Linear Regression

The classical linear model is

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where $\boldsymbol{Y}$ is $n \times 1$, $\boldsymbol{X}$ is a *fixed* $n \times p$ matrix, $\boldsymbol{\beta}$ is $p \times 1$ and

$$\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 \boldsymbol{I}).$$

We observe only $\boldsymbol{Y}$ and $\boldsymbol{X}$, and the unknown parameters are the *regression coefficients* $\boldsymbol{\beta}$ and the error variance σ^2.

We assume that $\boldsymbol{X}$ is of full column rank p.

Let's review frequentist estimation of $\boldsymbol{\beta}$ and σ^2.

The maximum likelihood estimate (MLE) of $\boldsymbol{\beta}$ is the *least squares estimate*, which is

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}.$$

Of course, $\widehat{\boldsymbol{\beta}}$ is the solution to the problem

$$\min_{\boldsymbol{b}} (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{b})^T(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{b}).$$

The MLE of σ^2 is

$$\begin{aligned} \widehat{\sigma}^2 &= n^{-1}(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})^T(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}) \\ &= n^{-1}\boldsymbol{Y}^T\left[\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\right]\boldsymbol{Y}. \\ &= n^{-1}\boldsymbol{Y}^T(\boldsymbol{I} - \boldsymbol{M})\boldsymbol{Y}. \end{aligned}$$

The Jeffreys noninformative prior for $\boldsymbol{\beta}$ is constant over all p-vectors, and hence improper.

A common noninformative prior for σ is $p(\sigma) = \sigma^{-1}$, which is also improper.

Assuming that $\boldsymbol{\beta}$ and σ are a priori independent, and using these noninformative priors, let's derive the posterior.

We have

$$p(\boldsymbol{\beta},\sigma|\boldsymbol{y}) \propto \frac{1}{\sigma^{n+1}} \exp\left[-\frac{(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})}{2\sigma^2}\right].$$

Simple algebra shows that

$$(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}) =$$

$$\boldsymbol{y}^T(\boldsymbol{I}-\boldsymbol{M})\boldsymbol{y} + (\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}})^T\boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}).$$

The posterior is thus proportional to

$$\sigma^{-(n+1)} \exp\left(-\frac{\boldsymbol{y}^T(\boldsymbol{I}-\boldsymbol{M})\boldsymbol{y}}{2\sigma^2}\right)$$

$$\times \exp\left[-\frac{(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}})^T\boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}})}{2\sigma^2}\right].$$

It now follows that the *conditional of $\boldsymbol{\beta}$ given σ and $\boldsymbol{y}$ is $N(\widehat{\boldsymbol{\beta}}, \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1})$.*

The posterior is proportional to

$$\sigma^{-(n+1)}\sigma^{p}\exp\left(-\frac{\boldsymbol{y}^T(\boldsymbol{I}-\boldsymbol{M})\boldsymbol{y}}{2\sigma^2}\right)$$

$$\times\frac{1}{\sigma^p}\exp\left[-\frac{(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}})^T\boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}})}{2\sigma^2}\right].$$

Now using the fact that

$$p(\boldsymbol{\beta},\sigma|\boldsymbol{y}) = p(\sigma|\boldsymbol{y})p(\boldsymbol{\beta}|\sigma,\boldsymbol{y}),$$

it follows that σ^2 given $\boldsymbol{y}$ is inverse-gamma with parameters $(n-p)/2$ and $\boldsymbol{y}^T(\boldsymbol{I}-\boldsymbol{M})\boldsymbol{y}/2$.

The marginal posterior of $\boldsymbol{\beta}$ is proportional to

$$\left[\boldsymbol{y}^T(\boldsymbol{I}-\boldsymbol{M})\boldsymbol{y}+(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}})^T\boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}})\right]^{-n/2} =$$

$$\left[(n-p)s^2+(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}})^T\boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}})\right]^{-n/2} \propto$$

$$\left[1+\frac{1}{(n-p)}(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}})^T\left(\frac{\boldsymbol{X}^T\boldsymbol{X}}{s^2}\right)(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}})\right]^{-n/2}.$$

The p-variate t-distribution with parameters ν, $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ has density proportional to

$$|\boldsymbol{\Sigma}|^{-1/2}\left[1+\frac{1}{\nu}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right]^{-(\nu+p)/2}.$$

It follows that $\boldsymbol{\beta}|\boldsymbol{y}$ has a multivariate t-distribution with parameters $n-p$, $\widehat{\boldsymbol{\beta}}$ and $s^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}$.

We found the full conditional for $\boldsymbol{\beta}$ on p. 214N.

From the posterior on p. 215 we see that the full conditional of $1/\sigma^2$ is gamma with parameters $n/2$ and

$$\frac{1}{2}\left[(n-p)s^2+(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}})^T\boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}})\right].$$

Example 16 Maximal oxygen uptake

The data in this example were simulated from a model reported in a 1994 article in *Research Quarterly for Exercise and Sport*.

The study in that article involved 50 male runners. A kinesiologist wanted to know if maximal oxygen uptake could be predicted from easily measured explanatory variables.

The variables in the study were as follows:

$y =$ maximal oxygen uptake (in liters per minute)

$x_1 =$ weight (in kilograms)

$x_2 =$ age (in years)

$x_3 =$ time to walk 1 mile (in minutes)

$x_4 =$ heart rate at end of walk (in beats per minute)

