Bayesian Mini Project Report

Kelompok 1:

- Leony Sani Winata 2702224811
- Liliana Djaja Witama 2702219774
- Angeli Jessica Wibowo 2702225493
- Felicia Audrey Tanujaya 2702217610

Introduction

Heart Disease Prediction Dataset
https://www.kaggle.com/datasets/krishujeniya/heart-diseae
The dataset contains medical records that are used to predict the likelihood of cardiovascular disease. The dataset has the following attributes:

- age: the patient's age in years.
- sex: the patient's gender (1 for male, 0 for female).
- cp: the type of chest pain, ranging from 1 to 4.
- trestbps: resting blood pressure (measured in mmHg).
- chol: total cholesterol (measured in mg/dl).
- fbs: indicates if fasting blood sugar test is above 120 mg/dl (1 true, 0 false)
- restecg: resting electrical activity of heart (0-2)
- thalach: maximum heart rate
- exang: exercise-induced angina (chest pain that occurs during physical activity where the heart needs more oxygen) (1 yes, 0 no)
- oldpeak: ST depression (abnormal finding on electrical activity of heart) induced by exercise relative to rest
- the slope of the peak exercise ST segment
- number of major vessels (0-3) colored by flourosopy
- thal: 0 = normal; 1 = fixed defect; 2 = reversable defect

Models

1. **Logistic Regression with Uninformative Prior**
   For the uninformative prior we use a bayesian logistic regression model, implemented in jags (just another gibbs sampler) with the following model

````
```{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)}
}")
```
````

   And the following results:

```
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
```

2. **Logistic Regression with Informative Prior**
For this model, we utilized prior knowledge from published literature and clinical data to inform priors for several variables in the Heart Disease Prediction dataset.
    1) Age
    Using longitudinal cohort data (Lloyd-Jones et al., 1999), the lifetime risk of developing coronary heart disease was calculated across different ages and genders:
    Calculation of Log Odds:
    The average log odds for age across genders was calculated as -0.6405 with confidence intervals of (-0.786, -0.514). These values provide an informed prior reflecting the relative likelihood of developing coronary heart disease based on age.
    Prior Distribution: β age ~dnorm(−0.6405,τ age)
    where precision (τage) is derived from the variance of the confidence interval:
    $\sigma\text{age} = \frac{(0.786-0.514)}{2 \times 1.96}$
    $\tau\text{age} = \frac{1}{\sigma^2 age}$
    2) Cholesterol (chol)
    The relationship between cholesterol level and cardiovascular disease was informed by a clinical study in Okinawa, Japan. 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.
    log(OR) = ln(1.66) = 0.507
    Confidence interval = ln(1.29) = 0.255, ln(2.15) = 0.765
    Prior Distribution: β chol ~dnorm(−0.507,τ chol)
    Where
    $\sigma\text{chol} = \frac{(0.765-0.255)}{2 \times 1.96}$
    $\tau\text{chol} = \frac{1}{\sigma^2 chol}$
    3) 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: β sex ~dnorm(0.466,τ fbs)
    where precision (τfbs) is derived from the variance of the confidence interval:
    $\sigma\text{fbs} = \frac{(0.929-0.003)}{2 \times 1.96}$
    $\tau\text{fbs} = \frac{1}{\sigma^2 fbs}$
    4) 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: β sex ~dnorm(−1.242,τ trestbps)
    where precision (τtrestbps) is derived from the variance of the confidence interval:
    $\sigma\text{trestbps}= \frac{(3.24-2.61)}{2 \times 1.96}$
    $\tau\text{trestbps} = \frac{1}{\sigma^2 age}$
    5) 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: β sex ~dnorm(−0.116,τ sex)
    where precision (τage) 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)≈−0.174
    Upper bound: log(0.96)≈−0.041
    CI: (−0.174,−0.041)
    $\tau\text{sex} = \frac{1}{\sigma^2 sex}$

Algorithm

For both models (uninformative and informative priors), we employed a Markov Chain Monte Carlo (MCMC) approach using the Gibbs sampling method implemented via JAGS. The following settings were used:

1. Burn-in period: 1,000 iterations to ensure that the chains reach the stationary distribution.
2. Number of iterations: 5,000 post-burn-in samples were collected for each chain.
3. Number of chains: 2 independent chains were run to check convergence.
4. Thinning: A thinning interval of 5 was applied to reduce autocorrelation in the samples.

