

# **Report Title: Example Report**

Group 13  
Department / Course

November 17, 2025

## Contents

|  |           |
|--|-----------|
| <b>1 Executive Summary</b>                               | <b>3</b>  |
| <b>2 Data Description</b>                                | <b>5</b>  |
| <b>3 Bayesian regression model (BRM)</b>                 | <b>7</b>  |
| 3.1 Reason for BRM . . . . .                             | 7         |
| 3.2 How BRM works . . . . .                              | 7         |
| 3.3 Model Selection . . . . .                            | 9         |
| 3.3.1 Key Variables . . . . .                            | 9         |
| 3.3.2 Fitting result . . . . .                           | 10        |
| 3.3.3 PPC Density . . . . .                              | 13        |
| 3.4 Model Testing . . . . .                              | 13        |
| <b>4 Example</b>   | <b>16</b> |
| <b>5 Widmark's Equation and <math>V_d</math></b>         | <b>18</b> |
| <b>6 Further research:</b>                               | <b>18</b> |
| 6.1 Variable Description . . . . .                       | 18        |
| 6.2 Preliminary analysis . . . . .                       | 18        |
| <b>7 Model selection regarding key variable 'beta60'</b> | <b>19</b> |
| 7.1 Advantage of current method . . . . .                | 19        |

## 1 Executive Summary

In the UK it is a criminal offence to drive a motor vehicle with a blood or breath alcohol concentration above the prescribed limit. When a person is arrested for driving under the influence of alcohol it is not usually possible to perform an accurate test of the level of alcohol in the blood or breath immediately. Breath tests can be used as an initial screening tool at the scene, but these are not sufficiently accurate for prosecution. Instead, people are taken to a police station or hospital, where the test can be carried out using proper laboratory protocols. As the body clears alcohol from the blood through time this means that if the individual was over the limit, the measured blood alcohol concentration (BAC) will be lower at the time of measurement than it was when the person was driving a motor vehicle. To deal with this situation, If the BAC after time  $t$  (hours) is measured as  $C_t$  (g/kg), the BAC at time 0 is estimated as  $C_0 = C_t - \beta t$ , where  $\beta$  (g/kg/h) is BAC elimination rate.

The key point is how to find a precise  $\beta$  to estimate  $C_0$ . Forensic scientists currently 2.5% percentile of  $\beta$  distribution constructing from samples as the estimated  $\beta$  value for every individuals. This method is obviously not rigorous enough for the courts, since:

- The courts will be forced to make decisions under estimated  $\beta$  if we only give a single estimation of *beta*, since the calculated  $C_0$  is either over or under the legal limit.
- Differences between individuals are ignored, for example age and sex, which may affect  $\beta$ .
- $\beta$  value at 2.5% is over conservative and most uncertainties are hiding.

In this report, we will introduce a Bayesian regression model with considering the heterogeneity between individuals, the model gives a posterior distribution of  $\beta$  by using both samples and prior knowledge. Then we randomly simulate 4000  $\beta$  values from posterior distribution and find out the probability of the person's BAC over limit while driving.

In real world,  $\beta$  estimation is quite difficult, it depends on the individuals' liver condition, drinking habits, genetic, diet, etc. As we don't have corresponding data, in our regression model of  $\beta$  is mainly dominate by gender, but we still find some useful variables:

- Gender: Female's BAC elimination rate is larger than male on average, it is explained by liver weight represents a greater fraction of lean body mass in the female gender [2].
- Drinking time after BAC peak: If the person keep drinking after BAC reaches peak, the measured  $\beta$  will be smaller since  $\beta$  is measured starting at BAC peak time till the end, there will be fluctuations on the BAC plots.

When it is too late to use a blood or breath test and the only information available is eyewitness

testimony of the quantity of alcohol consumed. We have to use Widmark's equation:

$$C_t = \frac{A}{Weight \times V_d} - \beta t$$

where  $A$  is Amount of Alcohol Consumed (g),  $V_d$  is the volume of distribution that need to be found. Forensic scientists use the same method again to estimate  $V_d$  separately with  $\beta$ . But  $V_d$  and  $\beta$  are not independent, so we build a joint Bayesian regression model, which can simulate them together to solve with the correlation. Then again after simulation, each  $C_t$  is calculated by each pair of  $\beta$  and  $V_d$ , so expert witness can still find a probability of  $C_t$  exceeding the limit.

To make it easier, we write the method in a function so other expert witness can also get the results by easily plugging new person's data.

## 2 Data Description

It is important to normalized all numerical variables. Centering the covariates can reduce autocorrelation  $\rho_k$  of lag  $k$  which is defined as:

$$\rho_k = \frac{Cov(\theta^i, \theta^{i+k})}{\theta^2},$$

$\rho_k$  measures how similar the sample at position  $i$  is to the sample at position  $i + k$ . As the random walk process,  $\theta_i$  is similar to  $\theta_{i+1}$ , less similar to  $\theta_{i+2}$ , independent with  $\theta_{i+k}$  for large  $k$ . So if  $\rho_k$  is large means the chains hardly move, can't converge.

