

Heart Disease Prediction

A Bayesian Data Analysis Approach

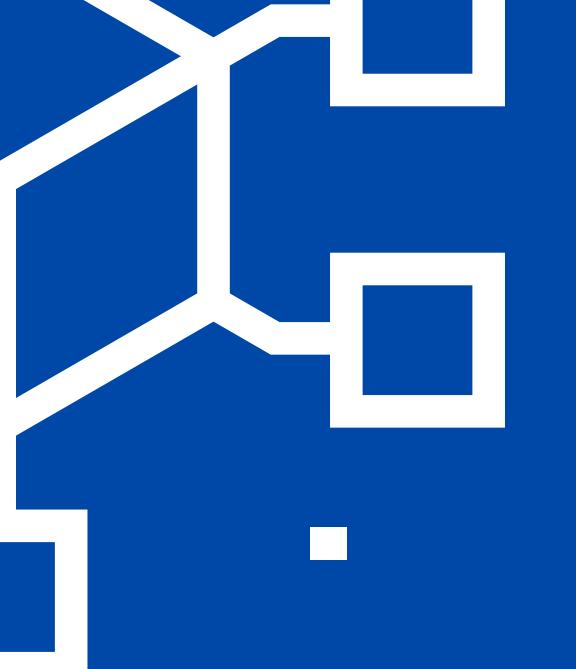
Kelompok 1:

Angeli Jessica Wibowo 2702225493

Felicia Audrey Tanujaya 2702217610

Leony Sani Winata 2702224811

Liliana Djaja Witama 2702219774



The Dataset

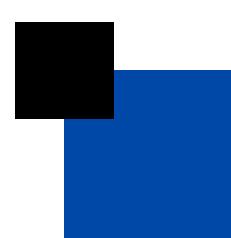
Heart Disease Prediction

Objective: Explore how covariates (*age, cholesterol, etc.*) affect heart disease and compare two predictive models.

Covariates: **303 rows** of age, sex, chest pain type (cp), resting blood pressure (trestbps), cholesterol (chol), fasting blood sugar (fbs), resting electrocardiographic results (restecg), maximum heart rate achieved (thalach), exercise-induced angina (exang), and ST depression induced by exercise relative to rest (oldpeak).

Why heart disease? Cardiovascular diseases (CVDs) are the leading cause of mortality, accounting for about one-third of all global deaths. Machine learning and statistical modeling can enhance prediction accuracy.

Source: **Heart Disease Dataset** from *kaggle* ([link](#))



The Dataset

Heart Disease Prediction

Covariates:

The names and social security numbers of the patients were recently removed from the database, replaced with dummy values.

Feature	Description	Example Values
Age	Age of the patient	12, 34
Sex	Gender of the patient (1 = Female, 0 = Male)	1, 0
Resting Blood Pressure (trestbps)	Resting blood pressure in mm Hg	200, 300
Fasting Blood Sugar (fbs)	Fasting blood sugar > 120 mg/dL (1 = True, 0 = False)	0, 1
Cholesterol (chol)	Serum cholesterol in mg/dL	200, 300
Maximum Heart Rate Achieved (thalach)	Maximum heart rate achieved during exercise	150, 170

The Dataset

Heart Disease Prediction

Covariates:

The names and social security numbers of the patients were recently removed from the database, replaced with dummy values.

Feature	Description	Example Values
Exercise-Induced Angina (exang)	Presence of exercise-induced chest pain (1 = Yes, 0 = No)	0, 1
ST Depression Induced by Exercise (oldpeak)	Abnormal finding on electrical activity of heart induced by exercise	1, 0
Slope of Peak Exercise ST Segment (slope)	Slope of the peak exercise ST segment	0, 1, 2
Number of Major Vessels (ca)	Number of major vessels (0-3) colored by fluoroscopy	0, 1, 2, 3
Thalassemia (thal)	Blood disorder status (e.g., normal, fixed defect, reversible defect)	0, 1, 2

The Dataset

Heart Disease Prediction

Covariates:

The names and social security numbers of the patients were recently removed from the database, replaced with dummy values.

Feature	Description	Example Values
Chest Pain Type (cp)	Types of chest pain experienced (0 = typical angina, 1 = atypical angina, 2 = non-anginal pain, 3 = asymptomatic)	0, 1, 2, 3
Resting ECG Results (restecg)	Resting electrocardiographic results (0 = normal, 1 = having ST-T wave abnormality, 2 = showing)	0, 1, 2

Model 1

Uninformative Prior

Bayesian **Logistic Regression** using JAGS

```
```{r}
mod <- textConnection("model{
 for(i in 1:n){
 Y[i] ~ dbern(pi[i])
 logit(pi[i]) <- beta[1] +
 X[i,1]*beta[2] + X[i,2]*beta[3] + X[i,3]*beta[4] +
 X[i,4]*beta[5] + X[i,5]*beta[6] + X[i,6]*beta[7] +
 X[i,7]*beta[8] + X[i,8]*beta[9] + X[i,9]*beta[10] +
 X[i,10]*beta[11] + X[i,11]*beta[12] + X[i,12]*beta[13] +
 X[i,13]*beta[14]
 }
 for(j in 1:14){beta[j] ~ dnorm(0,0.01)}
}")
```
```

The Model

Model 1

Uninformative Prior

Bayesian **Logistic Regression** using JAGS

```
Iterations = 2001:7000
Thinning interval = 1
Number of chains = 2
Sample size per chain = 5000
```

