Report Title: Example Report

Group 3
Department / Course

November 16, 2025

Contents

1 Executive Summary **2**

2 Data Description **3**

3 Bayesian regression model (BRM) **4**

3.1 Reason for BRM 4

3.2 How BRM works 5

3.3 Model Selection 5

3.3.1 Key Variables 5

3.3.2 Priors 5

3.4 Model Results 6

3.4.1 Fitting result 6

3.4.2 PPC Density 6

3.5 Model Testing 6

3.5.1 LOO-PIT QQ plot 6

4 New method: build a linear regression model for beta **7**

5 Calculate V_d: **7**

6 Further research: **7**

6.1 REPORT 7

7 Exclusive summary **7**

8 Overview of Datasets **8**

8.1 Data collection 8

8.2 Variable Description 8

8.3 Prelimary analysis 8

9 Model selection regarding key variable 'beta60' **8**

9.1 Advantage of current method 8

9.2 New model motivation 9

10 Bayesian model **9**

10.1 Model fitting example 9

11 Testing another assumption(condition): **9**

11.1 variable explanation 9

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 β $(g/kg/h)$ is BAC elimination rate.

The key point is how to find a precise β to estimate C_0. Forensic scientists currently 2.5% percentile of β distribution constructing from samples as the estimated β 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 β 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 β.
- β 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 β by using both samples and prior knowledge. Then we randomly simulate 4000 β values from posterior distribution and find out the probability of the person's BAC over limit while driving.

In real world, β 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 β 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 β will be smaller since β is measured start at BAC peak time till the end.

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 β. But V_d and β 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 β 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 very important to normalized all variables like weight and height. Since β is a small value, the coefficients of large value variables are pretty small, near 0, which will make β estimation worse. The other important reason is centering the covariates can reduce autocorrelation

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

then the Effective Sample Size (ESS):

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

is lager, which means the MCMC method converge faster (for example if we set 4 chains with each chains 4000 iterations, 1000 burn-in, the $n_{\text{eff}} = 10000$ means the number of independent samples is 10000 out of 16000). Larger n_{eff} in same number of iterations means smaller variance of estimated exoectation since:

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

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

Since β 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 β 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.

3 Bayesian regression model (BRM)

3.1 Reason for BRM

Comparing with fitting linear regression model for β_i, which can be expressed as:

$$\hat{\beta}_i = \hat{\gamma}_1 \,\text{female}_i + \hat{\gamma}_2 \,\text{male}_i + \hat{\gamma}_3 \,\text{weight}_i + \hat{\gamma}_4 \,\text{height}_i + \sigma^2$$

and giving a exact value of $\hat{\beta}$, BRM can model uncertainty from each coefficient γ, since β from each individuals may affect by variables in different level. We provide priors and sample likelihoods for each variables, BRM will apply bayesian rule to each variables and output simulation results of coefficients specifically for each individual:

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

Since we have simulation results of $\hat{\beta}$, we can offer a probability for the courts.

3.2 How BRM works

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

The principle theorem behind BRM is Bayes' Rule:

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

The θ in the equation is the coefficients of γ_j. For most real models, we can't compute this posterior $p(\theta \mid \text{data})$ 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 γ_j, we need specify a family for β 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 β values good without restrict samples' shape.

Also for the MCMC simulation method, we set 4 chains, each chains 2000 draws with 500 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 500 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 1 showed, it is an example of the MCMC process of 'sexfemale' when using 4 chains, 100 draws. Clearly that if we keep the first 50 draws, the results will be affected.

3.3 Model Selection

3.3.1 Key Variables

3.3.2 Priors

Before fitting the Bayesian regression model, appropriate prior distributions need to be specified for all regression coefficients and the residual standard deviation. We use weakly informative priors so that the observed data dominate the inference process. The selected prior distributions are summarized as follows:

All variables have been standardised, so choosing Normal$(0, \cdot)$ as the weak information prior is reasonable. Based on the review by Jones (2010), sex is considered one of the primary factors influencing alcohol elimination rates. Therefore, we set a relatively wide prior Normal(0, 2) for the sex coefficients, allowing their effects to vary across a broad and reasonable range without being overly constrained by the prior. For other standardized variables (such as weight and height), since their impact on is considered relatively small, we used a more shrinkage weakly informative prior Normal(0, 0.5) to prevent unreasonably large effects. The residual standard deviation is given an 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 dataset.

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

3.4 Model Results

3.4.1 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 numbers is enough. Since the β is log-transformed, the value of 'b_sexfemale' and 'b_sexmale' is not typical value of β like 0.19 when other coefficients are all near zero.

3.4.2 PPC Density

Figure 3 is the Posterior Predictive Check (PPC) density plot comparing with true value of β (Here we have taken exponential of both true value and predicted value). Notice that predicted posterior distribution matches β sample distribution pretty well, which get the mean, spread, skew both right even at right tail.

3.5 Model Testing

3.5.1 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:

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

where S is the number of posterior draws (we set it 10000 here). We will use it to check if our posterior is under or over dispersion, the good model's PIT plot should be flat over $Uniform(0, 1)$. For our model, the PIT histogram (Figure 4) is flat everywhere except probability near 0.55, which is 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.

4 New method: build a linear regression model for beta

Check if age, weight and height have linear relationships with beta. Or have any patterns?

5 Calculate V_d:

in the Widmark's equation assumption, beta and V_d are independent, so we need to check whether Cov(V_d, beta) = 0

6 Further research:

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

6.1 REPORT

7 Exclusive summary

background Goal - 1, 2, 3 main data + elimination rate explanation method - model we used result

8 Overview of Datasets

8.1 Data collection

8.2 Variable Description

- how many observations do we have? where we get the resourse, 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 inrevelant,

Maximum BAC/BrAC,BAC peak time factor police

8.3 Prelimary analysis

- initial formula: explain beta60 + - motivation to improve?

variable correlations? Why? hypothesis needed? why 2.5 quantiles?

9 Model selection regarding key variable 'beta60'

9.1 Advantage of current method

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

9.2 New model motivation

Current model disadvantage: 1. beta is fixed for all individuals, we need to consider the heterogeneity in individuals. 2. Deviated aim: Police Scotland aims at assessing the probabilities that a person is over the limit, while the parameter beta only concentrates on the range between 2.5% and 97.5%.

10 Bayesian model

intro New model (Bayesian) ## Model Selection and Evaluation - why Bayesian? Any advantage, any improvement? More accurate? … ##explanation ###why prior setted? ## how result reflect?

10.1 Model fitting example

How our new model works on the 70 year-old madam?

11 Testing another assumption(condition):

11.1 variable explanation

explain all variables we used explain needed references

##Testing whether formula reasonable? whether beta60 relate to Vt? how related? hows the result? how new model fit?

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

#N

#Reference

CTP2E1 beta 1 60
https://www.sciencedirect.com/science/article/pii/S0379073810000770?via%3Dihub

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