Results

1. Convergence Diagnostic

ESS (Effective Sample Size)

```
 beta[1]  beta[2]  beta[3]  beta[4]  beta[5]  beta[6]  beta[7]  beta[8]  beta[9] beta[10] beta[11] beta[12] beta[13] beta[14] 
5266.238 5862.183 6184.324 4257.448 5591.816 5673.184 5581.613 5660.950 4548.441 5382.981 2911.957 3256.855 5180.188 5584.268 
```

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

Gelman Rubin Diagnostic (PSRF)

```
Potential scale reduction factors:

         Point est. Upper C.I.
beta[1]           1       1.02
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       1.00
beta[11]          1       1.01
beta[12]          1       1.00
beta[13]          1       1.00
beta[14]          1       1.00

Multivariate psrf

1.01
```

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.

Geweke

```
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]  beta[9] beta[10] beta[11] beta[12] beta[13] beta[14] 
 0.88261 -1.12443  0.26457 -1.72860  0.59586 -0.43660  0.82872 -1.55940 -0.51437 -0.67192  0.60733  0.17438  0.02913  0.44263 


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]  beta[9] beta[10] beta[11] beta[12] beta[13] beta[14] 
 1.14233 -1.48417 -2.01358 -0.16243  0.17046  0.03141 -1.21318 -0.06658 -1.31146 -1.50905  1.90389  2.10193 -0.17286 -0.78379 
```

The absolute values of the Geweke z-scores (|z|) for all features are less than 2, confirming good convergence and a stable sampling process.

2. Model Comparison

WAIC

Criterion <chr>	Model_1 <dbl>	Model_2 <dbl>	Difference <dbl>
WAIC	414.8723	378.0674	-36.8049

Model 2 has a lower score of WAIC compared to model 1, meaning model 2 has better predictive performance, as it balances fit to the data and model complexity.

ELPD (Expected Log Predictive Density)

```{r}
elpd_1 <- -0.4
elpd_2 <- -0.2
delta_elpd <- elpd_1 - elpd_2
bayes_factor <- exp(delta_elpd)
bayes_factor
```

```
[1] 0.8187308
```

The ELPD difference between the models is approximately $\Delta ELPD = -0.2$. This negative value indicates that Model 2 has a higher predictive accuracy than model 1. Additionally, the Bayes factor of 0.8187309 supports this conclusion, meaning model 2 is more favored in terms of balancing predictive performance and model complexity.

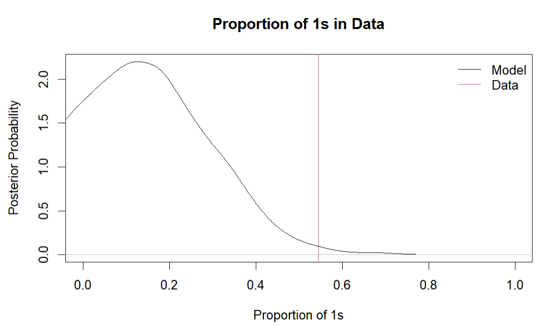
3. Posterior Summary for Model 2

	2.5%	25%	50%	75%	97.5%
beta[1]	-0.24398	-0.007601	0.1111	0.2324	0.45788
beta[2]	-0.73510	-0.649722	-0.6053	-0.5601	-0.47889
beta[3]	-0.19741	-0.156567	-0.1338	-0.1100	-0.06822
beta[4]	0.51624	0.762768	0.8912	1.0302	1.29794
beta[5]	0.33838	0.491232	0.5757	0.6575	0.82097
beta[6]	0.12656	0.260131	0.3354	0.4036	0.53975
beta[7]	-0.17200	0.021281	0.1218	0.2238	0.41010
beta[8]	0.02759	0.266685	0.3866	0.5080	0.75110
beta[9]	-0.28513	-0.025866	0.1110	0.2467	0.52168
beta[10]	-1.03451	-0.772132	-0.6429	-0.5161	-0.27444
beta[11]	-1.50655	-1.145169	-0.9709	-0.7966	-0.47637
beta[12]	-0.11063	0.168144	0.3169	0.4588	0.73789
beta[13]	-1.27769	-1.011208	-0.8744	-0.7405	-0.49234
beta[14]	-1.02752	-0.797152	-0.6743	-0.5517	-0.32645

