Report Title: Example Report

Group 13
Department / Course

November 17, 2025

Contents

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 like 'gender' and 'drinking time'

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 like Table 4 by easily plugging new person's data.

2 Data Description

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

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

ρ_k measures how similar the sample at position i is to the sample at position $i+k$. As the random walk process, θ_i is similar to θ_{i+1}, less similar to θ_{i+2}, independent with θ_{i+k} for large k. So if ρ_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_{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_{eff} in same number of iterations means smaller variance of estimated exoectation since:

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

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

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 1, β 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 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.

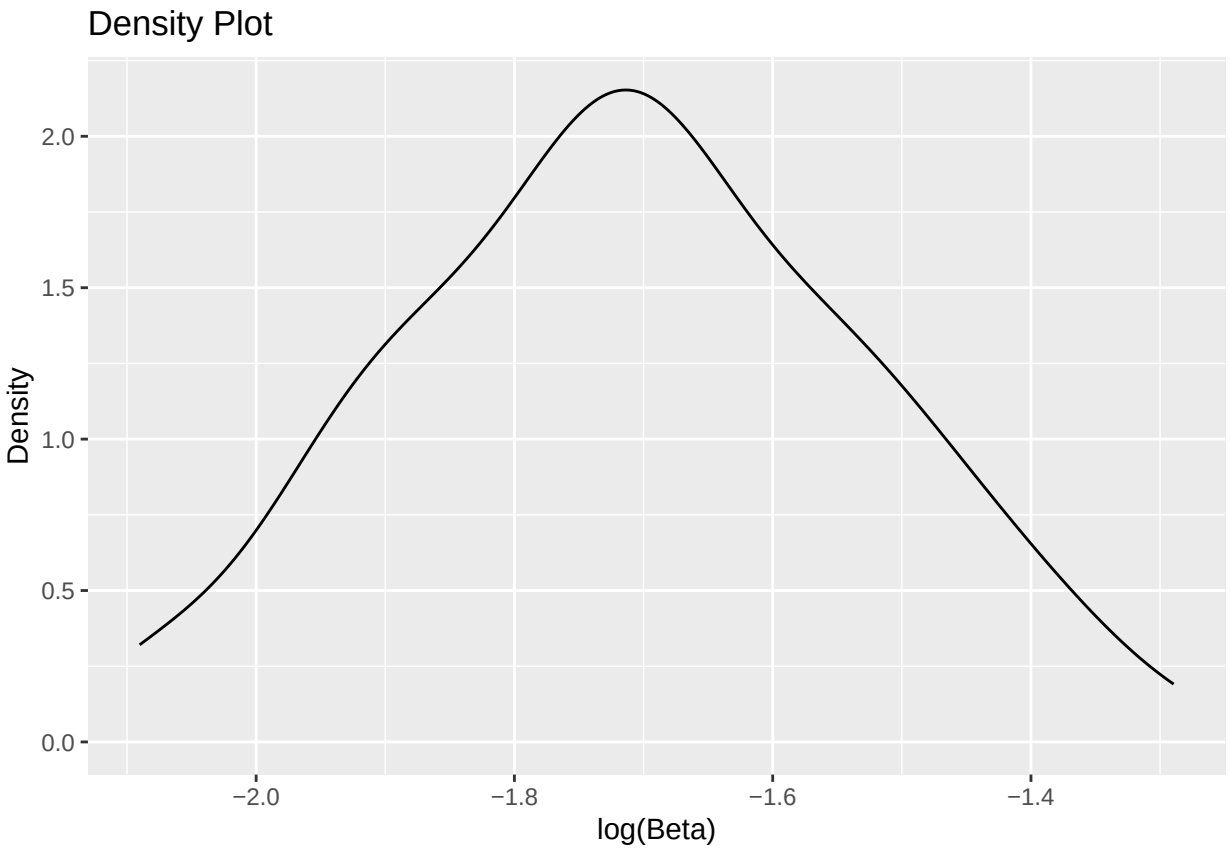


Figure 1: *log(beta) distribution.*

3 Bayesian regression model (BRM)

3.1 Reason for BRM

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

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

and giving a exact value of $\hat{\beta}_i$ for i_{th} estimation of $\hat{\beta}$ under σ_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:

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

where

$$\varepsilon_{\beta,i} \sim \mathcal{N}(0,\, \sigma_\beta^2).$$

Since we have log-transformed β data points, we will take exponential of $\widehat{\beta_{i,j}}$ 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:

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

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 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.