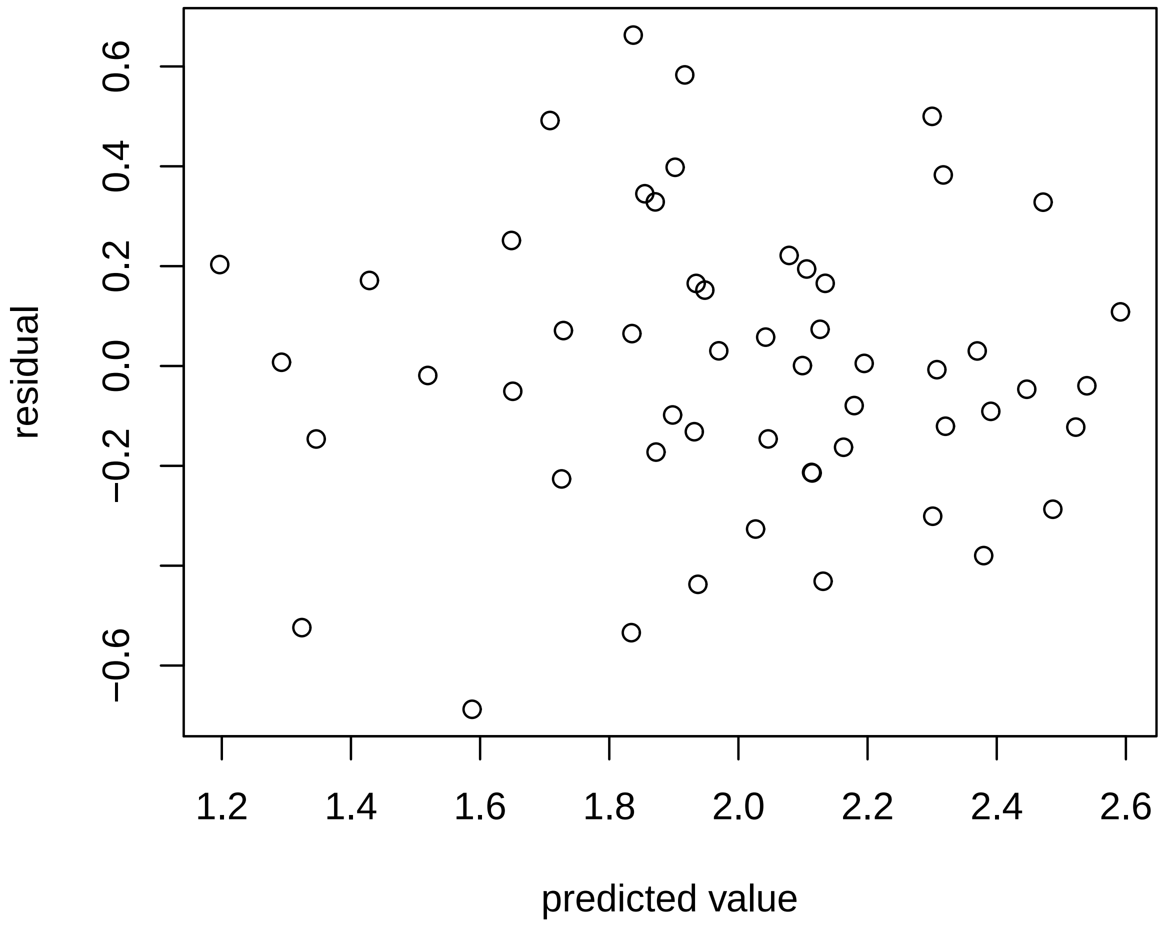
Least squares estimates of β_js:

$$\widehat{\beta}_0 = 5.588 \quad \widehat{\beta}_1 = 0.0129 \quad \widehat{\beta}_2 = -0.0830$$

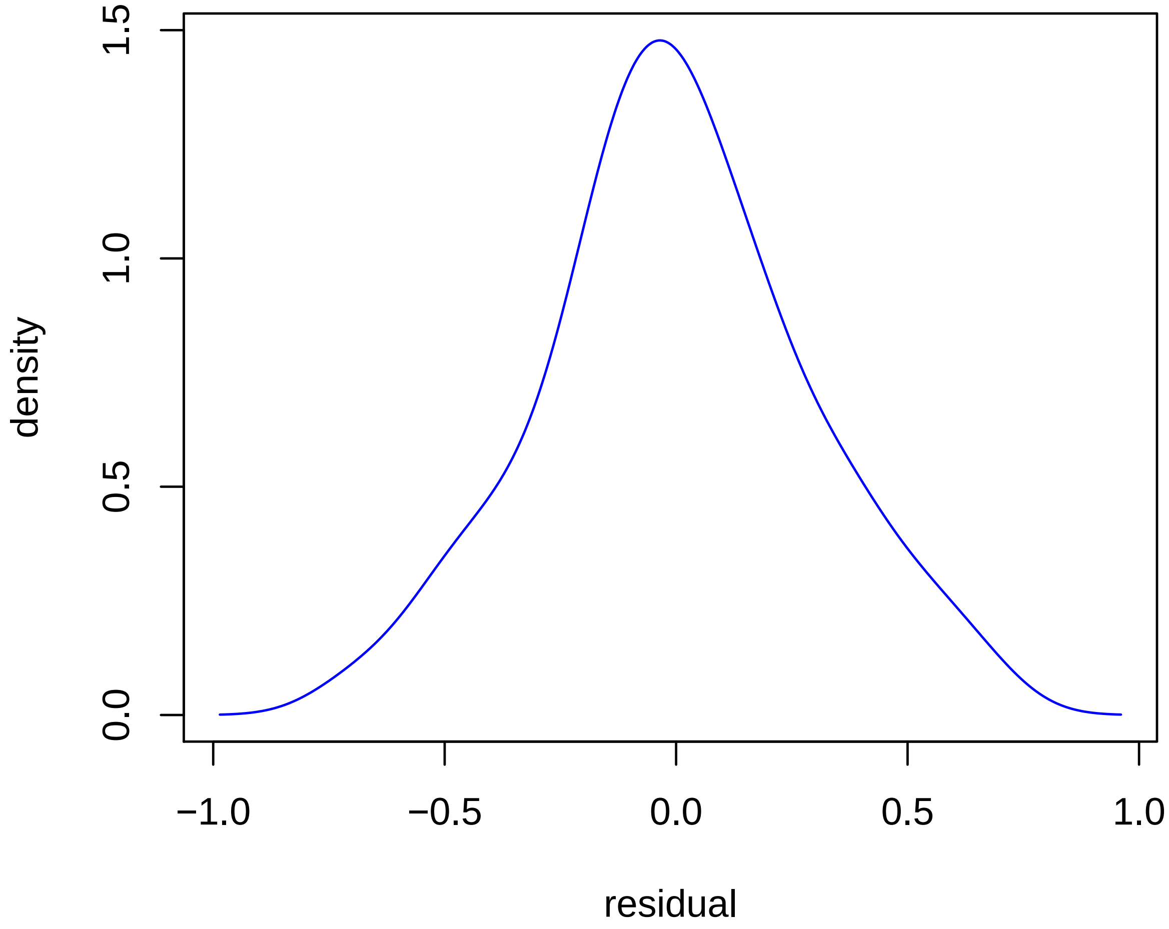
$$\widehat{\beta}_3 = -0.158 \quad \widehat{\beta}_4 = -0.00911.$$

The MLE of σ^2 is $\widehat{\sigma}^2 = 0.0814$.

