Estimating $P(\beta_j \neq 0|\boldsymbol{y})$ in a Linear Regression Model

We use the algorithm at the bottom of p. 167 and the top of p. 168 in Hoff.

We use the notation as in class: $\boldsymbol{z} = (z_1, \ldots, z_p)$ is a vector of 0s and 1s. If z_j is 1, then variable x_j is included in the model; otherwise it is not. The notation $\boldsymbol{z}_{-j}$ denotes a vector of length $p-1$ containing all components of $\boldsymbol{z}$ except z_j.

Outline of the algorithm

We use Gibbs sampling to draw samples from the *marginal* posterior distribution of $\boldsymbol{z}$ given the data.

Since each z_j is binary, it is obvious that the full conditional of z_j given $\boldsymbol{z}_{-j}$ is Bernoulli. We just need to figure out the Bernoulli success probability of each full conditional.

We have $P(z_j = 1|\boldsymbol{y}, \boldsymbol{z}_{-j}) = o_j/(1 + o_j)$, where o_j is defined in the middle of p. 167 of Hoff.

So, in doing the Gibbs sampling, component z_j is updated by generating one value from the Bernoulli distribution with success probability $o_j/(1 + o_j)$.

Once all the components of $\boldsymbol{z}$ have been updated, one draws $(\boldsymbol{\beta}, \sigma^2)$ from the posterior distribution $p(\boldsymbol{\beta}, \sigma^2|\boldsymbol{y}, \boldsymbol{z})$, where $\boldsymbol{z}$ is the updated model.

In the R program at eCampus, I have used a g-prior ($g = n$) for the regression coefficients of each model, as defined at the top of p. 165.

As prior for σ^2 given model $\boldsymbol{z}$, I use inverse-gamma$(1/2, s_{\boldsymbol{z}}^2/2)$, where $s_{\boldsymbol{z}}^2$ is the estimated error variance for model $\boldsymbol{z}$.

For the regression model with normal errors, these priors imply that σ^2 given $(\boldsymbol{y}, \boldsymbol{z})$ is inverse-gamma with parameters $(n+1)/2$ and $(s_{\boldsymbol{z}}^2 + \mathrm{SSR}_n^{\boldsymbol{z}})/2$, where $\mathrm{SSR}_n^{\boldsymbol{z}}$ is defined in the middle of p. 165.

The posterior of $\boldsymbol{\beta}$ given $(\sigma^2, \boldsymbol{y}, \boldsymbol{z})$ is multivariate normal with mean vector

$$\frac{n}{n+1}(\boldsymbol{X}_{\boldsymbol{z}}^T \boldsymbol{X}_{\boldsymbol{z}})^{-1} \boldsymbol{X}_{\boldsymbol{z}}^T \boldsymbol{y}$$

and covariance matrix

$$\frac{n}{n+1}\sigma^2(\boldsymbol{X}_{\boldsymbol{z}}^T \boldsymbol{X}_{\boldsymbol{z}})^{-1}.$$

Summary of Output for Exercise 9.2(b)

The following results are based on 10,000 replications.

Variable	$P(\beta_j \neq 0 \mid \boldsymbol{y})$
npreg	0.1026
bp	0.1636
skin	0.0955
bmi	0.9826
ped	0.6823
age	1.0000

Variable	95% credible interval for β_j
npreg	[-1.05, 0.00]
bp	[0.00, 0.32]
skin	[0.00, 0.36]
bmi	[0.40, 1.33]
ped	[0.00, 17.16]
age	[0.49, 1.02]