Bayesian Statistics Assignment_Hsuan Lee

h.lee1@students.uu.nl

Research Question:

The issue of concern in the study is: To what extent male "body fat" could be predicted by "age" and "height".

Data Description:

The data lists estimates of the percentage of body fat determined by underwater weighing and various body circumference measurements for 252 men without missing data. The data were supplied by Dr. A. Garth Fisher who gave permission to freely distribute the data and use for non-commercial purposes.

Variables Description:

Three variables were used in the study, which were "Body Fat", "Age", and "Height" respectively. Body Fat was the dependent variable in this study, "Age", and "Height" were the independent variables. To facilitate the operation of Markov Chain Monte Carlo, all three variables were centered. The variables descriptions of them are:

Body Fat – Participants' percent of body fat, which was computed from Siri's body fat equation in 1956

Age – Age of participants measured by year

Height – Height of participants measured by inch

Statistical Method:

In this study, the Bayesian approach was used for the analysis, specifying the Bayesian linear regression model with the variable *Body Fat* as y, variable *Age* as x_1, variable *Height* as x_2.

The model could be expressed as:

$$y_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + \sigma^2$$

To estimate the parameters of concern for this study, Metropolis-Hasting sampler was used, as the study was based on an uninformative prior, the non-conjugate prior "non-standardized t distribution" is capable of allowing more extreme values and was therefore assigned to the coefficients b_1 and b_2. Whereas, such a conditional distribution did not possess the proper form that the Gibbs Sampler could sample from, leading to the employment of the Metropolis-Hasting sampler. Thereafter, the convergence was examined by the trace plots, density plots, autocorrelation plots, and Gelman-Rubin plots. Following this was the interpretation of the estimates and intervals of the model.

In model selection section, DIC was used to compare three models which were: a model containing Age and Height as predictors (the model used for the analysis); a model with Age only as a predictor; and a null model with intercept. As well, Bayes factor would be implemented to compare the informative hypotheses on whether Age and Height were positively and nearly equally associated with Body Fat in male.

The comparison of Bayesian and frequent methods is the last section, which first discusses the main differences between Bayesian and frequent methods, followed by their comparison based on the example of this study.

Estimation & Metropolis-Hasting Sampler:

Prior Distribution Specification:
In this study, an uninformative prior distribution was used for each parameter, since no data from previous studies were available and accessible. Furthermore, since the normal distribution of the data was known, it could be inferred that the parameters have conjugate priors. The intercept (b_0) and coefficients (b_1,b_2) of the model would follow a normal distribution with mean (μ_{j0}) equal to 0 and variance (ζ^2_{j0}) equal to 100000 (non-informative prior, variance should be large); the residual variance (σ^2) would follow an inverse gamma distribution with shape (α_0) equal to 0.001 and scale (β_0) equal to 0.001. They could be expressed as:

$$b_j \sim N\left(\mu_{j0}, \zeta^2_{j0}\right), j=0,1,2$$
$$\sigma^2 \sim Ig\left(\alpha_0, \beta_0\right)$$

Gibbs Sampler:
The conditional posterior distributions were derived and employed in Gibbs sampler to obtain the joint posterior distribution instead of directly utilizing the joint posterior distribution. Each parameter would be sampled from the conditional posterior distribution to which it belongs, a specific number of iterations would be performed. The derived conditional posterior distributions for each parameter in the model used in the study were:

The conditional posterior mean(μ_{01}) and variance(ζ^2_{01}) of the intercept (b_0) from normal distribution:

$$\mu_{01} = \frac{\sum_{i=1}^{N}\left(y_i - b_1 x_{1i} - b_2 x_{2i}\right) / \sigma^2 + \left(\mu_{00} / \zeta^2_{00}\right)}{\left(N / \sigma^2\right) + \left(1 / \zeta^2_{00}\right)};\ \zeta^2_{01} = \frac{1}{\left(N / \sigma^2\right) + \left(1 / \zeta^2_{00}\right)}$$

The conditional posterior mean(μ_{j1}) and variance(ζ^2_{j1}) of the coefficients (b_j, j = 1, 2) from normal distribution:

$$\mu_{j1} = \frac{\sum_{i=1}^{N} x_{ji}\left(y_i - b_0 - b_{3-j} x_{(3-j)i}\right) / \sigma^2 + \left(\mu_{j0} / \zeta^2_{j0}\right)}{\sum\left(x^2_{ji}\right) / \sigma^2 + \left(1 / \zeta^2_{j0}\right)};\ \zeta^2_{j1} = \frac{1}{\sum\left(x^2_{ji}\right) / \sigma^2 + \left(1 / \zeta^2_{j0}\right)}$$