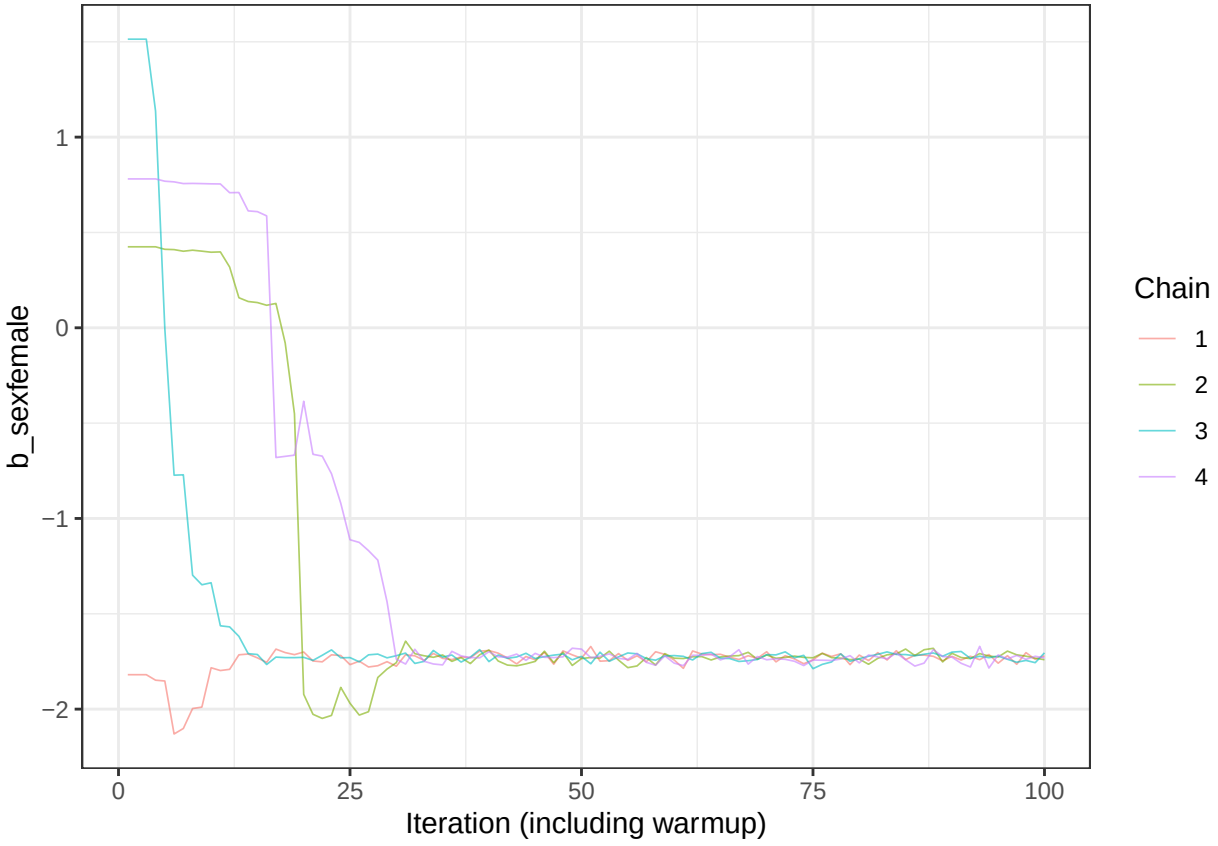
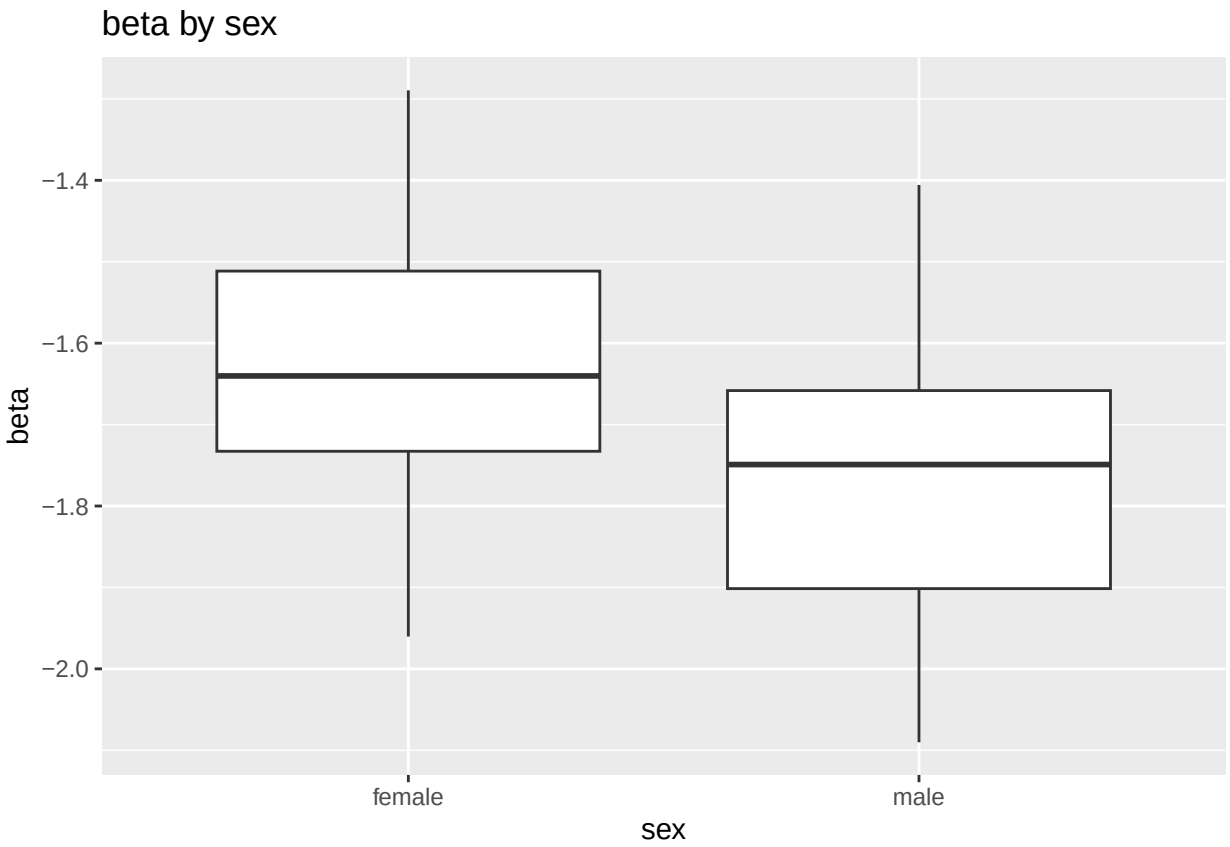