Residual plot



Kernel estimate of density of ϵ_j



The estimate is computed from the residuals

$$\widehat{\epsilon}_i = y_i - \left[\widehat{\beta}_0 + \sum_{j=1}^{4} \widehat{\beta}_j x_{ij}\right], \qquad i = 1, \ldots, n.$$

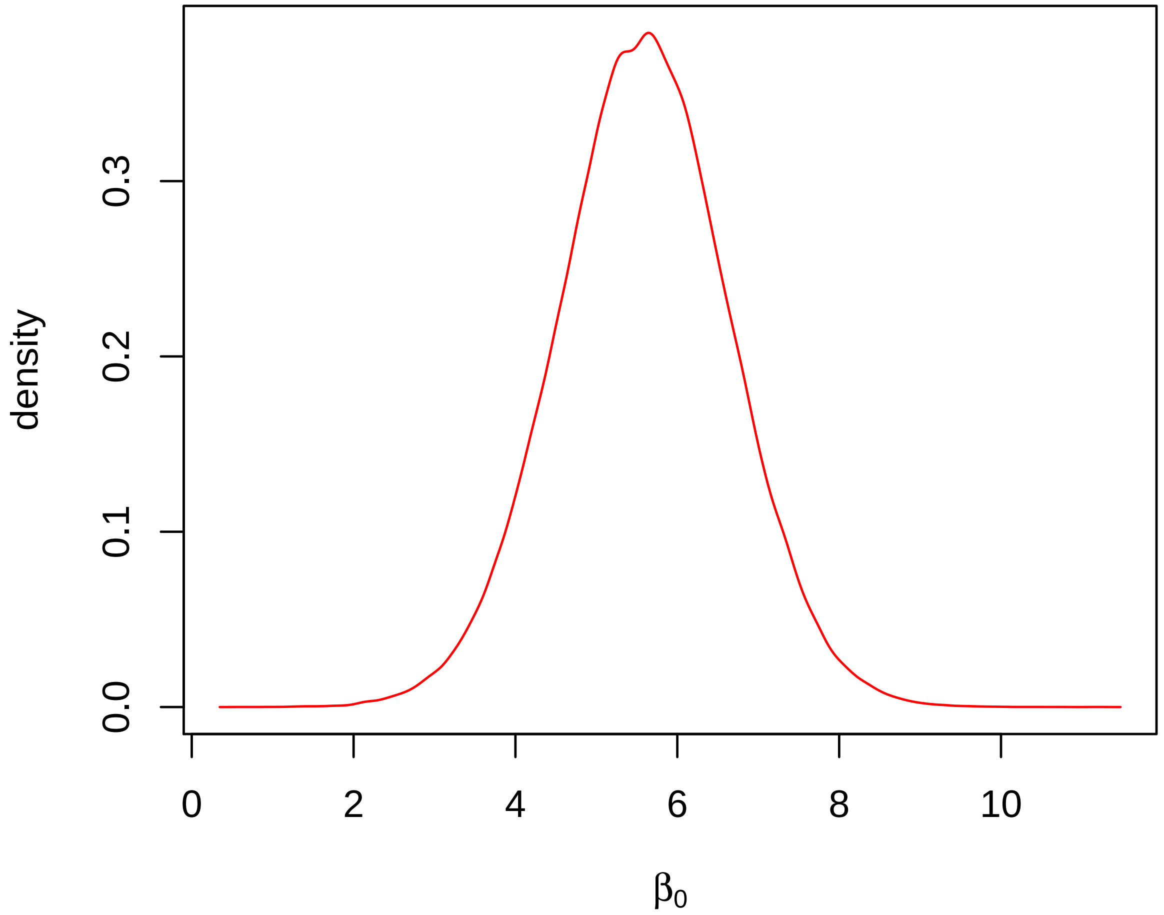
Based on the residual plot and the kernel estimate, the assumption of i.i.d. normal error terms seems reasonable.

- I used Gibbs sampling to generate 100,000 observations from the posterior based on the noninformative prior on p. 213N.

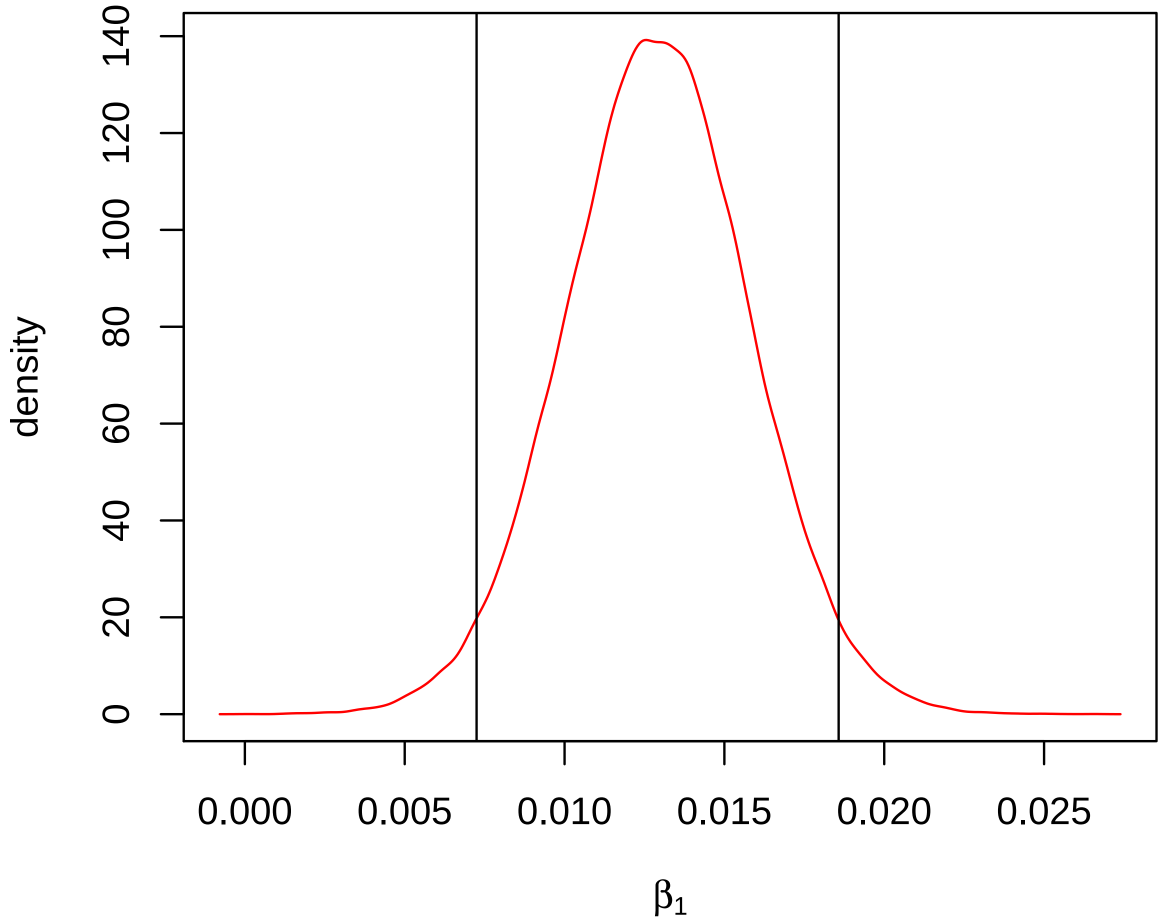
- Sample autocorrelation functions for the output showed essentially no serial correlation among the generated $\boldsymbol{\beta}$s.