The conditional posterior shape parameter(α_1) and scale parameter(β_1) of the residual variance(σ^2) from inverse-gamma distribution:

$$\alpha_1 = \frac{N}{2} + \alpha_0;\ \beta_1 = \frac{S}{2} + \beta_0$$

, where S is the sum of squared residuals:

$$S = \sum_{i=1}^{N}\left(y_i - \left(b_0 + b_1 x_{1i} + b_2 x_{2i}\right)\right)^2$$

In order to manipulate the Gibbs sampling and evaluate the number of iterations(H) it should take, multiple chains were needed to inspect convergence, each chain should have a different value chosen at random as the starting value of each parameter. In this study two chains were used.

After specifying the starting values and the number of iterations(H) for each parameter, sampling could begin. In this study, the number of iterations(H) was set to be 10000, ***h*** *represented* the order of the iterations. For each new iteration, h would be set equal to $h+1$. The sampler first sampled a value for $b_0^{(h)}$ from its conditional posterior distribution $p\left(b_0 \middle| b_1^{(h-1)}, b_2^{(h-1)}, \sigma^{2(h-1)}; y\right)$, next sampled for $b_1^{(h)}$ from its conditional posterior distribution $p\left(b_1 \middle| b_0^{(h)}, b_2^{(h-1)}, \sigma^{2(h-1)}; y\right)$, afterwards sampled for $b_2^{(h)}$ from its conditional posterior distribution

$p\left(b_2 \middle| b_0^{(h)}, b_1^{(h)}, \sigma^{2(h-1)}; y\right)$, followed by $\sigma^{2(h)}$ from its conditional posterior distribution $p\left(\sigma^2 \middle| b_0^{(h)}, b_1^{(h)}, b_2^{(h)}; y\right)$. Repeat these steps until $\boldsymbol{h} = \boldsymbol{H}$, the sampling was complete.

Metropolis-Hasting Sampler:
The Metropolis-Hasting Algorithm was employed when the normalizing constant cannot be found; or the derived conditional posterior cannot be identified the type of distribution. As this study was based on a weak uninformative prior, the non-conjugate prior "non-standardized t distribution" is capable of allowing more extreme values and was therefore assigned to the coefficients b_1 and b_2. However, such a conditional distribution did not possess the proper form that the Gibbs Sampler could sample from. The MH sampler therefore would be implemented.

The key difference between MH sampler and Gibbs Sampler is that instead of sampling from the conditional posterior distribution, MH sampler iteratively samples the values from a proposal distribution, which is a different distribution than the conditional posterior. In this study standard normal distribution was assigned as the proposal distribution for b_1 and b_2. As random walk MH sampler was used in this study, the proposal density was centered around the current value $b_{j,t-1}$ of the parameter. The tuning parameters τ^2 of the proposal distribution were both specified as 0.5 in this study. The proposal density could be expressed as:

$$q(b^*|b_{j,t-1}) = N(b^*|b_{j,t-1}, \tau^2)\ ,\ j=1,2$$

After sampling the candidate value from the proposal distribution, a random value *u* should be drawn from a uniform distribution on the interval from 0 to 1. Afterwards, the acceptance ratio *r* was computed, followed by a comparison between *r* and *u* to decide whether to accept the candidate value.

The acceptance ratio *r* for b_1 could be computed by:

$$r=\frac{p(b^*|x,y,b_{0,t},b_{2,t-1},\sigma^2_{t-1})}{p(b_{1,t-1}|x,y,b_{0,t},b_{2,t-1},\sigma^2_{t-1})}\frac{q(b_{1,t-1}|b^*)}{q(b^*|b_{1,t-1})} = \frac{p(b^*|x,y,b_{0,t},b_{2,t-1},\sigma^2_{t-1})}{p(b_{1,t-1}|x,y,b_{0,t},b_{2,t-1},\sigma^2_{t-1})}$$

The acceptance ratio *r* for b_2 could be computed by:

$$r=\frac{p(b^*|x,y,b_{0,t},b_{1,t},\sigma^2_{t-1})}{p(b_{2,t-1}|x,y,b_{0,t},b_{1,t},\sigma^2_{t-1})}\frac{q(b_{2,t-1}|b^*)}{q(b^*|b_{2,t-1})} = \frac{p(b^*|x,y,b_{0,t},b_{1,t},\sigma^2_{t-1})}{p(b_{2,t-1}|x,y,b_{0,t},b_{1,t},\sigma^2_{t-1})}$$

As the proposal distribution is symmetric, the ratio of the proposal distribution reduced to 1, i.e., the acceptance ratio *r* would be equal to the ratio of the proportional conditional posterior.