Figure 2: *MCMC Simulations Traceplot.*

3.3 Model Selection

3.3.1 Key Variables

- 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 starting at BAC peak time till the end, there will be fluctuations on the BAC plots.



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 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 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 are considered relatively small, we used more shrinkage weakly informative priors 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 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, 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.

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 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. Also σ value is estimated under log-transformed β.

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: Posterior Summary from BRM

	Estimate	Est.Error	l-95% CI	u-95% CI	Rhat	Bulk_ESS	Tail_ESS
sexfemale	-1.673	0.034	-1.739	-1.606	1	8677.139	8065.652
sexmale	-1.730	0.025	-1.778	-1.682	1	8090.728	8240.799
weight_s	-0.008	0.024	-0.054	0.038	1	9207.023	8348.131
height_s	-0.045	0.025	-0.094	0.004	1	8268.992	8280.094
drinkingtime_s	-0.041	0.017	-0.073	-0.008	1	11655.316	8532.767

3.3.3 PPC Density

Figure 5 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.

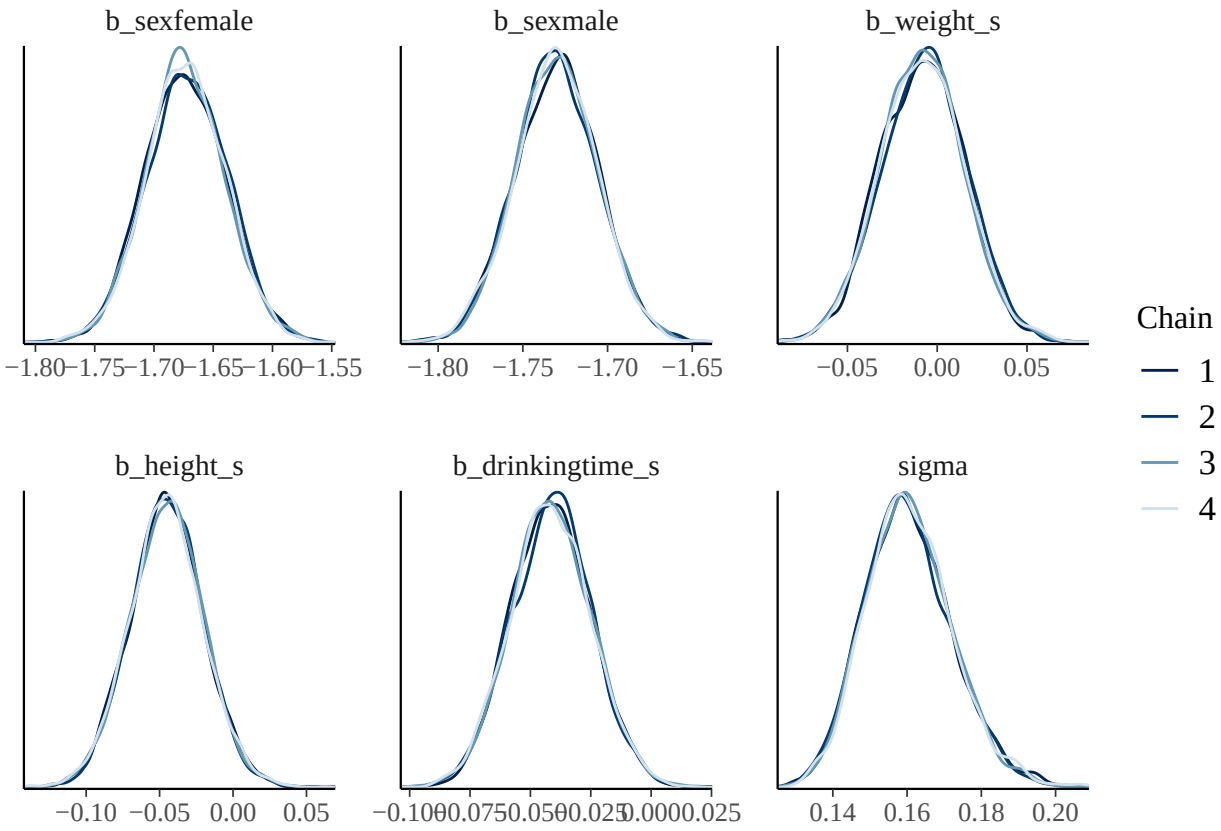


Figure 3: *PPC density of 500 predict results.*

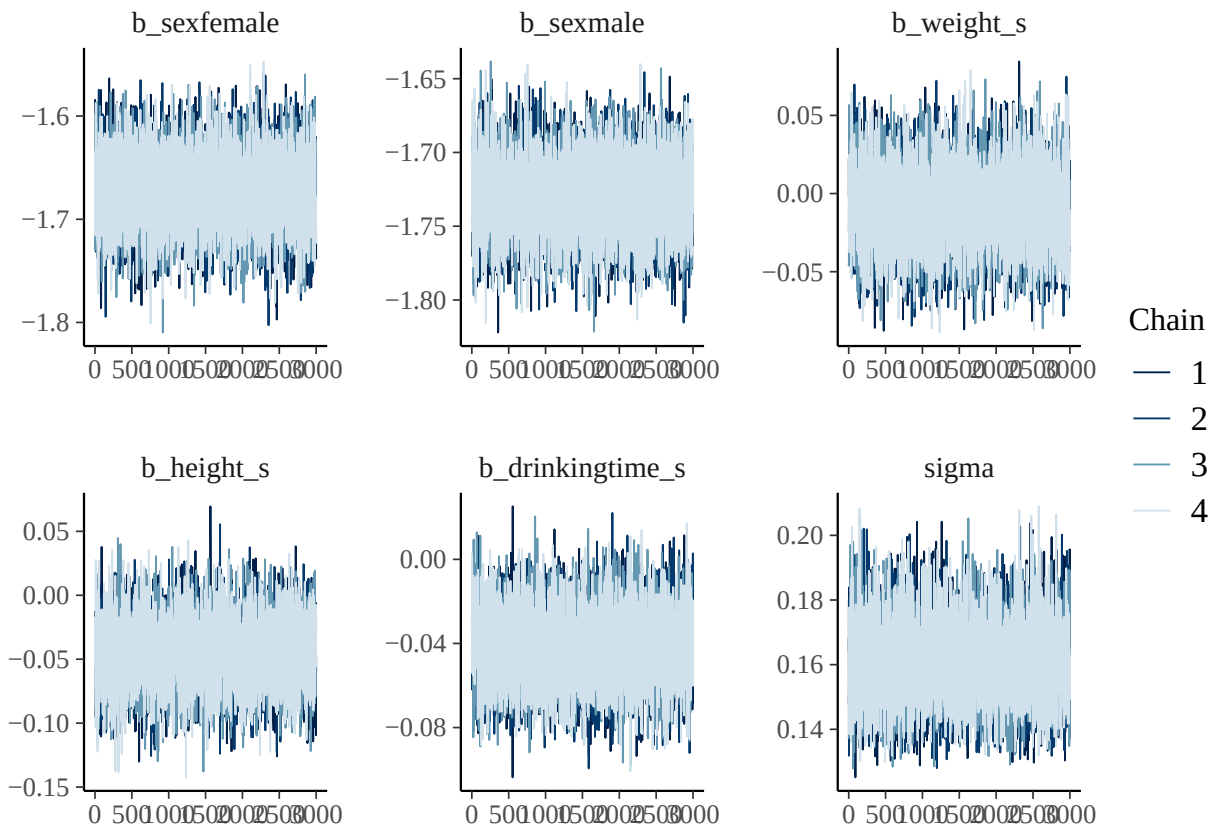


Figure 4: *Traceplot of all coefficients and variance sigma.*

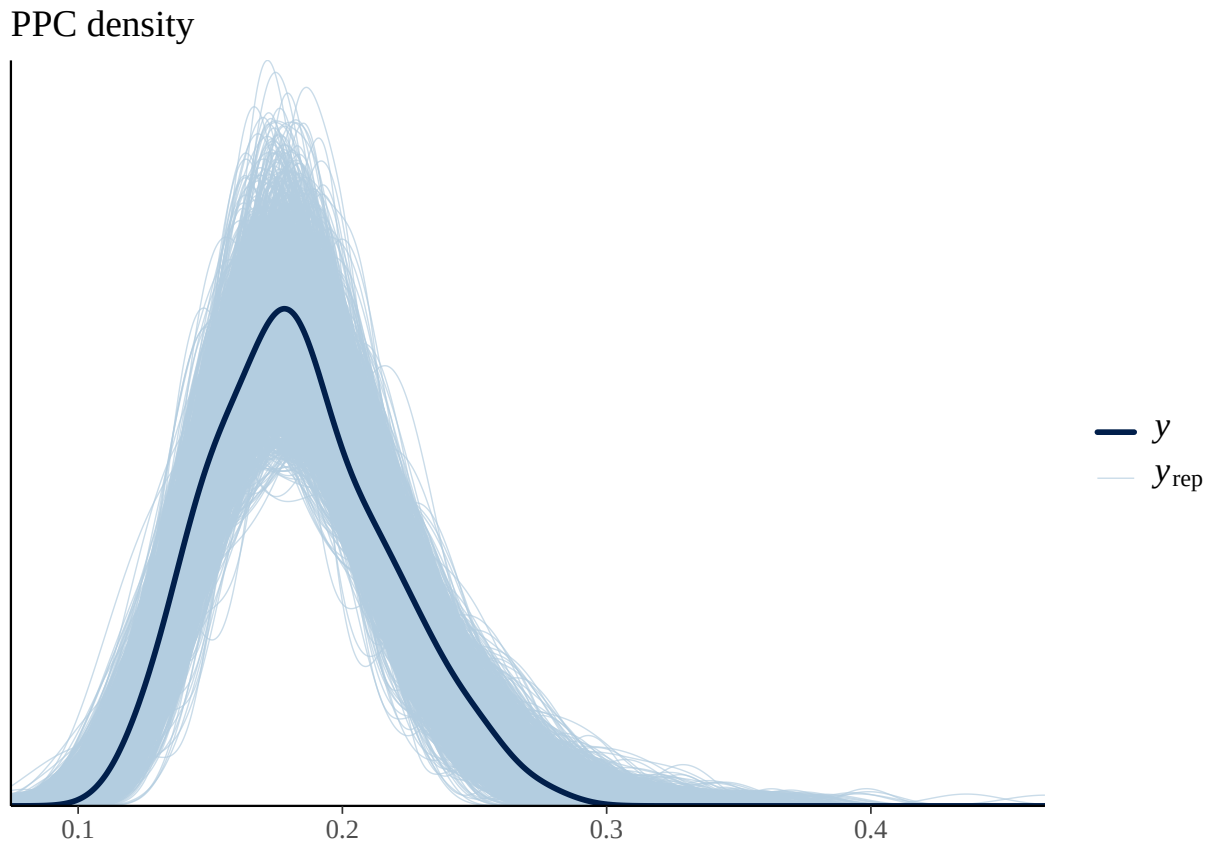


Figure 5: *PPC density of 2000 predict results from posterior.*

3.4 Model Testing

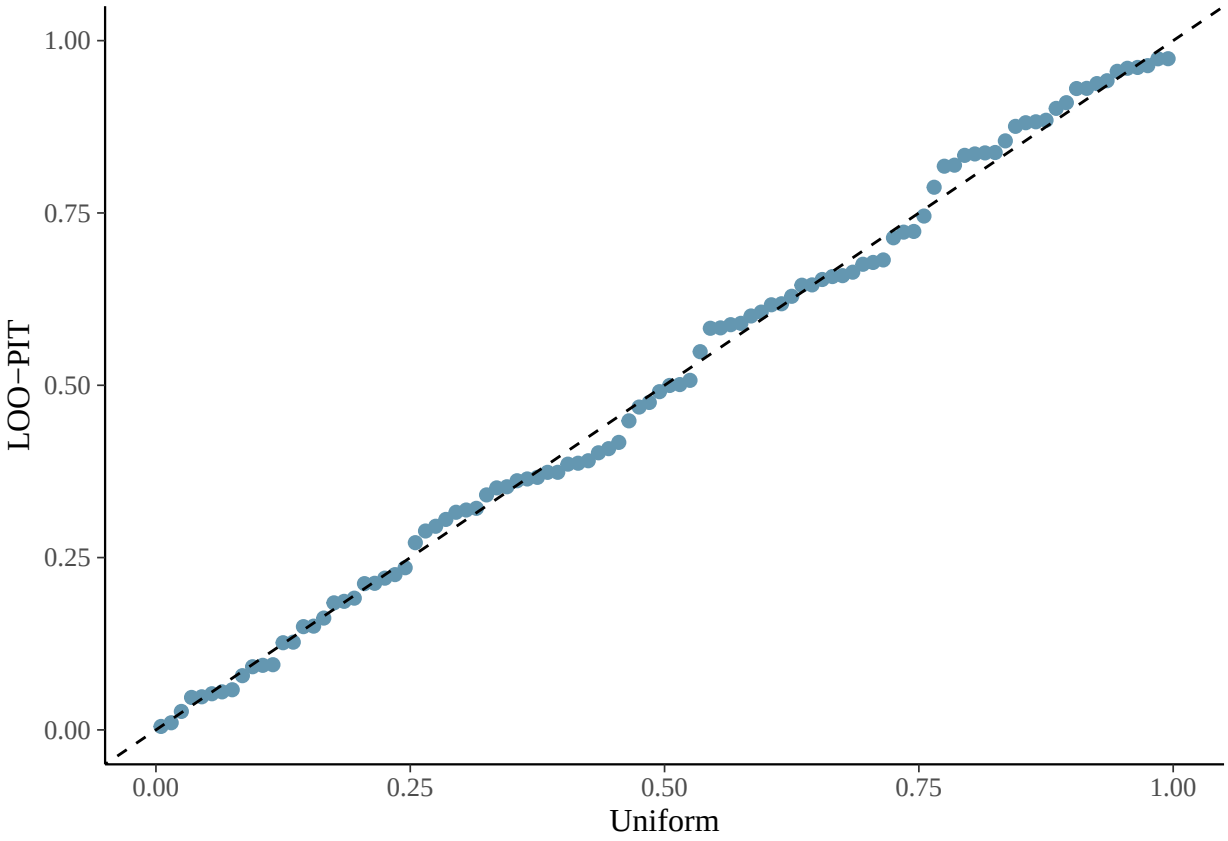
We take 12000 number of draws from the posterior distribution. ### LOO-PIT QQ plot The Leave-One-Out Probability Integral Transform Quantile-Quantile plot (LOO-PIT QQ plot) can check if the predictions calibrated correctly and if the predictive distributions biased, which is defined as:

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

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

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

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 $Uniform(0, 1)$. 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 β 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
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 Ct = 0.15g/kg. The legal limit is x = 0.47g/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.

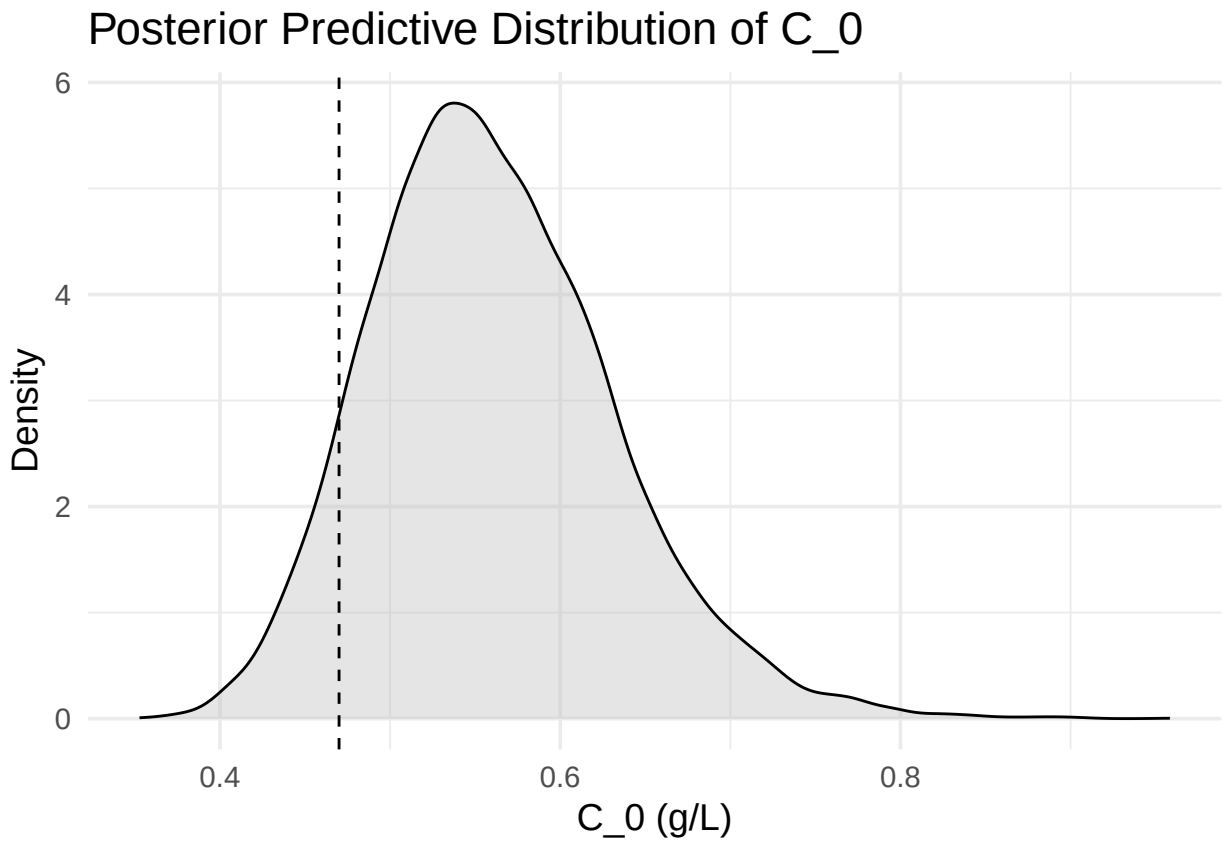


Figure 6: *Posterior Predictive Distribution of C_0, message=FALSE, warning=FALSE, with limit C_0 = 0.47 (dot line).*