- Minimal correlation was observed among the generated values of σ.

Kernel estimate of posterior of β_0

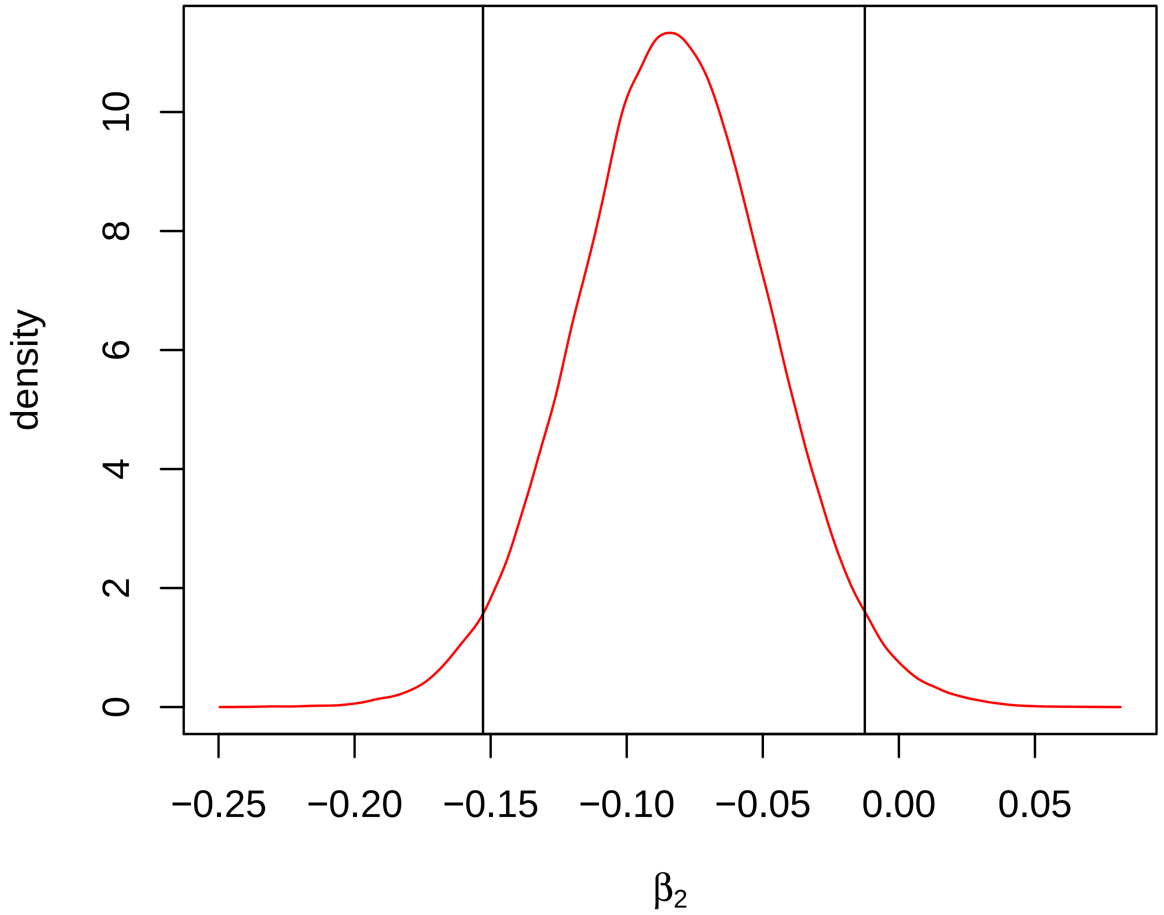


Kernel estimate of posterior of β_1

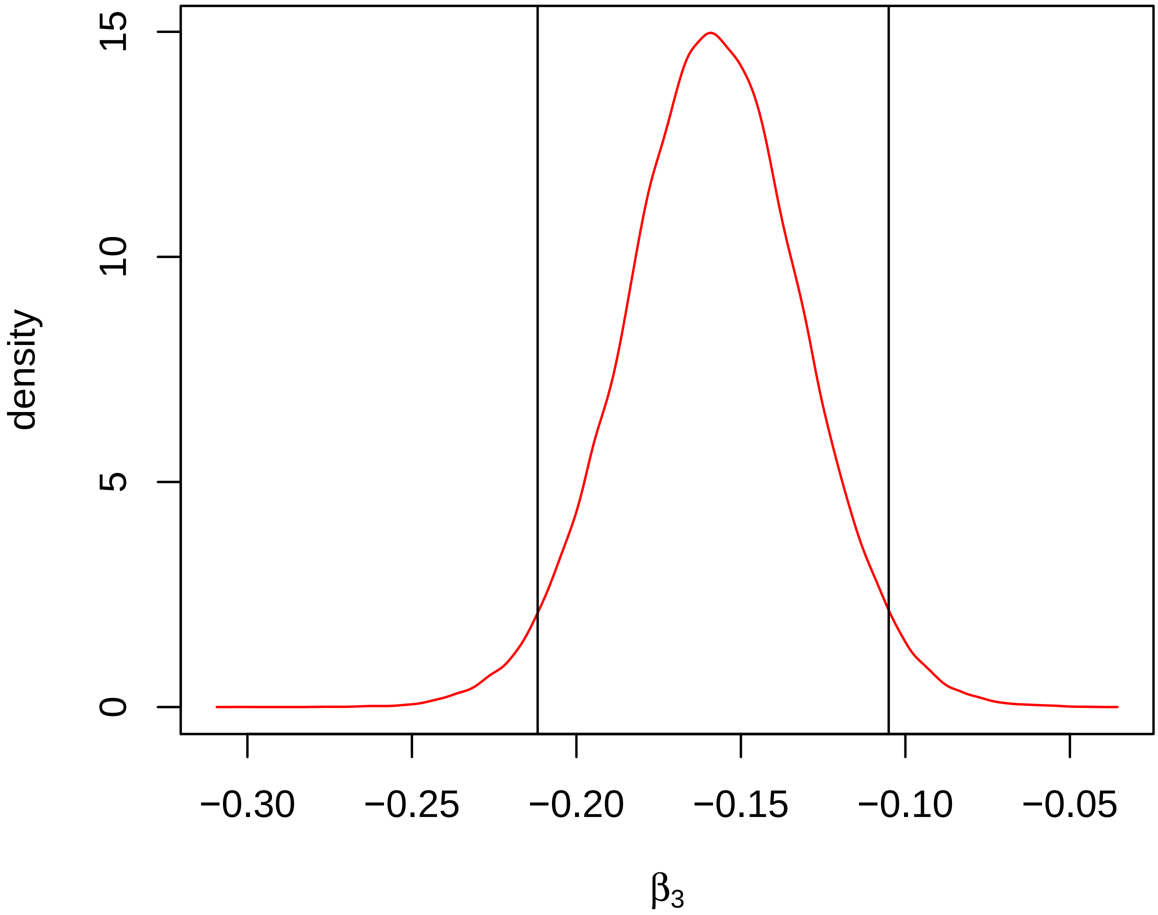


HPD region indicated by black lines.