As a more clear expression, we will introduce Effective Sample Size (ESS):

$$n_{\text{eff}} = \frac{n}{1 + 2 \sum_{k=1}^{\infty} \rho_k}$$

$n_{\text{eff}}$  is larger means the MCMC method converge faster (for example if we set 4 chains with each chains 2000 iterations, 500 burn-in, the  $n_{\text{eff}} = 4000$  means the number of independent samples is 4000 out of 8000). Larger  $n_{\text{eff}}$  in same number of iterations means smaller variance of estimated exoecptation since:

$$V(\hat{E}(\beta)) \approx \frac{\sigma^2}{n_{\text{eff}}}.$$

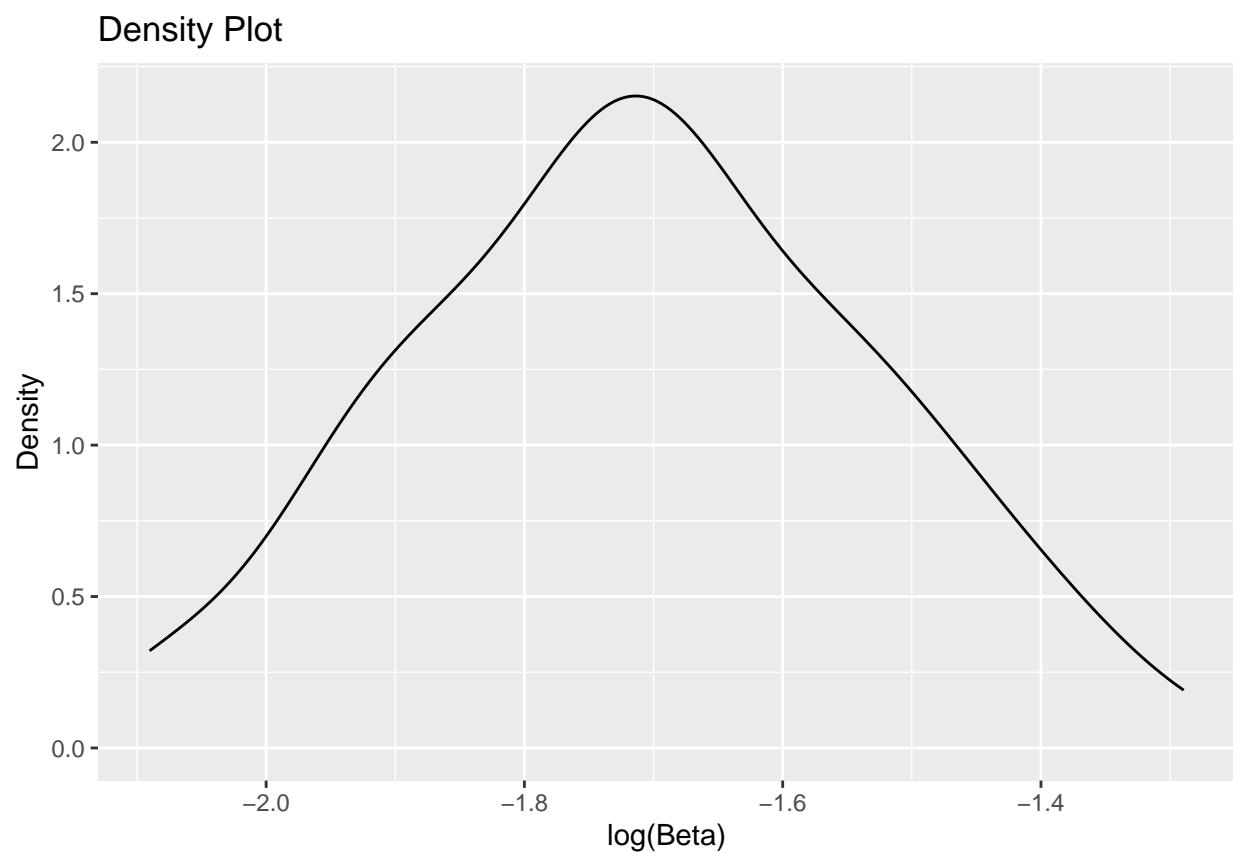
Also Since  $\beta$  is a small value, the coefficients of large value variables are pretty small, near 0, which will make  $\beta$  estimation worse.

Then we prefer to  $\beta$  to  $\log(\beta)$  since the original  $\beta$  distribution has heavy tail, all outliers (if have) can be scaled. As shown in figure 1,  $\beta$  forms a good normal distribution shape after log-transferring. After simulation we can take exponential of the results.

'BAC Peak time' is converted to hours.

Since  $\beta$  is a constant rate, measured when the BAC curve reaches peak and starts to decrease, so variables like AAC and Maximum BAC will not be used, since they are correlated with the value of BAC, not BAC elimination rate.

From the dataset, most data comes from age smaller than 30 (85/100), so even though age is sufficient variable for  $\beta$  the model can't really recognize it. Research shows that age doesn't really matter unless the subject has age-related liver disease, but because of moral and ethical, we don't find much research of alcohol test on elderly person with liver disease.



**Figure 1:**  $\log(\beta)$  distribution.

### 3 Bayesian regression model (BRM)

#### 3.1 Reason for BRM

Comparing with fitting linear regression model for  $\beta_i$ , which can be expressed as:

$$\hat{\beta}_{i,j} = \hat{\gamma}_1 \text{female}_j + \hat{\gamma}_2 \text{male}_j + \hat{\gamma}_3 \text{weight}_j + \hat{\gamma}_4 \text{height}_j + \sigma_i^2$$

and giving a exact value of  $\hat{\beta}_i$  for  $i_{th}$  estimation of  $\hat{\beta}$  under  $\sigma_i^2$  for  $j_{th}$  individuals, BRM can model uncertainty from each coefficient  $\hat{\gamma}$ . We provide priors and sample likelihoods for each variables' coefficients, BRM will apply bayesian rule to each and output simulation results of each coefficients, not only for  $\hat{\beta}$ .  $\hat{\beta}$ s are calculated from simulation results of  $\hat{\gamma}_j$ , which is:

$$\hat{\beta}_{i,j} = \hat{\gamma}_{1,i} \text{female}_j + \hat{\gamma}_{2,i} \text{male}_j + \hat{\gamma}_{3,i} \text{weight}_j + \hat{\gamma}_{4,i} \text{height}_j + \sigma_i^2.$$

Since we have log-transformed  $\beta$  data points, we will take exponential of  $\hat{\beta}_{i,j}$  to get actual value.