1. Empirical mean and standard deviation for each variable,
plus standard error of the mean:

| | Mean | SD | Naive SE | Time-series SE |
|----------|----------|--------|----------|----------------|
| beta[1] | 0.11897 | 0.1865 | 0.001865 | 0.002615 |
| beta[2] | -0.04092 | 0.2171 | 0.002171 | 0.003771 |
| beta[3] | -0.88276 | 0.2292 | 0.002292 | 0.003923 |
| beta[4] | 0.97373 | 0.1979 | 0.001979 | 0.003192 |
| beta[5] | -0.37628 | 0.1887 | 0.001887 | 0.002669 |
| beta[6] | -0.25756 | 0.2077 | 0.002077 | 0.003313 |
| beta[7] | 0.01013 | 0.1947 | 0.001947 | 0.002849 |
| beta[8] | 0.26457 | 0.1903 | 0.001903 | 0.002607 |
| beta[9] | 0.58088 | 0.2474 | 0.002474 | 0.004228 |
| beta[10] | -0.48864 | 0.1996 | 0.001996 | 0.002806 |
| beta[11] | -0.68588 | 0.2536 | 0.002536 | 0.004466 |
| beta[12] | 0.37947 | 0.2246 | 0.002246 | 0.003895 |
| beta[13] | -0.85105 | 0.2057 | 0.002057 | 0.002985 |
| beta[14] | -0.59446 | 0.1844 | 0.001844 | 0.002629 |

2. Quantiles for each variable:

| | 2.5% | 25% | 50% | 75% | 97.5% |
|----------|----------|-----------|-----------|---------|----------|
| beta[1] | -0.24739 | -0.006632 | 0.117424 | 0.2450 | 0.48680 |
| beta[2] | -0.46211 | -0.186207 | -0.038351 | 0.1037 | 0.38349 |
| beta[3] | -1.35458 | -1.029638 | -0.878027 | -0.7256 | -0.44939 |
| beta[4] | 0.59328 | 0.838860 | 0.970110 | 1.1066 | 1.37167 |
| beta[5] | -0.75026 | -0.501821 | -0.375372 | -0.2483 | -0.01545 |
| beta[6] | -0.66344 | -0.398062 | -0.259426 | -0.1198 | 0.15110 |
| beta[7] | -0.36172 | -0.121448 | 0.007229 | 0.1383 | 0.40181 |
| beta[8] | -0.10983 | 0.137567 | 0.263602 | 0.3918 | 0.63981 |
| beta[9] | 0.10158 | 0.413678 | 0.579856 | 0.7467 | 1.07711 |
| beta[10] | -0.88373 | -0.621252 | -0.486846 | -0.3551 | -0.09972 |
| beta[11] | -1.21597 | -0.851843 | -0.676774 | -0.5136 | -0.20615 |
| beta[12] | -0.05955 | 0.229351 | 0.376429 | 0.5332 | 0.82142 |
| beta[13] | -1.26503 | -0.988614 | -0.849130 | -0.7105 | -0.45549 |
| beta[14] | -0.95966 | -0.716513 | -0.591573 | -0.4702 | -0.24046 |

The Result

Model 2

Informative Prior

The Features

We utilized prior knowledge from published literature and clinical data to inform priors for several variables in the Heart Disease Prediction dataset.

Age:

The average log odds for age across genders in developing coronary heart disease was calculated as -0.6405, with confidence intervals of (-0.786, -0.514). (*Lloyd-Jones et al., 1999, The Lancet, 353(9147), 89-92*)

Prior Distribution: $\beta_{age} \sim dnorm(-0.6405, \tau_{age})$

where precision (τ_{age}) is derived from the variance of the confidence interval:

$$\sigma_{age} = \frac{(0.786 - 0.514)}{2 \times 1.96}$$

$$\tau_{age} = \frac{1}{\sigma^2_{age}}$$

Cholesterol (chol):

The adjusted odds ratio (95% confidence interval) of the observed serum levels of cholesterol was 1.66 (1.29-2.15) with a reference serum cholesterol <167mg/dl. (*Wakugami, K et al., 1998, Japanese circulation journal, 62(1), 7-14*)

$$\log(OR) = \ln(1.66) = 0.507$$

$$\text{Confidence interval} = \ln(1.29) = 0.255, \ln(2.15) = 0.765$$

Prior Distribution: $\beta_{chol} \sim dnorm(-0.507, \tau_{chol})$

Where

$$\sigma_{chol} = \frac{(0.765 - 0.255)}{2 \times 1.96}$$

$$\tau_{chol} = \frac{1}{\sigma^2_{chol}}$$

Model 2

Informative Prior

The Features

We utilized prior knowledge from published literature and clinical data to inform priors for several variables in the Heart Disease Prediction dataset.