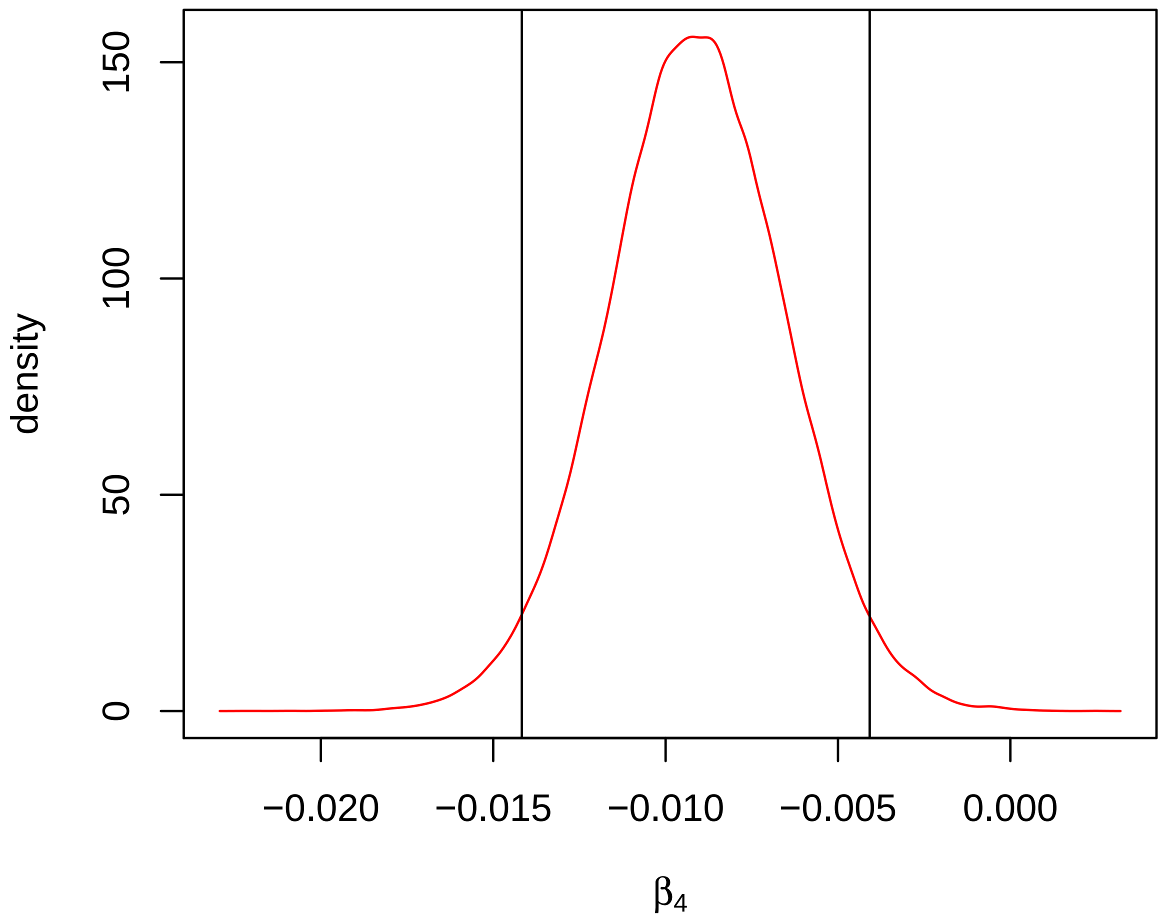
Kernel estimate of posterior of β_2



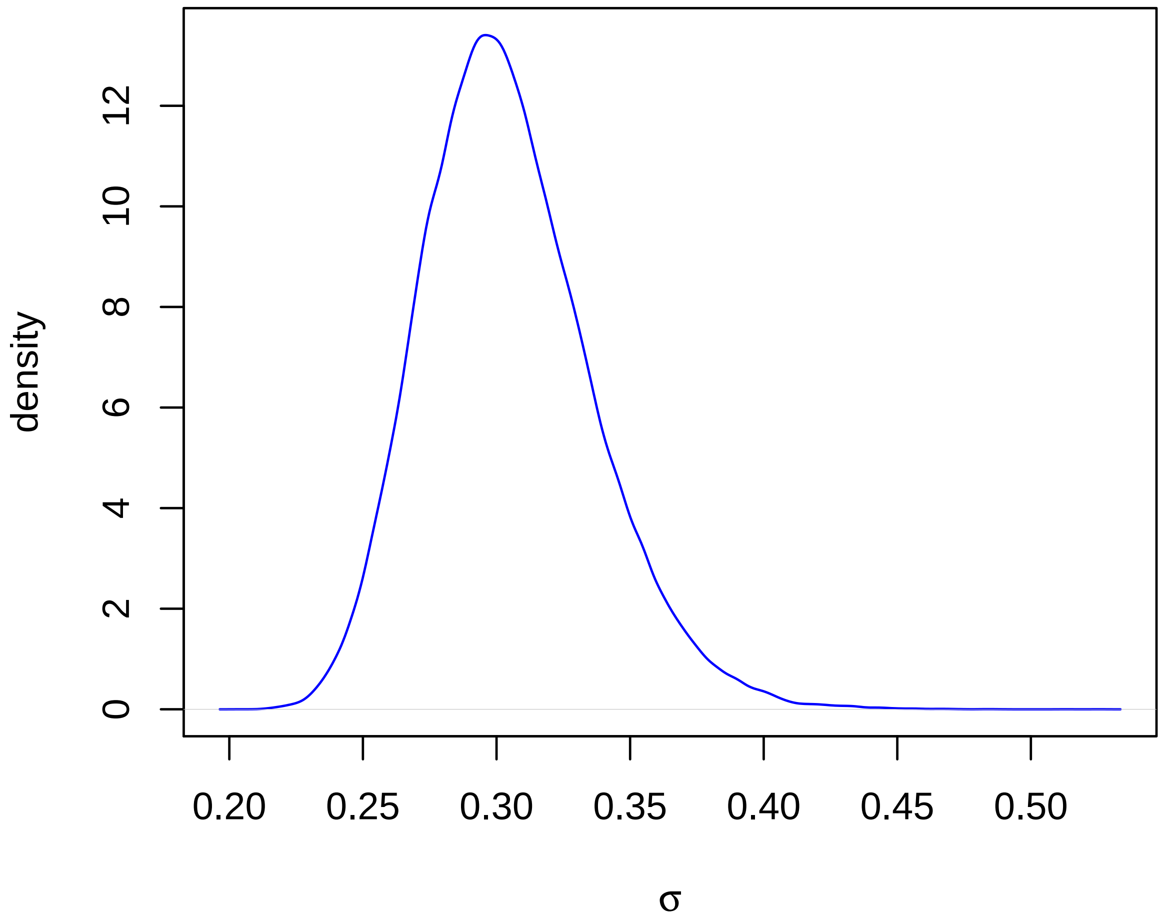
Kernel estimate of posterior of β_3



Kernel estimate of posterior of β_4



Kernel estimate of posterior of σ



Correlation matrix of β from the Gibbs output

$$\begin{matrix}
1.00 & -0.42 & -0.71 & -0.45 & -0.38 \\
-0.42 & 1.00 & 0.07 & -0.01 & -0.12 \\
-0.71 & 0.07 & 1.00 & -0.06 & -0.01 \\
-0.45 & -0.01 & -0.06 & 1.00 & 0.26 \\
-0.38 & -0.12 & -0.01 & 0.26 & 1.00
\end{matrix}$$

Note that there is little correlation among β_1, β_2, β_3 and β_4.

This is good because inference about any one regression coefficient is not affected by that for other coefficients.

The lack of correlation also shows that we don't need to worry about collinearity of predictors.

A large positive correlation between two predictors translates into a large negative correlation between the corresponding βs.

Selecting a linear regression model

To this point we have not addressed the question of model selection. *We have always assumed that we know the correct parametric model.*

Linear regression is a natural place to address model selection.

Often investigators throw in “all variables but the kitchen sink,” and then go through a process of finding a best subset of variables.

Bayesian methods provide a means of estimating a best subset, i.e., a best model.

Let's first consider the general problem of model selection, and then apply what we learn to linear regression.

Let $M_1, M_2, \ldots, M_N$ be the models under consideration. These might be of different dimensions, i.e., they might have different numbers of parameters.

- $\boldsymbol{\theta}_k$: vector of parameters for model M_k

