, 1997. 3

Adrian Raftery
University of Washington, Seattle
raftery@stat.washington.edu

Ian Painter
Ian.Painter@gmail.com

Chris Volinsky
Volinsky@research.att.com