#### 3.2 How BRM works

We use *brm* package in R to do this.

The principle theorem behind BRM is Bayes' Rule:

$$p(\theta | \text{data}) = \frac{p(\text{data} | \theta) p(\theta)}{p(\text{data})}.$$

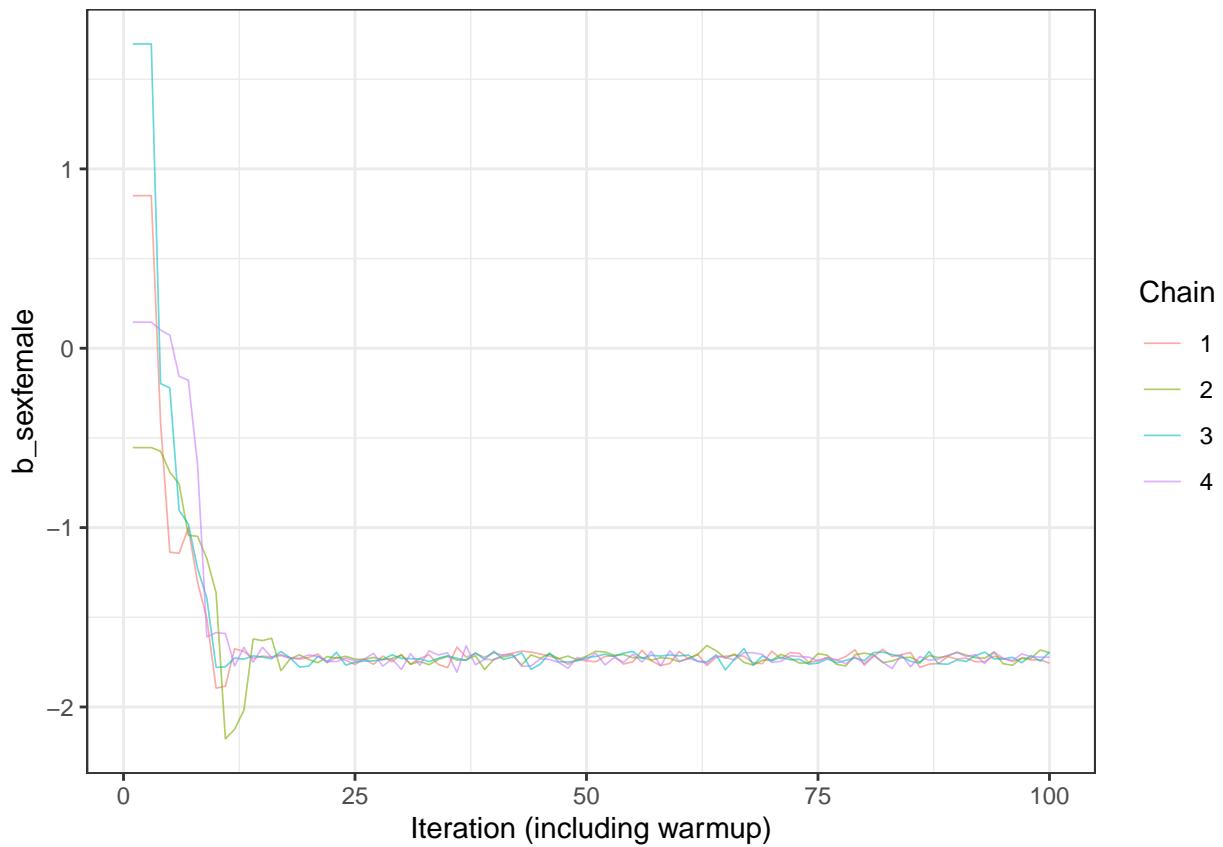
The  $\theta$  in the equation is the coefficients of  $\gamma_j$ . For most real models, we can't compute this posterior  $p(\theta | \text{data})$  analytically. Markov Chain Monte Carlo (MCMC) helps us draw samples from it, which we can then summarize. BRM uses Hamiltonian Monte Carlo (One of MCMC), which is better than standard MCMC (like Metropolis-Hastings) since it is faster and cost less.

After setting priors for  $\gamma_j$ , we need specify a family for  $\beta$  as the the likelihood of the observed data. In our sample, we have already taken  $\log(\beta)$ , so it is better to use the most common normal distribution likelihood, it is normal and can express  $\beta$  values good without restrict samples' shape.

Also for the MCMC simulation method, we set 4 chains, each chains 4000 draws with 1000 burn-in draws. Since MCMC is a random walk method through parameter space. One chain is a single run of the MCMC algorithm that generates a sequence of samples, so we need more chains to check the convergence and ensure that all chains convergent to same mode.