To decide whether to accept the candidate value or retain the current value, the random value *u* drawn from a uniform distribution on the interval from 0 to 1 was compared with the acceptance ratio *r*. If $u \le r$ the candidate value became the current value, otherwise the current value was retained. That is:

$$b_{j,t} = \begin{cases} b^*, \text{ if } u \le r \\ b_{j,t-1}, \text{ if } u > r \end{cases},\ j=1,2$$

After the above steps, the simulation of the target posterior distribution were expected to be completed.

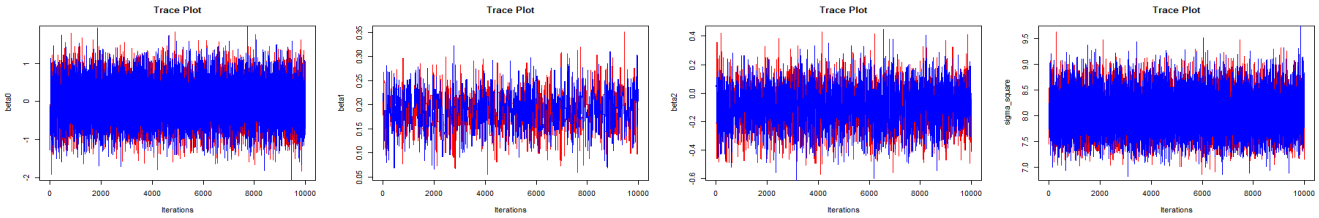
Convergence:

In the study, two chains were employed, the number of iterations were set to 10000 with 500 burn-in periods. That is, each parameter has 10500 sampled values, the first 500 samples are dropped to be used for convergence.

To inspect the convergence, trace plot, density plot, autocorrelation plot, and Gelman-Rubin plot were used. The burn-in periods were not considered in these plots as they have already been removed.

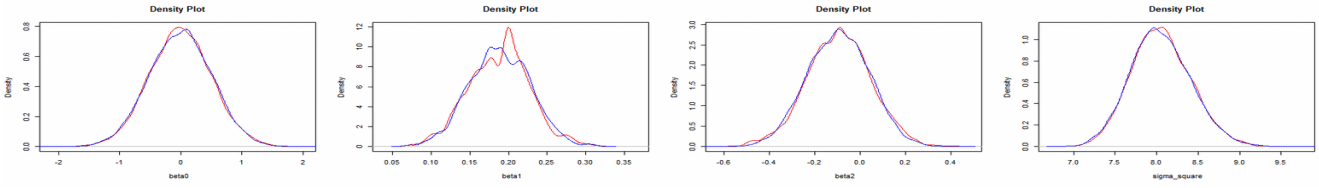
Trace Plot:

Trace plot show the sampled parameters over the iterations. In this study the intercept b_0, coefficient b_2 and residual variance σ^2 were stable, as the two chains resemble a fat hairy caterpillar. The coefficient b_1, however, did not achieve a perfect convergence, yet it was not bad either. One notable point was that b_1 experienced a bit of autocorrelation, as a few iterations were suspended at the same value, but not severely, which could be studied more closely in the autocorrelation plot.



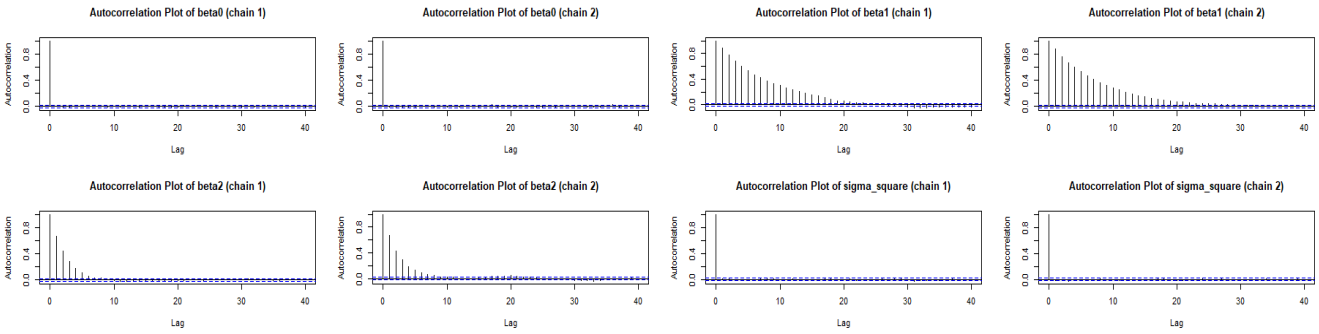
Density Plot:

The two chains of the intercept b_0, coefficient b_2 and residual variance σ^2 were overlapped, denoting that good convergence were reached for the three parameters. The two chains of coefficient b_1, however, did not exhibit good overlap at the peak of the plot, yet again, the non-overlap situation was not remarkably severe.