Fasting Blood Sugar (FBS):

Based on the China-PAR project done by Tong et all., the persistency of FBS for a cardiovascular risk is 1.594 of 1.003 to 2.532 with a 95% confidence interval

Prior Distribution: $\beta_{\text{sex}} \sim \text{dnorm}(0.466, \tau_{\text{fbs}})$

where precision (τ_{fbs}) is derived from the variance of the confidence interval:

$$\sigma_{\text{fbs}} = \frac{(0.929 - 0.003)}{2 \times 1.96}$$

$$\tau_{\text{fbs}} = \frac{1}{\sigma^2_{\text{fbs}}}$$

Blood Pressure (trestbps):

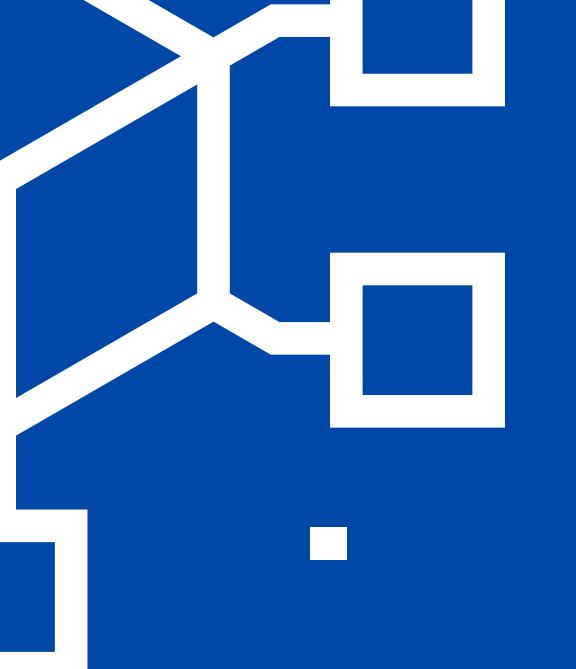
Based on a study done by Zhang H et all., 2020, the relationship between blood pressure and cardiovascular-related events is 23.8% with a 95% confidence interval (17.9% to 28.8%)

Prior Distribution: $\beta_{\text{sex}} \sim \text{dnorm}(-1.242, \tau_{\text{trestbps}})$

where precision (τ_{trestbps}) is derived from the variance of the confidence interval:

$$\sigma_{\text{trestbps}} = \frac{(3.24 - 2.61)}{2 \times 1.96}$$

$$\tau_{\text{trestbps}} = \frac{1}{\sigma^2_{\text{age}}}$$



Model 2

Informative Prior

The Features

We utilized prior knowledge from published literature and clinical data to inform priors for several variables in the Heart Disease Prediction dataset.

Sex:

Based on the study done by Leifheit-Limson et all., 2015 that was published on 'Journal of the American College of Cardiology', Women were less likely to be told at-risk with the relative risk of 0,89-0,96 with a 95% confidence interval.

Prior Distribution: $\beta_{\text{sex}} \sim \text{dnorm}(-0.116, \tau_{\text{sex}})$

where precision (τ_{age}) is derived from the variance of the confidence interval:

$$\sigma_{\text{sex}} = \frac{((-0.041) - (-0.174))}{2 \times 1.96}$$

Lower bound: $\log(0.84) \approx -0.174$

Upper bound: $\log(0.96) \approx -0.041$

CI: $(-0.174, -0.041)$

$$\tau_{\text{sex}} = \frac{1}{\sigma_{\text{sex}}^2}$$

Model 2

Informative Prior

The Features

| Feature | Description | Prior Distribution |
|-----------------------------------|---|-------------------------------------|
| Age | Age of the patient | $d\text{norm}(-0.6405, \tau_{age})$ |
| Sex | Gender of the patient (1 = Female, 0 = Male) | $d\text{norm}(-0.116, \tau_{sex})$ |
| Cholesterol (chol) | Serum cholesterol in mg/dL | $d\text{norm}(-0.507, \tau_{chol})$ |
| Fasting Blood Sugar (fbs) | Fasting blood sugar > 120 mg/dL (1 = True, 0 = False) | $d\text{norm}(0.466, \tau_{fbs})$ |
| Resting Blood Pressure (trestbps) | Resting blood pressure in mm Hg | $d\text{norm}(-1.242, \tau_{age})$ |

Model 2

Informative Prior

Bayesian **Logistic Regression** using JAGS with informative priors from published literature and clinical data

```
informative_model <- textConnection("model{
  for(i in 1:n){
    Y[i] ~ dbern(pi[i])
    logit(pi[i]) <- beta[1] +
      x[i,1]*beta[2] + x[i,2]*beta[3] + x[i,3]*beta[4] +
      x[i,4]*beta[5] + x[i,5]*beta[6] + x[i,6]*beta[7] +
      x[i,7]*beta[8] + x[i,8]*beta[9] + x[i,9]*beta[10] +
      x[i,10]*beta[11] + x[i,11]*beta[12] + x[i,12]*beta[13] +
      x[i,13]*beta[14]
  }
  #informative
  beta[1] ~ dnorm(0, 0.01)                      # Intercept
  beta[2] ~ dnorm(-0.6405 , 207.69896)          # Age
  beta[3] ~ dnorm(-0.1160 , 868.69806)          # Sex
  beta[5] ~ dnorm(1.2420, 38.71605)              # BP
  beta[6] ~ dnorm(0.507, 59.07882)                # Cholesterol
  beta[7] ~ dnorm(0.4660 , 17.92050)              # BP

  #uninformative
  beta[4] ~ dnorm(0, 0.01)                      #cp
  for(j in 8:14){beta[j] ~ dnorm(0, 0.01)}      # Uninformative for others
}")

")
```

The Model

Model 2

Informative Prior

Bayesian **Logistic Regression** using JAGS with informative priors from published literature and clinical data

```
Iterations = 2001:7000  
Thinning interval = 1  
Number of chains = 2  
Sample size per chain = 5000
```