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	0.554
Mode	0.538
2.5%	0.437
25%	0.510
75%	0.604
97.5%	0.716
P(C_0 > 0.47)	0.913

5 Widmark's Equation and V_d

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:

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

Since there are many versions of Widmark's equation, it can also write as:

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

where

$$\rho = \frac{TBW}{weight \times F_{\text{water}}}$$

is known as Widmark's rho factor.

F is the fraction of the dose that reaches the systemic circulation, F_{water} is know as the water content of the blood sample, TBW is Total Body Water.

We assume that all different part between two equations are included in V_d. The unit of C_t is g/kg in our dataset, the unit of the first term is g/L, but it is not a problem since the density of water is 1, we just ignore the 1.

5.1 V_d, TBW and ρ

All information in this part are from 'Total body water is the preferred method to use in forensic blood-alcohol calculations rather than ethanol's volume of distribution'[5].

TBW is calculated by:

$$\text{TBW}_{\text{Men}}(\text{L}) = 2.447 - (0.09516 \times \text{age}) + (0.1074 \times \text{height}) + (0.3362 \times \text{weight})$$

$$\text{TBW}_{\text{Women}}(\text{L}) = -2.097 + (0.1069 \times \text{height}) + (0.2466 \times \text{weight}).$$

Now subbing the ρ expression in the C_t equation, and comparing with our equation, V_d is actually calculated by:

$$V_d = \frac{\text{TBW}}{\text{weight}} \times \frac{1}{F_{\text{water}}} \times \frac{1}{F} = \eta \times \frac{\text{TBW}}{\text{weight}}$$

and we need to estimate η value since TBW is a value that can be determined in our dataset. we

will use:

$$\hat{V}_{d_{i,j}} = \hat{\eta}_{1,i}(\frac{\text{TBW}}{\text{weight}}_{\text{male}})_j + \hat{\eta}_{2.i}(\frac{\text{TBW}}{\text{weight}}_{\text{female}})_j + \varepsilon_i$$

to estimate $\hat{V}_d$ where i is the i_{th} iterations (we will again use BRM, but as a joint model with β to deal with the correlation, we will explain it in next part.), j is the j_{th} individuals, η is the coefficients that need to be predicted.

5.2 Joint BRM for β and V_d

V_d is calculated by true value β (with our log-transformed). We will not log-transformed V_d since the density of V_d can't be improved by taking log.

Figure 7: *Density plot of V_d*

5.3 Correlations between β and V_d

To show that the original method is not good because of the correlation can't be ignored, figure 9 gives the correlations between true β and V_d is -0.415, which is relatively big.