Autocorrelation Plot:

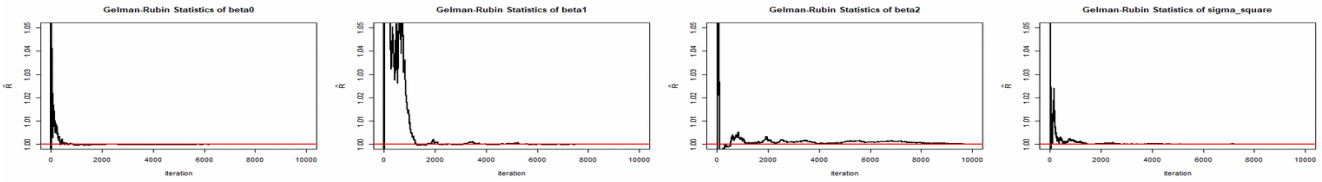
The autocorrelation was expected to be smaller as the lags increased. The intercept b_0 and residual variance σ^2 were zero after lag 0, denoting that satisfactory autocorrelation case were achieved for the two parameters. The coefficient b_1 and b_2, however, did not exhibit satisfied autocorrelation cases, as the autocorrelation issues happened per 20 lags for b_1, and per 8 lags for b_2, which yielded a high degree of correlation between the draws and slow mixing. Yet, due to the fact that the MH sampler was performed in both parameters, such extent of autocorrelation issues were acceptable.



Gelman-Rubin Plot:

Gelman-Rubin Statistics compares an estimate of the total variance of the samples in all the chains to the average variance within the chains, which is supposed to be near 1, given that this represents almost no "between-chain" variance. As the Gelman-Rubin plots shown below, in this study, all four parameters have a

seemingly satisfactory Gelman-Rubin statistics, reaching up to 1, signifying that the convergence was achieved for each parameter.



Interpretation of Estimates & Intervals

As shown in *table 1*, the two variables of concern for the study, *Age* and *Height*, only *Age* displayed a significant relation with the dependent variable *Body Fat*, which could be observed from the 95% credible intervals. As the coefficient b_1 for *Age* landed between 0.108 and 0.268, implying a 95% probability that b_1 lie between this interval excluding 0; The coefficient of *Height*, b_2, however, fell between -0.369 and 0.174, signifying that with 95% probability b_2 was located between the interval containing 0, namely, the lack of significant relation between *Height* and *Body Fat*.

With the above explanation, the study could be interpreted to be that, for a male one year older than the average, his body fat would be expected to be 0.188 percent higher than average. As well, with an average age male who is one inch taller than average, his body fat would be expected to be 0.094 percent lower than the average. Yet, as height is not significantly connected to body fat, the second interpretation is negligible.

	Mean (95% CI)	*Standard Deviation*	*Monte Carlo error*
Intercept - b_0	*-0.004 (-0.990-0.998)*	*0.508*	*0.000025*
Age - b_1	*0.188(0.108-0.268)*	*0.040*	*0.000002*
Height - b_2	*-0.094(-0.369-0.174)*	*0.139*	*0.000007*
Residual Variance - σ^2	*8.052(7.381-8.800)*	*0.362*	*0.000018*

Table 1. Estimates of the Parameters

Model Selection Using DIC

To ensure the best model was chosen for analysis in the study, DIC was employed, as it minimizes the loss in decision or deviance. In this study three different models were compared, i.e., a model containing *Age* and *Height* as predictors; a model consisting of *Age* as a predictor only; and a null model with intercept. DIC consists of two sections, model misfit and model complexity; in the model misfit section, log-likelihood assessed on the posterior mean of the parameters is used; in model complexity section, the effective number of parameters in the model are estimated. The lower the DIC value, the better the model.