For 1000 burn-in draws, since MCMC starts from an initial value which often far from the true posterior region, so the chains need some steps to walk to the correct trace. As figure 2 shows, it

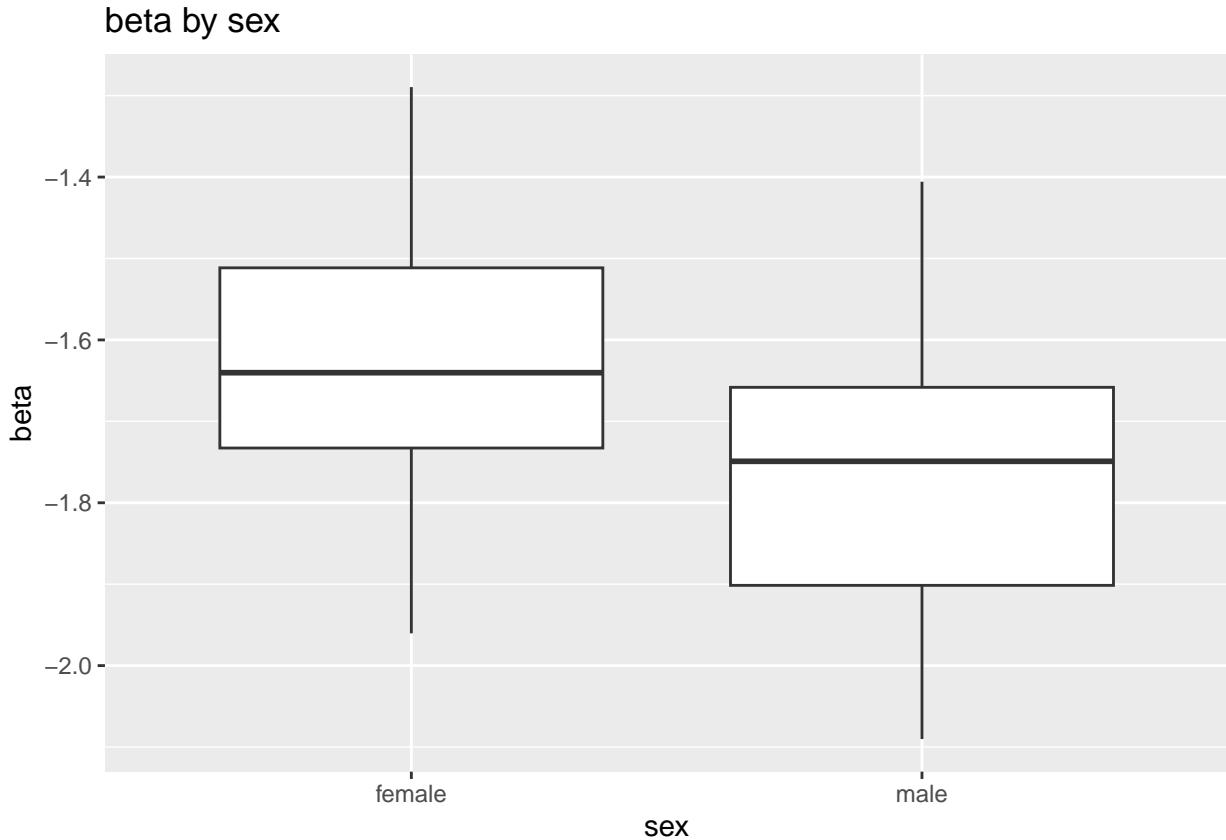
is an example of the MCMC process of ‘sexfemale’ when using 4 chains, 100 draws. Clearly that if we keep the first 25 draws, the results will be affected.



**Figure 2:** MCMC Simulations Traceplot.

### 3.3 Model Selection

#### 3.3.1 Key Variables



### Priors

Before fitting the Bayesian regression model, appropriate priors need to be specified for all regression coefficients and the residual standard deviation. We use weakly informative priors, which have small effects on the posterior. The selected priors are summarized as follows:

**Table 1:** Weakly informative priors used in the Bayesian regression model

| Parameter                         | Prior          |
|-----------------------------------|----------------|
| Regression coefficient for male   | Normal(0, 2)   |
| Regression coefficient for female | Normal(0, 2)   |
| Regression coefficients (others)  | Normal(0, 0.5) |
| Residual SD                       | Exponential(1) |

All variables have been standardized and we do not expect  $\log_{10}$  to have a linear relationship with the variables, so zero-mean normal priors are appropriate. Based on the review by Jones (2010), sex is

considered one of the main factors influencing alcohol elimination rates. Therefore, we set relatively wide priors  $\text{Normal}(0, 2)$  for the sex coefficients, allowing their effects to vary within a reasonable range. For other standardized variables (such as weight and height), since their impacts on  $\beta$  are considered relatively small, we used more shrinkage weakly informative priors  $\text{Normal}(0, 0.5)$  to prevent unreasonably large effects. The residual standard deviation is given an  $\text{Exponential}(1)$  prior, which is a commonly used positive weakly informative prior that can avoid unreasonably high noise.

If future research involves different populations, larger sample sizes, or additional biological information, the prior distributions can be adjusted accordingly to ensure they are applicable to any new datasets.

reference Jones, A. W. (2010). Evidence-based survey of the elimination rates of ethanol from blood with applications in forensic casework. *Forensic Science International*, 200, 1–20. <https://doi.org/10.1016/j.forsciint.2010.02.021> ## Model Results