- $p(\boldsymbol{\theta}_k|k)$: prior for $\boldsymbol{\theta}_k$ assuming that M_k is the correct model

- p_k: prior probability of M_k

- $p(\boldsymbol{y}|\boldsymbol{\theta}_k, k)$: likelihood of data assuming true model is M_k and parameters are $\boldsymbol{\theta}_k$

We have

$$p(\boldsymbol{\theta}_k, k|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\boldsymbol{\theta}_k, k)p(\boldsymbol{\theta}_k|k)p_k}{m(\boldsymbol{y})},$$

where

$$m(\boldsymbol{y}) = \sum_{k=1}^{N} p_k m_k(\boldsymbol{y}),$$

and

$$m_k(\boldsymbol{y}) = \int_{\Theta_k} p(\boldsymbol{y}|\boldsymbol{\theta}_k, k)p(\boldsymbol{\theta}_k|k)\, d\boldsymbol{\theta}_k.$$

Why? *Because of the law of total probabilities.*

$$\begin{aligned} m(\boldsymbol{y}) &= \sum_{k=1}^{N} m(\boldsymbol{y}|M_k)p_k \\ &= \sum_{k=1}^{N} \int_{\Theta_k} p(\boldsymbol{y}, \boldsymbol{\theta}_k|k)\, d\boldsymbol{\theta}_k \cdot p_k \\ &= \sum_{k=1}^{N} \int_{\Theta_k} p(\boldsymbol{y}|\boldsymbol{\theta}_k, k)p(\boldsymbol{\theta}_k|k)\, d\boldsymbol{\theta}_k \cdot p_k \\ &= \sum_{k=1}^{N} p_k m_k(\boldsymbol{y}). \end{aligned}$$

An apparently reasonable way to select a model is to choose one that maximizes posterior probability:

$$\begin{aligned} P(M_k|\boldsymbol{y}) &= \int_{\Theta_k} p(\boldsymbol{\theta}_k, k|\boldsymbol{y})\, d\boldsymbol{\theta}_k \\ &= \frac{m_k(\boldsymbol{y})p_k}{m(\boldsymbol{y})}. \end{aligned}$$

A problem that arises is that when using $P(M_k|\boldsymbol{y})$ as a criterion for model selection, *one should* **not** *use improper priors for any of* $p(\boldsymbol{\theta}_1|1), \ldots,$ $p(\boldsymbol{\theta}_N|N)$.

Why?

When considering a single model, the arbitrary constant in front of the improper prior cancels out when computing the posterior.