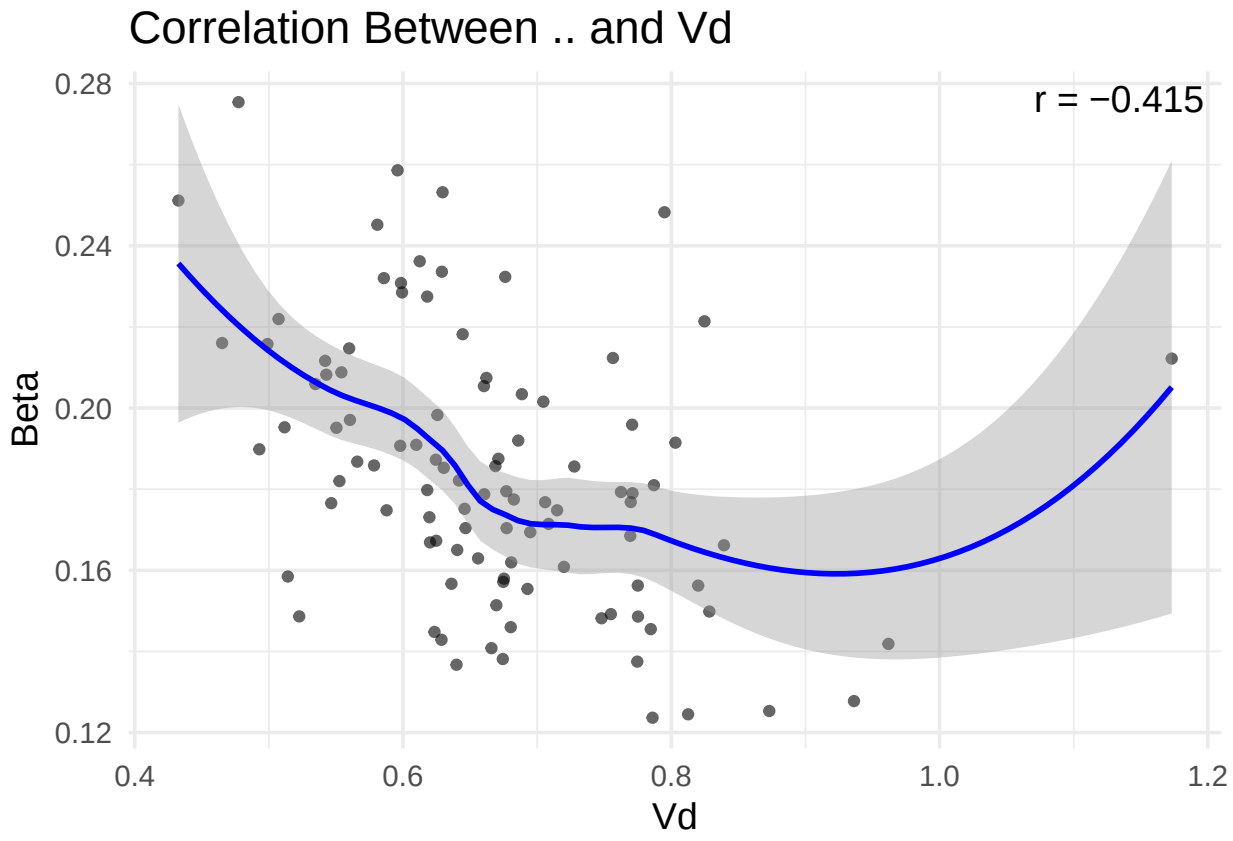


Figure 8: *Correlation plot betweem V_d and Beta*

5.4 Fitting model and Results

The mathematical expression of the joint model can be expressed as:

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

and

$$\hat{V}_{d_{i,j}} = \hat{\eta}_{1,i}(\frac{\text{TBW}}{\text{weight}}_{\text{male}})_j + \hat{\eta}_{2.i}(\frac{\text{TBW}}{\text{weight}}_{\text{female}})_j + \varepsilon_{i,V_d}$$

where

$$\begin{pmatrix} \varepsilon_{\beta,i} \\ \varepsilon_{V_d,i} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2_\beta & \rho\,\sigma_\beta\sigma_{V_d} \\ \rho\,\sigma_\beta\sigma_{V_d} & \sigma^2_{V_d} \end{pmatrix} \right).$$

5.4.1 Model Selection (priors)

The prior for β's variables' coefficients is unchanged. And the prior for η is easy to calculated since $\eta = \frac{1}{F_{\text{water}}} \times \frac{1}{F}$.

The Blood water content F_{water} is easy to determine by desiccation and the mean reported values for men and women are 0.825% w/v and 0.838% w/v respectively. The small sex difference is mainly attributed to lower haematocrit in female blood samples[].

The fraction of the alcohol dose that reaches the systemic circulation F is typically 0.7−0.9. So:

- Prior for $\frac{\text{TBW}}{\text{weight}}_{\text{male}}$ is $Normal(\frac{1}{0.825} \times \frac{1}{0.8}, 2) = Normal(1.51, 2)$
- Prior for $\frac{\text{TBW}}{\text{weight}}_{\text{male}}$ is $Normal(\frac{1}{0.848} \times \frac{1}{0.8}, 2) = Normal(1.49, 2)$

The mean value of two priors is actually near to the linear regression results of single V_d model (1.41 for male and 1.3 for female).

Priors for others are still weak informative priors. Setting for others are not changed (4 chains, 4000 iterations, 1000 warm-up).

5.4.2 Results

Table 5 belew is the results table, all Rhat and ESS are in good range. The expected estimated correlation −0.46 in the model is also near the sample's V_d-β correlation −0.415 (the log-transformation of β doesn't affect the correlation.)

Table 5: Joint model results

Parameter	Estimate	Est.Error	Rhat	Bulk_ESS	Tail_ESS	Group
beta_sexfemale	-1.6663	0.0330	1.0002	7569.217	7719.132	Beta
beta_sexmale	-1.7340	0.0239	1.0003	7828.005	7590.979	Beta
beta_weight_s	-0.0007	0.0210	1.0004	7825.627	7441.216	Beta
beta_height_s	-0.0467	0.0220	1.0001	7393.980	6788.047	Beta
beta_drinkingtime_s	-0.0310	0.0149	1.0001	10080.999	6941.147	Beta
Vd_T_Vd:sexfemale	1.2966	0.0358	1.0002	8797.854	7030.693	Vd
Vd_T_Vd:sexmale	1.4093	0.0300	1.0000	9356.247	7496.418	Vd
sigma_beta	0.1609	0.0117	1.0002	8337.962	7107.995	Beta
sigma_Vd	0.1140	0.0083	1.0003	10044.474	7045.928	Vd
rescor(beta,Vd)	-0.4623	0.0795	1.0004	8932.569	6864.751	Correlation

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 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

#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%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