### 3.3.2 Fitting result

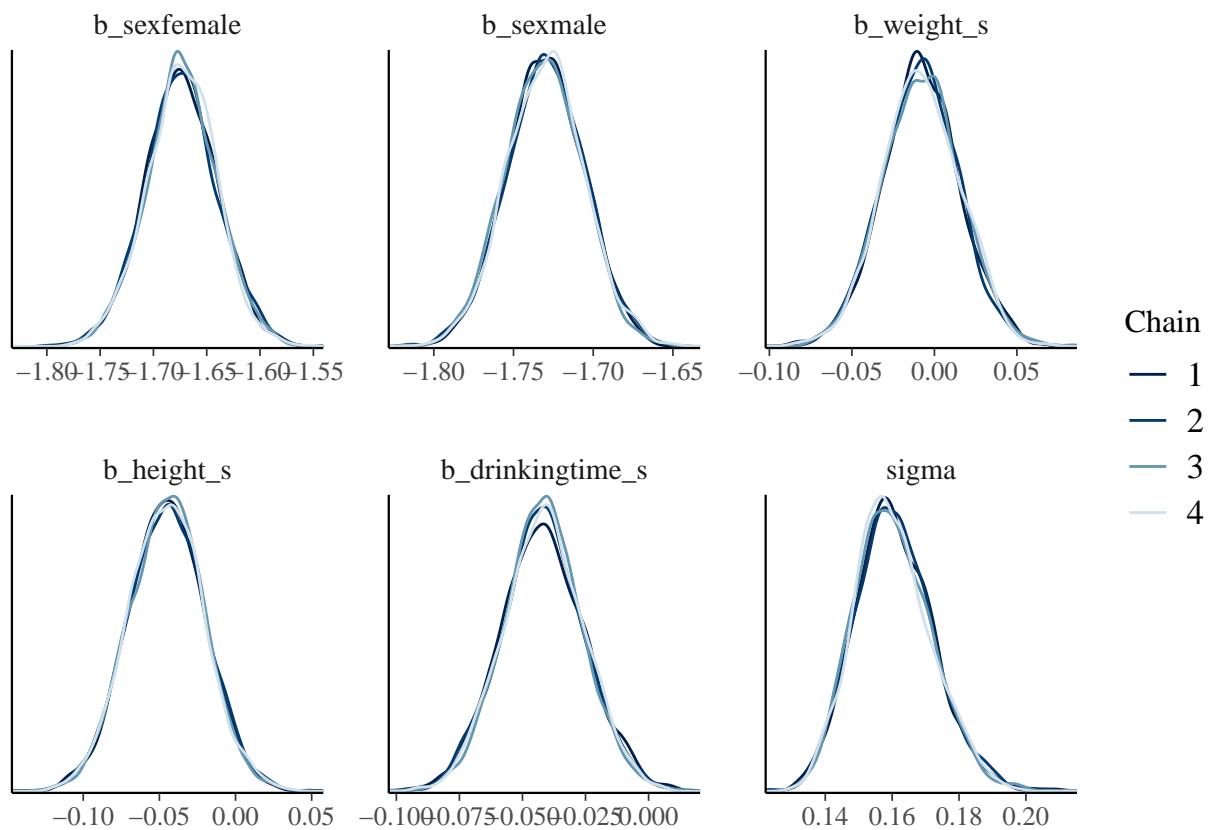
After 4000 iterations in 4 chains, figure 3 shows posterior results for each variables' coefficients, all coefficients show good normal patterns since we all chains have converged and warm-up length is enough. Since the  $\beta$  is log-transformed, the value of 'b\_sexfemale' and 'b\_sexmale' is not typical value of  $\beta$  like 0.19 when other coefficients are all near zero. Also  $\sigma$  value is estimated under log-transformed  $\beta$ .

Figure 4 shows the trace plots for all variable coefficients, it can be seen that all 4 chains mix well and all values are picking after convergent.

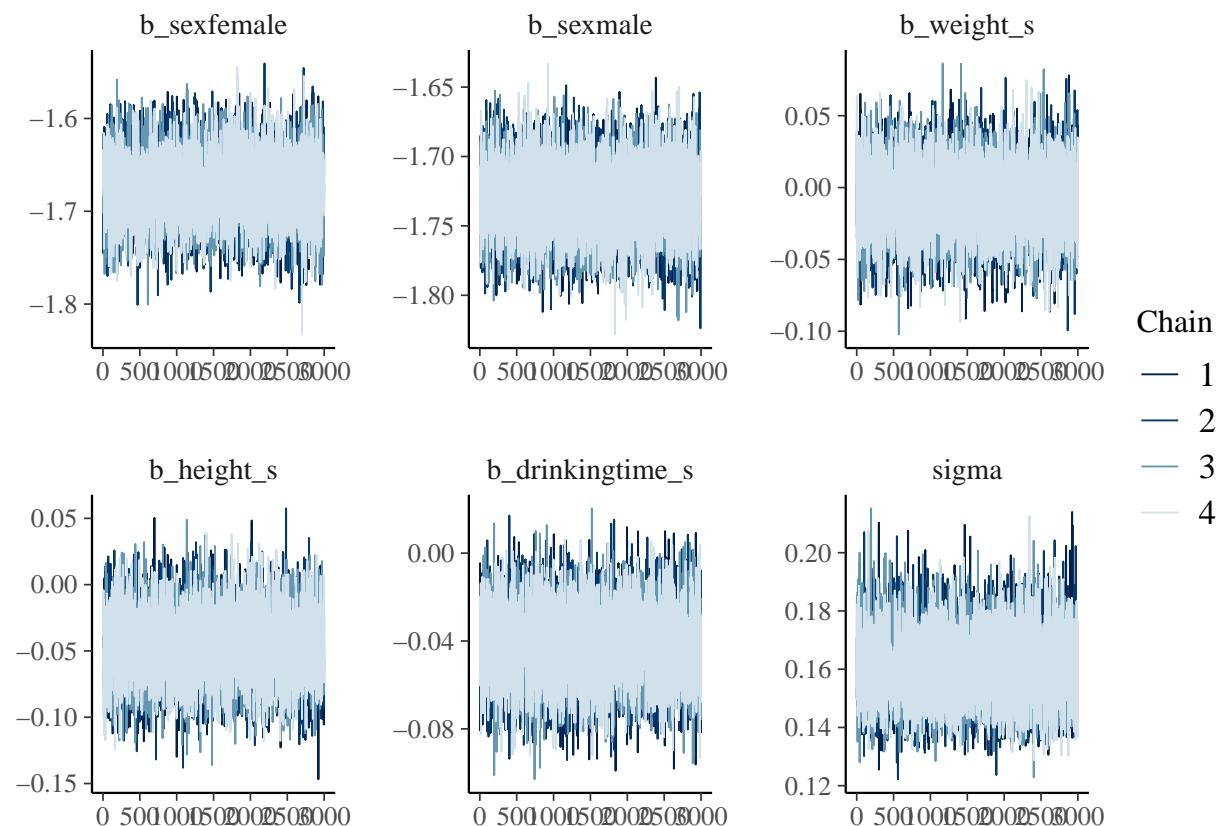
Table 2 gives the posterior summary. Rhat compares within-chain variance and between-chain variance, all less than 1.01 means chains are well-mixed.