1. Empirical mean and standard deviation for each variable, plus standard error of the mean:

| | Mean | SD | Naive SE | Time-series SE |
|----------|---------|---------|-----------|----------------|
| beta[1] | 0.1067 | 0.18225 | 0.0018225 | 0.0024044 |
| beta[2] | -0.6035 | 0.06602 | 0.0006602 | 0.0009282 |
| beta[3] | -0.1338 | 0.03337 | 0.0003337 | 0.0004310 |
| beta[4] | 0.8949 | 0.19231 | 0.0019231 | 0.0030497 |
| beta[5] | 0.5779 | 0.12191 | 0.0012191 | 0.0016426 |
| beta[6] | 0.3325 | 0.11014 | 0.0011014 | 0.0014541 |
| beta[7] | 0.1198 | 0.14799 | 0.0014799 | 0.0020537 |
| beta[8] | 0.3943 | 0.18461 | 0.0018461 | 0.0026427 |
| beta[9] | 0.1199 | 0.20927 | 0.0020927 | 0.0033647 |
| beta[10] | -0.6445 | 0.19219 | 0.0019219 | 0.0027871 |
| beta[11] | -0.9850 | 0.25882 | 0.0025882 | 0.0046254 |
| beta[12] | 0.3009 | 0.21942 | 0.0021942 | 0.0039017 |
| beta[13] | -0.8770 | 0.20440 | 0.0020440 | 0.0030423 |
| beta[14] | -0.6689 | 0.17888 | 0.0017888 | 0.0023530 |

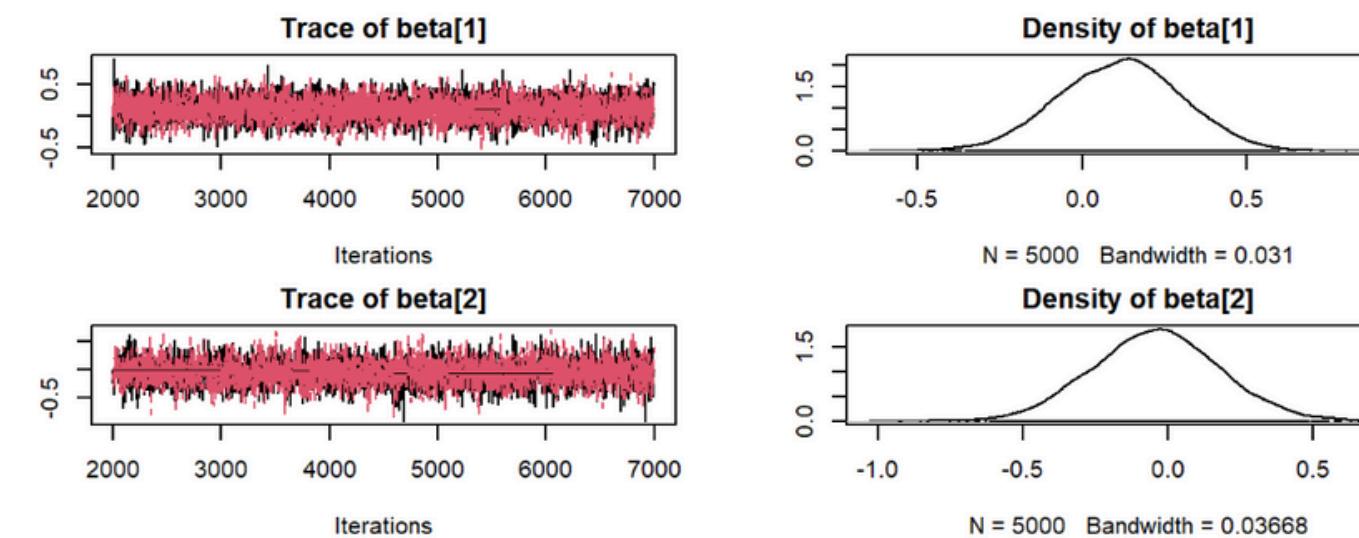
2. Quantiles for each variable:

| | 2.5% | 25% | 50% | 75% | 97.5% |
|----------|----------|----------|---------|---------|----------|
| beta[1] | -0.24982 | -0.01703 | 0.1069 | 0.2270 | 0.46642 |
| beta[2] | -0.73350 | -0.64816 | -0.6034 | -0.5594 | -0.47510 |
| beta[3] | -0.19706 | -0.15674 | -0.1343 | -0.1112 | -0.06775 |
| beta[4] | 0.51743 | 0.76569 | 0.8931 | 1.0221 | 1.27887 |
| beta[5] | 0.34287 | 0.49580 | 0.5785 | 0.6577 | 0.82123 |
| beta[6] | 0.11671 | 0.25744 | 0.3311 | 0.4064 | 0.55745 |
| beta[7] | -0.16623 | 0.02079 | 0.1179 | 0.2216 | 0.40960 |
| beta[8] | 0.03371 | 0.27301 | 0.3909 | 0.5189 | 0.75768 |
| beta[9] | -0.28102 | -0.02384 | 0.1204 | 0.2569 | 0.54235 |
| beta[10] | -1.02942 | -0.77616 | -0.6406 | -0.5160 | -0.27181 |
| beta[11] | -1.49682 | -1.15808 | -0.9800 | -0.8136 | -0.47735 |
| beta[12] | -0.13462 | 0.15199 | 0.3023 | 0.4515 | 0.72571 |
| beta[13] | -1.28783 | -1.01180 | -0.8744 | -0.7388 | -0.48335 |
| beta[14] | -1.02677 | -0.79002 | -0.6671 | -0.5461 | -0.32456 |