The DIC of the model used for the analysis (*Age* and *Height* as predictors) was 1769.95 with estimated parameters of 3.91; the DIC of the model consisting of *Age* only as a predictor was 1768.38 with estimated parameters of 2.92; the DIC of the empty model was 1788.90 with estimated parameters of 2.00. Evidently, the null model was the worst model with the highest DIC value. Yet, the model with *Age* and *Height* as predictors and the model with *Age* only as predictor possessed similar DIC values, which results from the fact that *Height* is not an influential predictor. With nearly identical DIC values, coupled with approximately the equivalent estimate for the variable *Age* ($b_1 = 0.194$ for the *Age*-only model), the model employed for the analysis (with *Age* and *Height* as predictors) could be maintained, however.

the Bayes Factor

As the Bayes factor quantifies the relative support in the data for two hypotheses by the ratio of two marginal likelihoods, the informative hypotheses were tested by the Bayes factor in this study. Given that the marginal likelihood ratio between the informative(m_i) and uninformative(m_u) hypotheses could be derived as the ratio of the fit(f_i) and the complexity(c_i) of the informative hypothesis. The Bayes factor could be expressed as:

$$BF_{iu} = \frac{m_i}{m_u} \approx \frac{f(y|b,\sigma^2,x)h_i(b_1,b_2|[y,x]^d)}{g_i(b_1,b_2|y,x)} \Big/ \frac{f(y|b,\sigma^2,x)h_u(b_1,b_2|[y,x]^d)}{g_u(b_1,b_2|y,x)}$$
$$= \frac{1/c_i \times h_u(b_1,b_2|[y,x]^d)}{1/f_i \times g_u(b_1,b_2|y,x)} \Big/ \frac{h_u(b_1,b_2|[y,x]^d)}{g_u(b_1,b_2|y,x)} = \frac{f_i}{c_i}$$

,where ***h*** *represents the fractional prior of the hypothesis;* ***g*** *represents the posterior of the hypothesis;* ***d*** *denotes a fraction of the density of the data used to specify a prior distribution.*

The informative hypotheses test in this study were whether *Age* and *Height* were positively and nearly equally associated with *Body Fat* in male. That is,

$$H_1: b_1 \approx b_2, \text{ that is }, |b_1 - b_2| < 0.1$$
$$H_2: b_1 > 0, b_2 > 0$$
$$H_u: b_1, b_2$$

As *Age* and *Height* were not measured in the same scale, the estimates b_1 and b_2 were standardized. On another note, it was known that the complexity(c_i) and the fit(f_i) were computed based on sampling from the prior and posterior distributions. To obtain the Bayes factor, a normal approximation of the posterior distribution of b_1, b_2 were sampled, as well as the fractional prior distribution with the mean equaled 0 (the boundary of the considered hypotheses). Following the observation of the data, it was evident that H_1 was 4.28 times more likely than H_u as $BF_{1u} \approx 4.28$; H_2 was 2.64 times more likely than H_u as $BF_{2u} \approx 2.64$; H_2 was 0.62 times more likely than H_1 as $BF_{21} \approx 0.62$. Overall, first: the effect of *Age* and *Height* on *Body Fat* were not differed by much; second: both predictors were likely to have a positive link to *Body Fat*, albeit the second finding was not as robust as the first, as $BF_{21} \approx 0.62$.

Comparison of Bayesian and Frequentist Approach

Bayesian approach and Frequentist approach are fundamentally distinct from each other. In Bayesian approach, the parameters of study concern are considered to be random variables, whereas the data is fixed. Namely, all relevant information required to perform inferences is embedded in the observed data. Provided that the prior probability model is properly specified, validity is maintained irrespective of the pre-specified experimental design. With the prior specified, the given data are bound with the prior to construct a posterior distribution of each parameter, from which samples are drawn to acquire 95% credible intervals and estimates for each parameter. Notably, as data is fixed in Bayesian approach, the 95% credible intervals could be interpreted as the range in which the researcher has 95% belief that it contains the parameters that are estimated. In frequentist approach the parameters are fixed but unknown quantities, while the data is treated as random variable. The estimates of the true value of the parameters depends on the data, as the data is random, the estimates obtained are diverse for each repetition of the study. Accordingly, the 95% confidence intervals could be interpreted that among all the intervals computed at the 95% level, 95% of them are supposed to include the true values of the parameters.

Consider this study as an example, though the uninformative prior was used, if the relevant data from previous studies were accessible, the present study could have used them as a reference to obtain an informative prior, which could have led to different results than the present one. Yet, the parameter estimates obtained from the Bayesian linear regression model in this study were approximately equal to those of the frequentist linear model due to the specification of an uninformative prior.