However, these constants *do not* cancel out when computing $P(M_k|\boldsymbol{y})$.

Each improper prior *could* be $C_k p(\boldsymbol{\theta}_k|k)$, where C_k is arbitrary. That means we can write

$$\begin{aligned} P(M_k|\boldsymbol{y}) &= \frac{C_k m_k(\boldsymbol{y}) p_k}{\sum_{j=1}^{N} p_j C_j m_j(\boldsymbol{y})} \\ &= \frac{m_k(\boldsymbol{y}) p_k}{\sum_{j=1}^{N} p_j (C_j/C_k) m_j(\boldsymbol{y})}, \end{aligned}$$

and hence $P(M_k|\boldsymbol{y})$ is not well-defined.

So, when doing model selection, we need to use proper priors.

Unfortunately, even weakly informative priors, such as unit information priors, don't always work very well for model selection. This will be seen in Example 17.

In the linear regression setting, consider the problem of selecting a best subset of predictors.

Suppose we have a total of p predictors, and let $\boldsymbol{z} = (z_1, \ldots, z_p)$ be a p-vector of 0s and 1s.

We have $z_j = 1$ if predictor x_j is in the model and 0 otherwise.

For a given model $\boldsymbol{z}$, let $\boldsymbol{X}_{\boldsymbol{z}}$ and $\boldsymbol{\beta}_{\boldsymbol{z}}$ be the corresponding design matrix and vector of regression coefficients.

For example, suppose $\boldsymbol{z} = (1, 0, 1, 0)$. Then

$$\boldsymbol{\beta}_{\boldsymbol{z}}^T = (\beta_1, \beta_3)$$

and $\boldsymbol{X}_{\boldsymbol{z}}$ is an $n \times 2$ matrix whose first and third columns contain $x_{11}, \ldots, x_{n1}$ and $x_{13}, \ldots, x_{n3}$, respectively.

For the prior $p(\boldsymbol{\beta}_{\boldsymbol{z}}, \sigma^2|\boldsymbol{z})$ suppose that we use

$$\boldsymbol{\beta}_{\boldsymbol{z}}|(\sigma^2, \boldsymbol{z}) \sim N(\mathbf{0}, n\sigma^2(\boldsymbol{X}_{\boldsymbol{z}}^T\boldsymbol{X}_{\boldsymbol{z}})^{-1})$$

and

$$\sigma^2|\boldsymbol{z} \sim \text{inverse gamma}\left(\frac{1}{2}, \frac{s_{\boldsymbol{z}}^2}{2}\right),$$

where $s_{\boldsymbol{z}}^2$ is the estimated error variance under model $\boldsymbol{z}$.

Notes:

- The mean vector $\mathbf{0}$ in the prior for the regression coefficients results from an invariance argument (pp. 156-57).

- Using $s_{\boldsymbol{z}}^2$ in the prior for σ^2 is "cheating," but is consistent with the principle of "unbiased but weak prior information." The weak part comes from using 1/2 as the first parameter.

For a model $\boldsymbol{z}$ define

$$\text{RSS}_{\boldsymbol{z}} = \boldsymbol{y}^T \left[\boldsymbol{I} - \left(\frac{n}{n+1}\right) \boldsymbol{X}_{\boldsymbol{z}} (\boldsymbol{X}_{\boldsymbol{z}}^T \boldsymbol{X}_{\boldsymbol{z}})^{-1} \boldsymbol{X}_{\boldsymbol{z}}^T \right] \boldsymbol{y}.$$

Note that the usual RSS, i.e., residual sum of squares, does not have the factor of $n/(n+1)$.

Hoff shows that using the prior on the previous page leads to

$$\frac{p(\boldsymbol{z}_a|\boldsymbol{y})}{p(\boldsymbol{z}_b|\boldsymbol{y})} = \frac{p(\boldsymbol{z}_a)}{p(\boldsymbol{z}_b)} \cdot \text{Bayes factor},$$

where $p(\boldsymbol{z})$ is the prior probability of model $\boldsymbol{z}$ and

$$\begin{aligned} \text{Bayes factor} &= (1+n)^{(p_b-p_a)/2} \left(\frac{s^2_{\boldsymbol{z}_a}}{s^2_{\boldsymbol{z}_b}}\right)^{1/2} \\ &\quad \times \left(\frac{s^2_{\boldsymbol{z}_b} + \text{RSS}_{\boldsymbol{z}_b}}{s^2_{\boldsymbol{z}_a} + \text{RSS}_{\boldsymbol{z}_a}}\right)^{(n+1)/2}, \end{aligned}$$

where p_a and p_b are the numbers of 1s in the vectors $\boldsymbol{z}_a$ and $\boldsymbol{z}_b$, respectively.

Example 17 Model selection in the oxygen uptake example

We'll use the priors of p. 234N and evaluate each of the 15 subsets of predictors. We'll use the full model (with all four predictors) as a benchmark.

In other words, z_b in the odds ratio on p. 235N will always be the full model.

I will assign an equal prior probability of $1/15$ to each of the 15 models, meaning that the Bayes factor on p. 235 is the ratio of the posterior probabilities of two models.

Results for oxygen uptake data

Model	Probability relative to full model	BIC
x_1	0.007	66.61
x_2	0.001	72.12
x_3	0.049	60.84
x_4	0.000	75.51
x_1, x_2	0.004	66.09
x_1, x_3	0.411	51.52
x_1, x_4	0.002	67.88
x_2, x_3	0.031	59.96
x_2, x_4	0.000	74.66
x_3, x_4	0.053	58.32
x_1, x_2, x_3	0.230	50.62
x_1, x_2, x_4	0.002	67.08
x_1, x_3, x_4	1.578	43.63
x_2, x_3, x_4	0.037	56.85
x_1, x_2, x_3, x_4	1.000	41.70

The procedure of p. 235 points to the model with the predictors x_1, x_3 and x_4 as being best.

Our Gibbs sampling suggested that all four predictors should be in the model.

An all purpose model selector with a Bayesian motivation is BIC.

According to this criterion we choose the model with the smallest value of

$$BIC(k) = -2 \log p(\boldsymbol{y}|\widehat{\boldsymbol{\theta}}_k, k) + d_k \log n,$$

where $\widehat{\boldsymbol{\theta}}_k$ is the MLE for model k and d_k is the number of parameters in model k.

The BIC values on p. 237 point to the full model as being the best, which agrees with the results of our Gibbs sampling.