**Table 2:** Posterior Summary from BRM

|                | Estimate | Est.Error | l-95% CI | u-95% CI | Rhat  | Bulk_ESS  | Tail_ESS |
|----------------|----------|-----------|----------|----------|-------|-----------|----------|
| sexfemale      | -1.674   | 0.034     | -1.741   | -1.606   | 1.000 | 9234.965  | 7914.372 |
| sexmale        | -1.730   | 0.025     | -1.779   | -1.682   | 1.001 | 8857.296  | 8083.851 |
| weight_s       | -0.008   | 0.024     | -0.055   | 0.039    | 1.001 | 9175.477  | 8027.627 |
| height_s       | -0.045   | 0.025     | -0.095   | 0.004    | 1.000 | 8165.013  | 8019.086 |
| drinkingtime_s | -0.042   | 0.016     | -0.074   | -0.009   | 1.001 | 11346.450 | 8084.544 |



**Figure 3:** PPC density of 500 predict results.

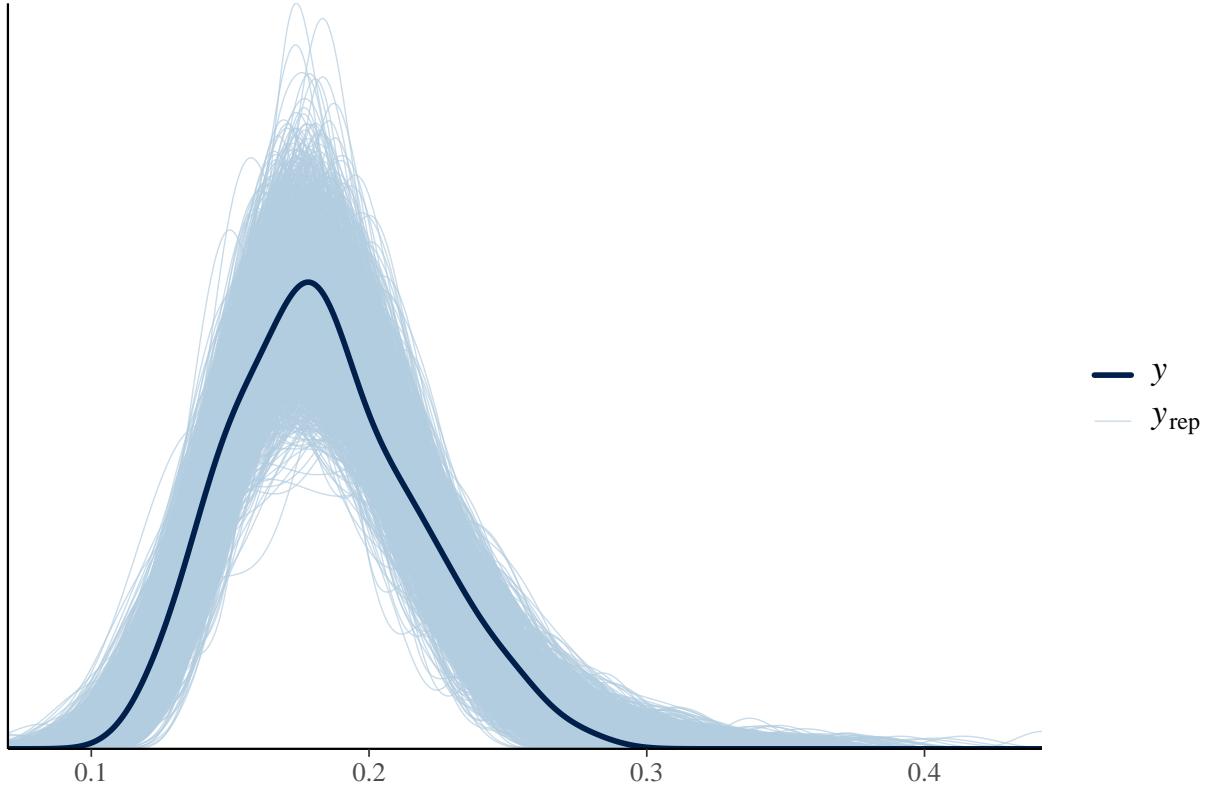


**Figure 4:** Traceplot of all coefficients and variance sigma.

### 3.3.3 PPC Density

Figure 5 is the Posterior Predictive Check (PPC) density plot comparing with true value of  $\beta$  (Here we have taken exponential of both true value and predicted value). Notice that predicted posterior distribution matches  $\beta$  sample distribution pretty well, which get the mean, spread, skew both right even at right tail.

