Practical 3


**Aim:** Selection of explanatory variables in a linear regression setting, through exact computation and Gibbs’ sampling.

**Reference:** *Bayesian Core – a practical approach to Computational Bayesian Statistics* by Jean-Michel Marin and Christian P. Robert, chapter 3.

In this practical, we would like to perform Bayesian linear regression to explain the weight at birth of babies.

Download the data:

```r
> install.packages("MASS")
> require(MASS)
> data(birthwt)
> data(birthwt)
```

There are 7 explanatory variables:

1. **lwt:** Mother’s weight before pregnancy
2. **race:** mother’s race (1=white, 2=black, 3=other); **categorical** variable
3. **smoke:** smoking status during pregnancy
4. **ptl:** number of previous premature labours
5. **ht:** history of hypertension
6. **ui:** presence of uterine irritability
7. **ftv:** number of physician visits during the first trimester

The output variable is **bwt**, the birth weight in grams. There is an additional output variable: **low**, which indicates whether the birth weight was less than 2500 grams.

Our linear model is

\[
\begin{align*}
  y_i &= \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_8 x_{8i} + \epsilon_i \\
  \epsilon_i &\sim N(0, \sigma^2)
\end{align*}
\]
1. Transform the categorical variable `race` so that it can be used in linear regression.

2. **Frequentist analysis** Give the frequentist estimates of $\beta$ and $\sigma^2$. You can use the R function `lm`.

   ```
   > summary(lm(y~X))
   ```

3. **Bayesian inference using Zellner’s G-prior** Consider the following prior:

   \[
   \begin{align*}
   \beta|\sigma^2, X &\sim N_{p+1}\left(\tilde{\beta}, g\sigma^2(tXX)^{-1}\right) \\
   \pi(\sigma^2) &\propto \sigma^{-2}
   \end{align*}
   \]

   (a) Take $\tilde{\beta} = 0_p$. Calculate the posterior distribution of $(\beta, \sigma^2)$.

   (b) For $g = 0.1, 1, 10, 100, 1000$, give $E[\sigma^2|y, X]$ and $E[\beta_0|y, X]$. What can you conclude about the impact of the prior on the posterior?

   (c) We would like to test the hypothesis $H_0 : \beta_7 = \beta_8 = 0$.

      i. Give the explicit form of the marginal likelihood $m(y|X)$.

      ii. Deduce the expression of the corresponding Bayes factor.

      iii. Compute this quantity given our data and conclude, using Jeffreys’ scale of evidence.

4. **Model choice: exact computation** In this question, we restrict ourselves to the first 3 explanatory variables.

   ```
   > X1 = X[,1:4]
   ```

   We would like to know which variables to include in our model, and assume that the intercept is necessarily included. We have $2^3 = 8$ possible models. To each model, we associate the variable $\gamma = (\gamma_1, \gamma_2, \gamma_3)$ where $\gamma_i = 1$ if $x_i$ is included in the model, and $\gamma_i = 0$ otherwise. Let $X_\gamma$ be the submatrix of $X$ where we only keep the columns $i$ such that $\gamma_i = 1$.

   Using the marginal probability formula of question 3(c)i, compute for each model its marginal likelihood $m(y|X_\gamma)$. Deduce the most likely model a posteriori.

5. **Non-informative prior** (Optional) We now move on to a non-informative setting. We still use Zellner’s prior, but the hyperparameter $g$ is no longer fixed: we take a hyperprior

   \[
   \pi(g) \propto g^{-1}1_{\mathbb{N}^+}(g)
   \]

   (a) Give the posterior distribution $\pi(\beta, \sigma^2|y, X)$. Compare your estimators with those you obtained for $g = 100$.

   (b) Test again the hypothesis $H_0 : \beta_7 = \beta_8 = 0$. Justify the use of a Bayes’ factor. Do you get the same conclusion?

6. **Model choice using Gibbs’ sampling** We now consider all 8 explanatory variables. We thus need to choose between $2^9$ models. Write a Gibbs’ sampler which samples from the posterior distribution of $\gamma$, and conclude on the most likely model.