Convergence Diagnostics

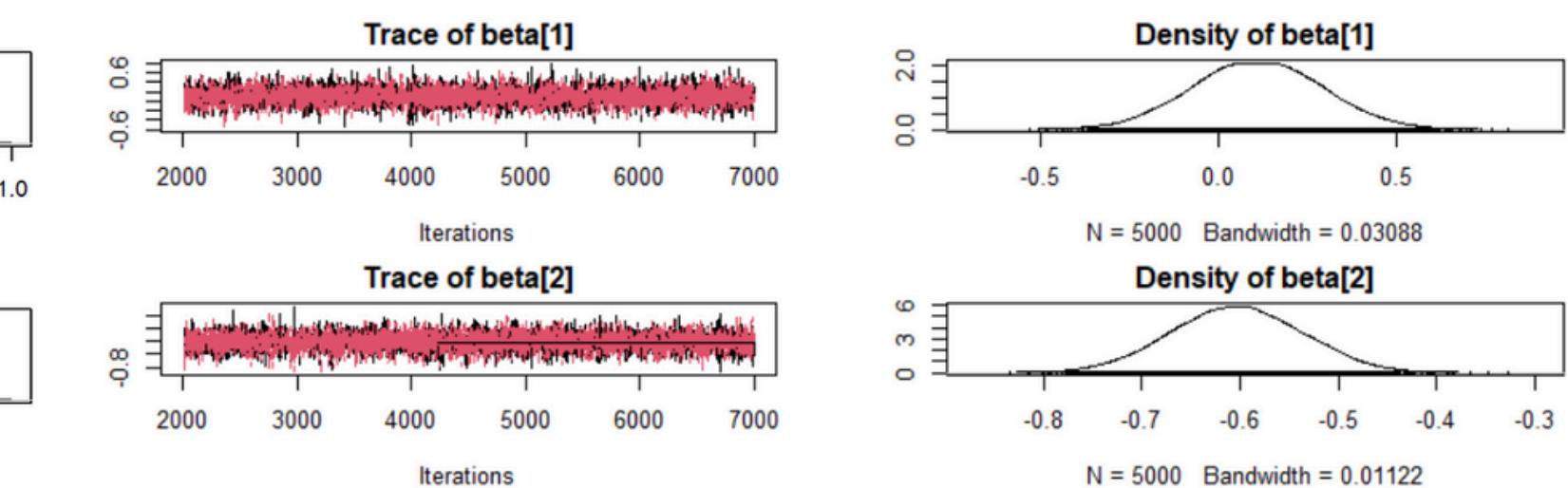
Plot Samples
Model 1

```
par(mar = c(4, 4, 2, 2))
plot(samps1)
```



Model 2

```
par(mar = c(4, 4, 2, 2))
plot(samps2)
```



The trace plots of all features appear stable, with the samples fluctuating around a consistent mean. This suggests that the MCMC chain for all features have likely converged.

Convergence Diagnostics

Autocorrelation Model 1

```
autocorr(samps1[[1]],lag=1)
```

```
, , beta[1]  
  
    beta[1]   beta[2]   beta[3]   beta[4]   beta[5]   beta[6]  
Lag 1 0.2687634 -0.110121 -0.1248178 -0.03687818 0.02349257 0.00207015  
    beta[7]   beta[8]   beta[9]   beta[10]  beta[11]  beta[12]  
Lag 1 0.04001203 -0.05496372 -0.138658 0.04163247 0.1356862 0.08274562  
    beta[13]  beta[14]  
Lag 1 0.09776163 0.07062796
```

```
, , beta[2]  
  
    beta[1]   beta[2]   beta[3]   beta[4]   beta[5]   beta[6]  
Lag 1 -0.119192 0.451756 0.1051748 -0.03775795 -0.2492575 -0.1563267  
    beta[7]   beta[8]   beta[9]   beta[10]  beta[11]  beta[12]  
Lag 1 -0.09707746 0.06929104 0.3985965 0.04449018 0.04458159 0.01237005  
    beta[13]  beta[14]  
Lag 1 -0.118871 -0.02422661
```

```
, , beta[3]
```

Autocorrelation score of all features are near 0, indicating good mixing, meaning there is no prolonged periods where the values are stuck in one region.

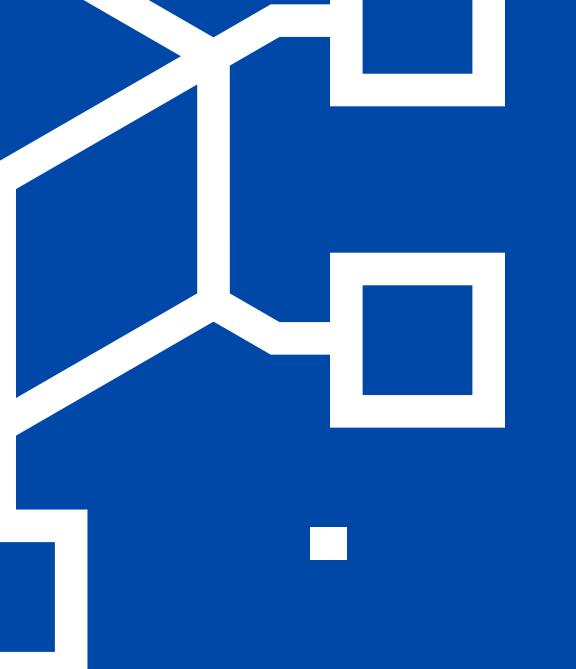
Model 2

```
autocorr(samps2[[1]],lag=1)
```

```
, , beta[1]  
  
    beta[1]   beta[2]   beta[3]   beta[4]   beta[5]   beta[6]  
Lag 1 0.2430762 -0.01268965 -0.01949307 -0.01231936 -0.01848734 -0.0026265  
    beta[7]   beta[8]   beta[9]   beta[10]  beta[11]  beta[12]  
Lag 1 0.01550287 -0.003939156 -0.01731265 0.01997397 0.1648376 0.08230633  
    beta[13]  beta[14]  
Lag 1 0.132114 0.04006681
```

```
, , beta[2]  
  
    beta[1]   beta[2]   beta[3]   beta[4]   beta[5]   beta[6]  
Lag 1 -0.003879257 0.2658891 -0.02434257 -0.01077814 -0.03400826 -0.001993879  
    beta[7]   beta[8]   beta[9]   beta[10]  beta[11]  beta[12]  
Lag 1 -0.02747256 0.01810433 0.1391447 0.05599395 0.008432148 -0.03007324  
    beta[13]  beta[14]  
Lag 1 -0.02280666 0.00961207
```

```
, , beta[3]
```