This example is to be continued . . .

It turns out that *if the prior variance for $\beta_{\mathcal{Z}}$ is too large, then there is a strong tendency for too small a model to have the largest posterior probability.*

This point has been made by a number of authors, including those below:

- Chipman, H., George, E.I. and McCulloch, R.E. (2001). The practical implementation of Bayesian model selection (with discussion). *IMS Lecture Notes - Monograph Series* **38** 65-116.

- Berger, J.O. and Pericchi, L.R. (1996). The intrinsic Bayes factor for model selection and prediction. *Journal of the American Statistical Association* **91** 109-122.

Instead of the prior on p. 235N, suppose we use

$$\boldsymbol{\beta}_{\boldsymbol{z}}|(\sigma^2, \boldsymbol{z}) \sim N(\mathbf{0}, g\sigma^2(\boldsymbol{X}_{\boldsymbol{z}}^T\boldsymbol{X}_{\boldsymbol{z}})^{-1})$$

and

$$\sigma^2|\boldsymbol{z} \sim \text{inverse gamma}\left(\frac{\nu_0}{2}, \frac{s^2}{2}\right),$$

where s^2 is the sample variance of $\boldsymbol{y}$.

The use of s^2 was suggested by Chipman, George and McCulloch (2001) (CGM), since it represents an upper bound for σ^2.

CGM recommend a small value, such as 2 or 3, for ν_0.

The prior for $\boldsymbol{\beta}_{\boldsymbol{z}}$ is the so-called g-prior proposed by Zellner (1986).

Previously we used $g = n$, but as we saw in our example this can produce a model smaller than what seems reasonable.

A common way of standardizing the independent variables is as follows:

$$v_{ij} = \frac{x_{ij} - \bar{x}_i}{s_i}, \quad i = 1, \ldots, p, \ j = 1, \ldots, n,$$

where $\bar{x}_i$ and s_i are the mean and standard deviation, respectively, of $x_{i1}, \ldots, x_{in}$.

If the predictors are so standardized, CGM recommend using $g = (2.85)^2$.

Using the prior on p. 241N, we have

$$\frac{p(\boldsymbol{z}_a|\boldsymbol{y})}{p(\boldsymbol{z}_b|\boldsymbol{y})} = \frac{p(\boldsymbol{z}_a)}{p(\boldsymbol{z}_b)} \cdot (1+g)^{(p_b - p_a)/2} \times \left(\frac{\nu_0 \sigma_0^2 + \mathrm{RSS}_{g,\boldsymbol{z}_b}}{\nu_0 \sigma_0^2 + \mathrm{RSS}_{g,\boldsymbol{z}_a}} \right)^{(n+\nu_0)/2},$$

where $\mathrm{RSS}_{g,\boldsymbol{z}}$ is defined as was $\mathrm{RSS}_{\boldsymbol{z}}$ but with g in place of n.

Example 17 continued

I recomputed the odds ratios, this time using the formula at the bottom of p. 242N. (I standardized the predictors as on p. 242N.)

I took $\nu_0 = 2$, $\sigma_0^2 = 0.198$ (the sample variance of the y_is) and $g = (2.85)^2$.

As before I gave equal prior probability to each of the fifteen models.

The resulting odds ratios are given on the next page.

Posterior probability of each model in the oxygen uptake example relative to that of the full model

Model	Probability relative to full model
x_1	0.000
x_2	0.000
x_3	0.000
x_4	0.000
x_1, x_2	0.000
x_1, x_3	0.009
x_1, x_4	0.000
x_2, x_3	0.000
x_2, x_4	0.000
x_3, x_4	0.001
x_1, x_2, x_3	0.021
x_1, x_2, x_4	0.000
x_1, x_3, x_4	0.324
x_2, x_3, x_4	0.002
x_1, x_2, x_3, x_4	1.000

This analysis places the highest probability on the full model. The model with x_1, x_3 and x_4 is reasonably probable, but none of the others seem viable.

Note that knowing all the odds ratios tells us the posterior probability of each model.

Let P_j be the posterior probability of model j. The odds ratios are $o_j = P_j/P_N$, and so

$$\sum_{j=1}^{N} o_j = \frac{1}{P_N}.$$

Therefore,

$$P_j = \frac{o_j}{\sum_{i=1}^{N} o_i}, \qquad j = 1, \ldots, n.$$

Using this fact, the posterior probability of the full model is 0.729 and that of the model with x_1, x_3 and x_4 is 0.245.

No other model has posterior probability higher than 0.0168.

It is interesting to look at how BIC compares with the Bayesian posterior probabilities. For our linear model,

$$\begin{aligned} BIC(\boldsymbol{z}) &= n\left[\log \hat{\sigma}^2_{\boldsymbol{z}} + 1 + \log(2\pi)\right] \\ &\quad +(p+1)\log n, \end{aligned}$$

where $\hat{\sigma}^2_{\boldsymbol{z}}$ is the MLE of σ^2 for model $\boldsymbol{z}$.

I can compare $\boldsymbol{z}_a$ and $\boldsymbol{z}_b$ by taking the difference of their BIC values. This gives

$$BIC(\boldsymbol{z}_b) - BIC(\boldsymbol{z}_a) =$$

$$n\log(\hat{\sigma}^2_{\boldsymbol{z}_b}/\hat{\sigma}^2_{\boldsymbol{z}_a}) + (p_b - p_a)\log n.$$

Now, from page 242N, if we take $p(\boldsymbol{z}_a) = p(\boldsymbol{z}_b)$, then

$$2\log\left[\frac{p(\boldsymbol{z}_a|\boldsymbol{y})}{p(\boldsymbol{z}_b|\boldsymbol{y})}\right] =$$

$$(n+\nu_0)\log\left[\frac{\nu_0\sigma_0^2 + \mathrm{RSS}_{g,\boldsymbol{z}_b}}{\nu_0\sigma_0^2 + \mathrm{RSS}_{g,\boldsymbol{z}_a}}\right]$$

$$+(p_b - p_a)\log(g+1).$$

So, there is a strong similarity between the difference of BIC values and

$$2 \log \left[\frac{p(\boldsymbol{z}_a | \boldsymbol{y})}{p(\boldsymbol{z}_b | \boldsymbol{y})} \right].$$

In fact, for large n they are essentially the same if we take $g = n$.

However, if we take g to be a fixed constant, as suggested by CGM, then the two criteria can be quite different.