PPC density



**Figure 5:** PPC density of 2000 predict results from posterior.

### 3.4 Model Testing

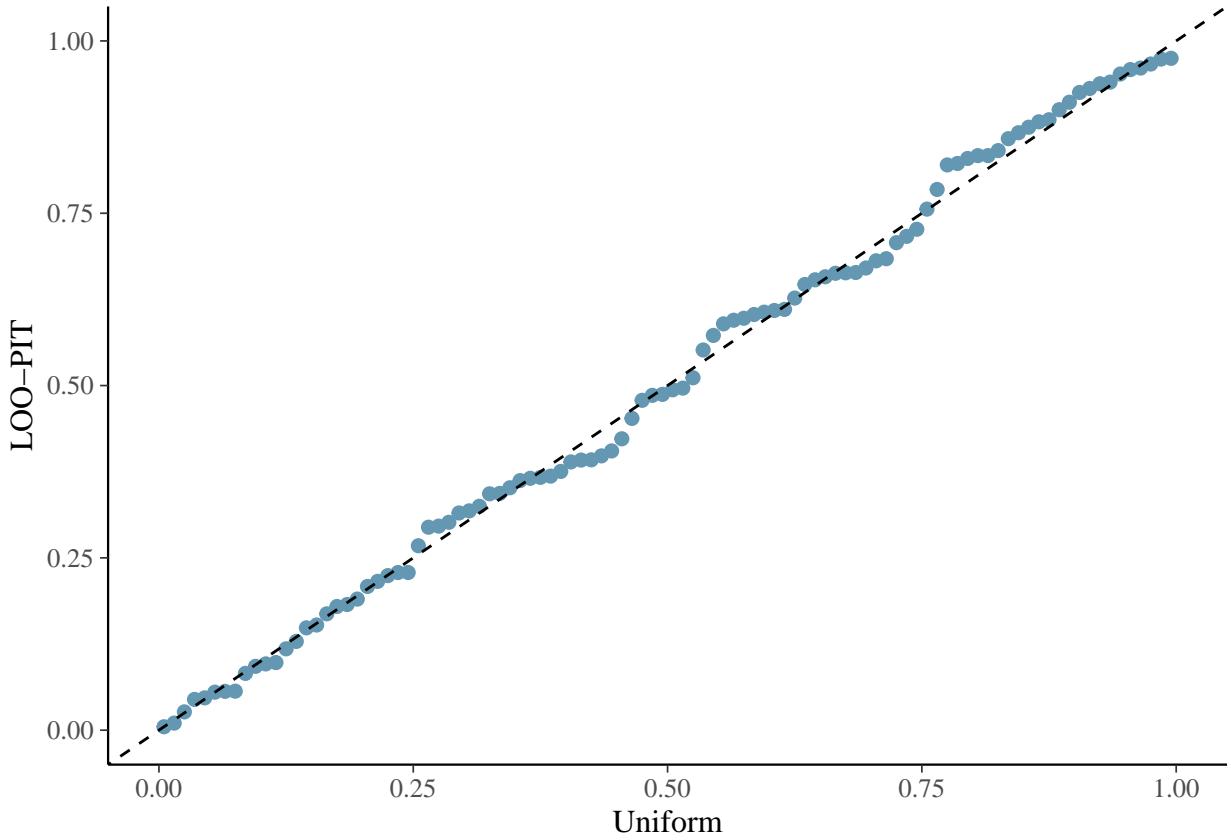
We take 12000 number of draws from the posterior distribution. #### LOO-PIT QQ plot The Leave-One-Out Probability Integral Transform Quantile-Quantile plot (LOO-PIT QQ plot) can check if the predictions calibrated correctly and if the predictive distributions biased, which is defined as:

$$\text{LOO-PIT}_i = P(\tilde{y}_i \leq y_i | y_{-i}).$$

Here we still use Monte Carlo method to simulate it, as:

$$\text{LOO-PIT}_i = \frac{1}{S} \sum_{s=1}^S \mathbf{1}(\hat{y}_{i,-i}^{(s)} \leq y_i)$$

where  $S$  is the number of posterior draws. We will use it to check if our posterior is under or over dispersion, the good model's PIT plot should be flat over  $\text{Uniform}(0, 1)$ . For our model, the PIT histogram (Figure 4) is flat everywhere except probability near 0.55 and 0.8, which are acceptable by the randomness of MCMC method. A U-shape (points over dot line at two tails) in PIT plots means overestimates variance, it doesn't appear in our plot gives evidence of our choices on variables and likelihood family.



### ### Coverage rate and Errors

Table 3 shows the coverage rate of 95% and 50% predicted intervals, which means 98 and 52 individuals' real  $\beta$  value is inside the intervals. The MAE and RMSE are both relatively small.

**Table 3:** Testing table

| Metric                           | Value |
|----------------------------------|-------|
| 95% predictive interval coverage | 0.980 |
| 50% predictive interval coverage | 0.510 |
| Mean Absolute Error              | 0.023 |

---

| Metric                 | Value |
|------------------------|-------|
| Root Mean Square Error | 0.029 |