Convergence Diagnostics

Effective Sample Size (ESS)

Model 1

```
effectiveSize(samps1)

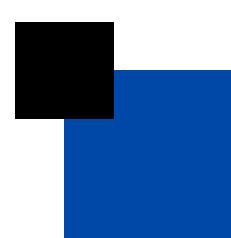
##  beta[1]  beta[2]  beta[3]  beta[4]  beta[5]  beta[6]  beta[7]  beta[8]
## 5362.433 3452.525 3398.164 4198.148 4763.223 3755.105 4734.752 4907.463
##  beta[9] beta[10] beta[11] beta[12] beta[13] beta[14]
## 3298.924 4771.133 3067.367 3134.955 4601.558 5235.926
```

Model 2

```
effectiveSize(samps2)

##  beta[1]  beta[2]  beta[3]  beta[4]  beta[5]  beta[6]  beta[7]  beta[8]
## 5911.973 5947.364 6467.355 4376.174 5740.640 6186.220 5421.869 5743.494
##  beta[9] beta[10] beta[11] beta[12] beta[13] beta[14]
## 4410.750 5333.056 2936.279 3431.407 4533.102 5335.022
```

Samples are effectively independent and have likely converged because all features has a high ESS.



Convergence Diagnostics

Gelman Rubin Diagnostic: Potential scale reduction factors (PSRF)

Model 1

```
gelman.diag(samps1)
```

Potential scale reduction factors:

| | Point est. | Upper C.I. |
|----------|------------|------------|
| beta[1] | 1 | 1.00 |
| beta[2] | 1 | 1.00 |
| beta[3] | 1 | 1.00 |
| beta[4] | 1 | 1.00 |
| beta[5] | 1 | 1.00 |
| beta[6] | 1 | 1.00 |
| beta[7] | 1 | 1.00 |
| beta[8] | 1 | 1.00 |
| beta[9] | 1 | 1.01 |
| beta[10] | 1 | 1.00 |
| beta[11] | 1 | 1.00 |
| beta[12] | 1 | 1.00 |
| beta[13] | 1 | 1.00 |
| beta[14] | 1 | 1.00 |

Multivariate psrf

1

PSRF value for all features are 1, which is less than 1.1, indicating that the MCMC chains have likely converged to the target posterior distribution.

Model 2

```
gelman.diag(samps2)
```

Potential scale reduction factors:

| | Point est. | Upper C.I. |
|----------|------------|------------|
| beta[1] | 1 | 1.00 |
| beta[2] | 1 | 1.00 |
| beta[3] | 1 | 1.00 |
| beta[4] | 1 | 1.00 |
| beta[5] | 1 | 1.00 |
| beta[6] | 1 | 1.00 |
| beta[7] | 1 | 1.00 |
| beta[8] | 1 | 1.00 |
| beta[9] | 1 | 1.00 |
| beta[10] | 1 | 1.01 |
| beta[11] | 1 | 1.00 |
| beta[12] | 1 | 1.01 |
| beta[13] | 1 | 1.02 |
| beta[14] | 1 | 1.01 |

Multivariate psrf

1.01

Convergence Diagnostics

Geweke's diagnostic

Model 1

```
for(i in seq_along(samps1)){
  print(geweke.diag(samps1[[i]]))
}
```

Fraction in 1st window = 0.1
Fraction in 2nd window = 0.5

beta[1] beta[2] beta[3] beta[4] beta[5] beta[6] beta[7] beta[8]
1.75386 2.22747 0.22225 -0.01006 -1.53144 0.49570 -0.34065 0.65102
beta[9] beta[10] beta[11] beta[12] beta[13] beta[14]
1.20148 0.90319 -1.21451 -1.69368 0.38233 -0.39950

Fraction in 1st window = 0.1
Fraction in 2nd window = 0.5

beta[1] beta[2] beta[3] beta[4] beta[5] beta[6] beta[7] beta[8]
0.1723 1.4734 0.5982 -1.1559 0.5426 -0.5809 1.8061 1.0272
beta[9] beta[10] beta[11] beta[12] beta[13] beta[14]
0.0544 -0.9171 0.9497 0.8650 0.7661 0.6012

```
for(i in seq_along(samps2)){
  print(geweke.diag(samps2[[i]]))
}
```

Fraction in 1st window = 0.1
Fraction in 2nd window = 0.5

beta[1] beta[2] beta[3] beta[4] beta[5] beta[6] beta[7] beta[8]
0.5091 0.7084 -0.6785 0.1535 -1.3652 -0.1191 -0.9736 -1.7198
beta[9] beta[10] beta[11] beta[12] beta[13] beta[14]
-0.7214 0.5114 0.5916 -0.3239 1.5231 -0.2307

Fraction in 1st window = 0.1
Fraction in 2nd window = 0.5

beta[1] beta[2] beta[3] beta[4] beta[5] beta[6] beta[7] beta[8]
-0.287698 -1.433303 0.005727 0.051982 -0.348273 -1.197923 0.182125 0.977683
beta[9] beta[10] beta[11] beta[12] beta[13] beta[14]
-0.316734 0.855738 0.163956 1.780198 -1.786717 -0.973307

Model 2

$|z|$ for all feataures are less than 2 indicating good convergence.

Comparison between Models

Bayes Factor

```
{r}
#fit model1
mod1 <- brm(target | trials(target) ~ 1, data = df, family = binomial())
#fit model2
mod2 <- brm(target | trials(target) ~ age + sex + cp + trestbps + chol + fbs +
restecg + thalach + exang + oldpeak + slope + ca + thal, data = df, family =
binomial())

loo_model1 <- loo(mod1)
loo_model2 <- loo(mod2)

print(loo_model1)
print(loo_model2)

loo_comparison <- loo_compare(loo_model1, loo_model2)
print(loo_comparison)
```

```
{r}
elpd_1 <- -0.4
elpd_2 <- -0.3
delta_elpd <- elpd_1 - elpd_2
bayes_factor <- exp(delta_elpd)
bayes_factor
```

Warning message:
In check_dep_version() : ABI version
lme4 was built with Matrix ABI vers
Current Matrix ABI version is 0
Please re-install lme4 from source
[1] 0.9048374

Bayes Factor of Model 1 (Uninformative)

| | Estimate
<S3: AsIs> | SE
<S3: AsIs> |
|----------|------------------------|------------------|
| elpd_loo | -0.4 | 0.0 |
| p_loo | 0.0 | 0.0 |
| looic | 0.8 | 0.0 |

Bayes Factor of Model 2 (Informative)

| | Estimate
<S3: AsIs> | SE
<S3: AsIs> |
|----------|------------------------|------------------|
| elpd_loo | -0.3 | 0.1 |
| p_loo | 0.2 | 0.1 |
| looic | 0.5 | 0.2 |

A Bayes factor < 1 indicates that model2 is more strongly supported.

Comparison between Models

```
criteria <- c("WAIC")
model1 <- c(414.8723)
model2 <- c(378.0674)

comparison_table <- data.frame(
  Criterion = criteria,
  Model_1 = model1,
  Model_2 = model2,
  Difference = model2 - model1
)
```

```
print(comparison_table)
```

```
##   Criterion Model_1 Model_2 Difference
## 1       WAIC  414.8723 378.0674    -36.8049
```

Which model fits better? Model 2 (informative) has a lower WAIC, indicating a better fit towards the data.

WAIC of Model 1 (Uninformative)

```
samps_like1 <- samps1[,1:p]
like1 <- rbind(samps1[[1]],samps1[[2]])
like1[like1 <= 0] <- 1e-6      #change negative values to 1e-6
fbar1 <- colMeans(like1)
Pw1 <- sum(apply(log(like1),2,var))
WAIC1 <- -2*sum(log(fbar1))+2*Pw1
WAIC1
```

```
## [1] 422.0567
```

WAIC of Model 2 (Informative)

```
samps_like2 <- samps2[,1:p]
like2 <- rbind(samps2[[1]],samps2[[2]])
like2[like2 <= 0] <- 1e-6      #change negative values to 1e-6
fbar2 <- colMeans(like2)
Pw2 <- sum(apply(log(like2),2,var))
WAIC2 <- -2*sum(log(fbar2))+2*Pw2
WAIC2
```

```
## [1] 376.5296
```

Best Model: Model 2

SIGNIFICANT PREDICTORS:

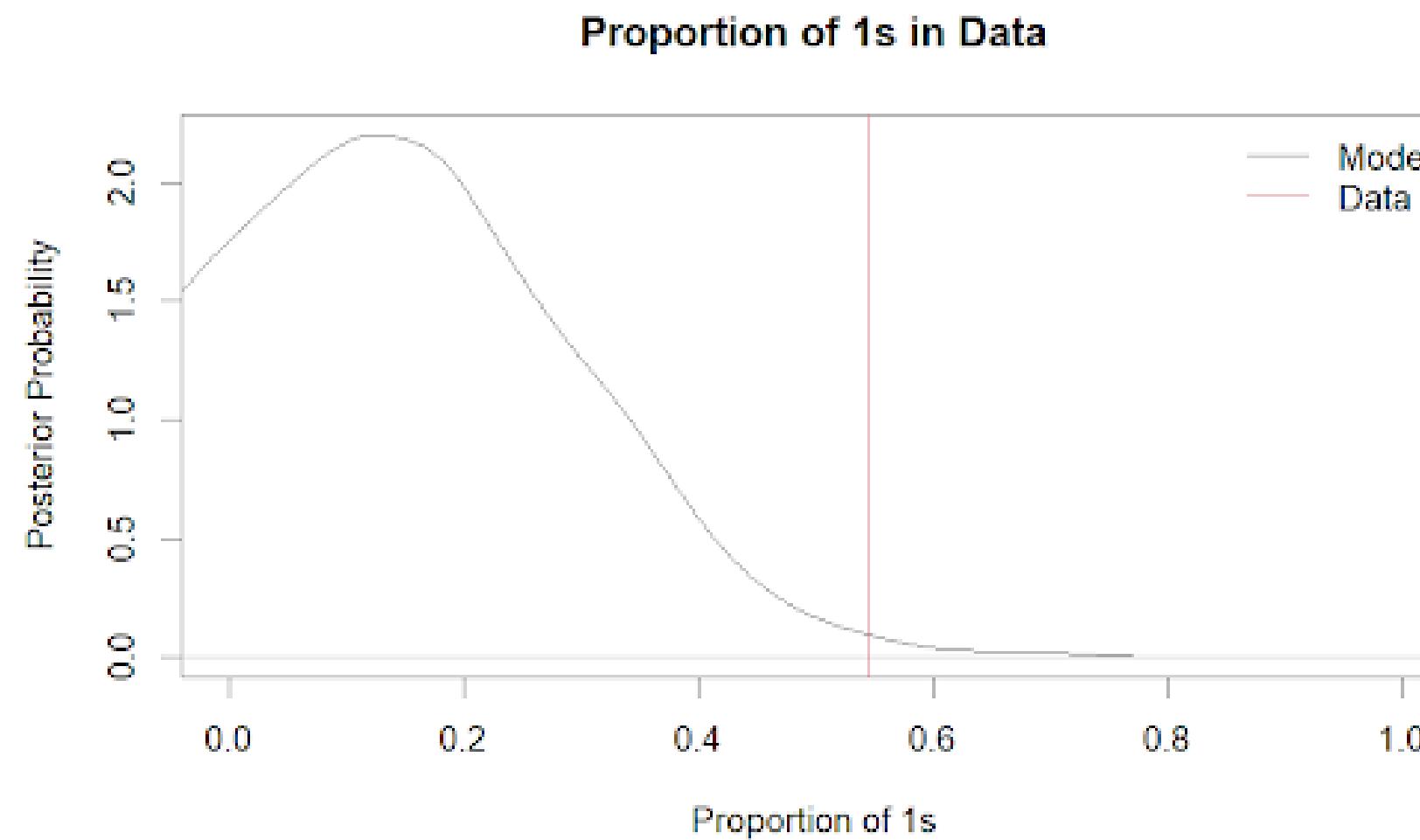
beta[2] (Age): CI (-0.73510 to -0.47889) → Significant.
beta[3] (Sex): CI (-0.19741 to -0.06822) → Significant.
beta[4] (Chest Pain Type - cp): CI (0.51624 to 1.29794) → Significant.
beta[5] (Resting BP - trestbps): CI (0.33838 to 0.82097) → Significant.
beta[6] (Cholesterol - chol): CI (0.12656 to 0.53975) → Significant.
beta[8] (Resting ECG - restecg): CI (0.02759 to 0.75110) → Significant.
beta[10] (Exercise-induced Angina - exang): CI (-1.03451 to -0.27444) → Significant.
beta[11] (ST Depression - oldpeak): CI (-1.50655 to -0.47637) → Significant.
beta[13] (Number of Major Vessels - ca): CI (-1.27769 to -0.49234) → Significant.
beta[14] (Thalassemia - thal): CI (-1.02752 to -0.32645) → Significant.

INTERPRETATION:

- A one-unit increase in resting bp, cholesterol, resting ECG, and slope of peak exercise increases the odds of having heart disease.
- A one-unit increase in age and number of major vessels decreases the odds of having heart disease.
- The odds of having heart disease for females are lower than for males.
- The odds of having heart disease for patients with higher chest pain types are higher than for patients with lower chest pain types.
- The odds of having heart disease for patients with exercise induced angina and higher thalassemia type are lower.

Posterior Predictive Check

```
D <- samps2[[1]]  
p_data <- mean(Y)  
D0 <- c(p_data)  
Dnames <- c("Proportion of 1s in Data")  
  
pval <- rep(0, 1)  
names(pval) <- Dnames  
  
for (j in 1:1) {  
  posterior_density <- density(D[,j])  
  plot(posterior_density, xlim = c(0, 1), xlab = "Proportion of 1s", ylab = "Posterior Probability", main = Dnames[j])  
  abline(v = D0[j], col = 2)  
  legend("topright", c("Model", "Data"), lty = 1, col = 1:2, bty = "n")  
  
  pval[j] <- mean(D[,j] > D0[j])  
}
```



pval

Proportion of 1s in Data
0.0096



Conclusion

In this analysis, we compared two Bayesian logistic regression models, one with uninformative priors and another with informative priors, to predict heart disease risk based on clinical and demographic factors. **The model with informative priors demonstrated superior predictive accuracy, reflected in a lower WAIC (365.37 vs. 414.87).** This highlights the value of incorporating prior domain knowledge into Bayesian analysis, though posterior predictive checks revealed some room for improvement in model fit.

Significant predictors of heart disease included age, sex, chest pain type, blood pressure, cholesterol, resting ECG, exercise-induced angina, ST depression, major vessel count, and thalassemia. Future efforts should focus on:

- **refining prior distributions with additional clinical data (incorporating lifestyle factors or family history as covariates).**
- **validating results with external datasets to enhance generalizability and robustness.**

Thank You

Apakah ada pertanyaan?

Kelompok 1