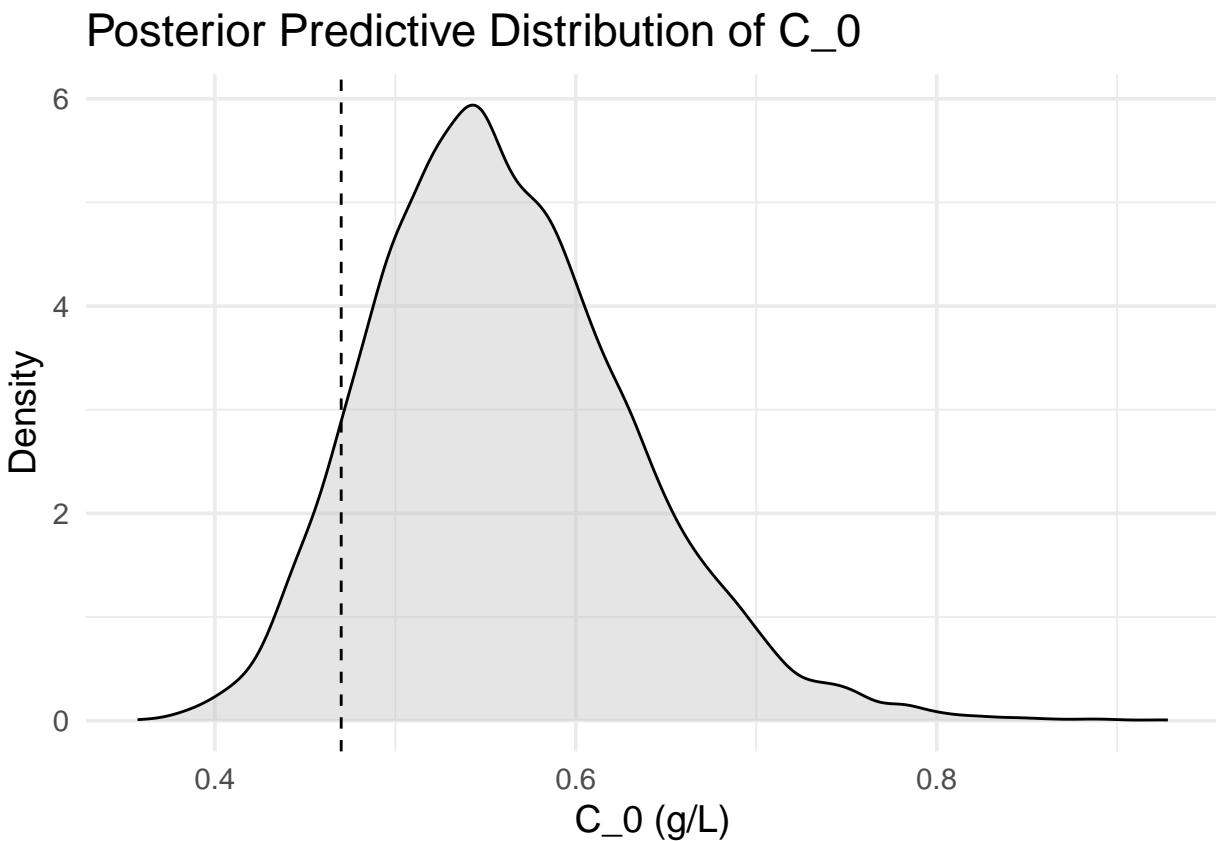
---

## 4 Example

Now we will apply our model on an individual example: A 70 year old female (weight: 70kg, height: 160cm) is arrested after being stopped by the police while driving. She provides a blood sample to the police 2 hours after her arrest which gives a reading of  $C_t = 0.15\text{g/kg}$ . The legal limit is  $x = 0.47\text{g/kg}$ .

Since there is no ‘drinking time’ data here, we will use a simplified model with only ‘sex’, ‘height’, ‘weight’ variables. We have standardized height and weight by using sample’s mean and variance.

As figure 7 shows, this is the posterior density of  $C_0$  calculated by presicted posterior density of  $\hat{\beta}$  the dot line is the legal limit  $C_0 = 0.47$ , it can be seen that most density is larger than 0.47, so obviously the 70 year old female is over the drink driving limit.



**Figure 6:** Posterior Predictive Distribution of  $C_0$ ,  $message=FALSE$ ,  $warning=FALSE$ , with limit  $C_0 = 0.47$  (dot line).

To make the result clear, it is better to give Table 4 to the courts and illustrate:

There is 91.4% probability that the 70 year old female is over the drink driving limit, with both mean, median and mode value over the limit.

The results also show that the original method is not reasonable since 2.5% percentile is over-conservative even in this case.

**Table 4:** Example results

| Statistic     | Value |
|---------------|-------|
| Mean          | 0.560 |
| Median        | 0.553 |
| Mode          | 0.542 |
| 2.5%          | 0.439 |
| 25%           | 0.510 |
| 75%           | 0.604 |
| 97.5%         | 0.715 |
| P(C_0 > 0.47) | 0.913 |

## 5 Widmark's Equation and $V_d$

## 6 Further research:

casual effect: biased data selection, (e.g. inner correlations between sex and height)

### 6.1 Variable Description

- how many observations do we have? where we get the resource, reference? explain each variables in table

The variables identified for the analysis are demographic and physiological characteristics from the tested individuals after drinking alcohol. Our variable selection is informed by their established relevance to human liver metabolic function, especially sex, age, weight and height are included due to the significant influence on liver metabolism.

- Sex: Biological sex of the individual, male or female
- Age:
- Weight
- Height
- Beta60:
- Co:

Page, Participant number, FigureOnPage, Sample Type... clearly irrelevant,

Maximum BAC/BrAC,BAC peak time      factor police

### 6.2 Preliminary analysis

- initial formula: explain beta60 + - motivation to improve?
- variable correlations? Why? hypothesis needed? why 2.5 quantiles?

## 7 Model selection regarding key variable ‘beta60’

### 7.1 Advantage of current method

simpler model with only one parameter(beta), easier to calculate, suitable for no dataset situations

#Limitation our model limit data collected reason - cannot reflect the real world model reason  
other reason: variables

#Reference CTP2E1 beta 1  
<https://www.sciencedirect.com/science/article/pii/S0379073810000770?via%3Dhub>

P.Y. Kwo, V.A. Ramchandani, S. O'Connor, D. Amann, L.G. Carr, K. Sandrasegaran, K.K. Kopecky, T.K. Li Gender differences in alcohol metabolism: relationship to liver volume and effect of adjusting for body mass Gastroenterology, 115 (1998), pp. 1552-1557