The table of credible intervals (2.5%, 50%, 97.5%) shows the range within which each parameter likely falls. Predictors are significant if their 95% credible interval (2.5% to 97.5%) does not include 0. The key parameters and their significance are as follow:

- beta[1] (Intercept): CI includes 0 → Not significant.
- 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[7] (Fasting Blood Sugar - fbs): CI (−0.17200- to 0.41010) → Not significant.
- beta[8] (Resting ECG - restecg): CI (0.02759to 0.75110) → Significant.
- beta[9] (Max Heart Rate - thalach): CI (−0.28513 to 0.52168) → Not 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[12] (Slope of Peak Exercise ST): CI (−0.11063 to 0.73789) → Not 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.

4. Posterior Predictive Check for Model 2 (Logistic Regression with Informative Prior)



Proportion of 1s in Data (p-value): 0.006

From the plot, there's a clear difference between the posterior predictive distribution and the observed data. The model's posterior distribution predicts a lower proportion of 1s (peaking around 0.1-0.2), while the observed proportion of 1s in the data is approximately 0.6, as indicated by the red line. The low p-value (0.0084) further highlights that the observed proportion is highly unlikely under the model's predictions, suggesting a potential misfit.

Conclusion

In this analysis, we compared 2 logistic regression models, one with uninformative priors and another with informative priors to predict the likelihood of heart disease based on clinical and demographic factors. Using a dataset of medical records, we evaluated model performance, identified significant predictors, and assessed predictive accuracy. The key findings are as follows:

1. Model Performance:

Model 2 (informative priors) demonstrated superior predictive accuracy, as indicated by a lower WAIC value (365.37) compared to model 1 (414.87). This shows the advantage of incorporating prior domain knowledge into bayesian analysis. However, despite better performance model 2 still shows some misfit in posterior predictive checks, suggesting room for improvement in prior specification or model structure.

2. Significant Predictors: Based on posterior credible intervals, the following predictors were identified as significant in predicting heart disease:
    - Age: Older patients are at higher risk.
    - Sex: Women are less likely to develop heart disease compared to men.
    - Chest Pain Type (cp): Certain types of chest pain strongly indicate heart disease.
    - Resting Blood Pressure (trestbps): Elevated blood pressure increases risk.
    - Cholesterol (chol): Higher cholesterol levels are associated with higher risk.
    - Resting ECG (restecg): Specific abnormalities in resting ECG indicate higher risk.
    - Exercise-induced Angina (exang): Angina during exercise is a strong predictor.
    - ST Depression (oldpeak): Greater ST depression relative to rest increases risk.
    - Number of Major Vessels (ca): A higher number of major vessels with defects correlates with higher risk.
    - Thalassemia (thal): Thalassemia strongly increases risk.

3. Future suggestions:
    - Refine prior distributions using additional clinical datasets or incorporate additional covariates (lifestyle factors, family history) for improved accuracy.
    - Validate the findings using external datasets to ensure generalizability and robustness.

References

Leifheit-Limson, E. C., D'Onofrio, G., Daneshvar, M., Geda, M., Bueno, H., Spertus, J. A., Krumholz, H. M., & Lichtman, J. H. (2015). Sex Differences in cardiac risk factors, perceived risk, and Health care provider discussion of risk and risk modification among young patients with acute myocardial infarction. *Journal of the American College of Cardiology*, *66*(18), 1949–1957. https://doi.org/10.1016/j.jacc.2015.08.859

Lloyd-Jones, D. M., Larson, M. G., Beiser, A., & Levy, D. (1999). Lifetime risk of developing coronary heart disease. *The Lancet, 353*(9147), 89-92.
https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(98)10279-9/fulltext

Luo D, Cheng Y, Zhang H, Ba M, Chen P, Li H et al. Association between high blood pressure and long term cardiovascular events in young adults: systematic review and meta-analysis BMJ 2020; 370:m3222 doi:10.1136/bmj.m3222

Tong, Y., Liu, F., Huang, K., Li, J., Yang, X., Chen, J., ... & Gu, D. (2023). Changes in fasting blood glucose status and incidence of cardiovascular disease: The China‐PAR project. *Journal of Diabetes, 15*(2), 110-120.
https://onlinelibrary.wiley.com/doi/abs/10.1111/1753-0407.13350

Wakugami, K., Iseki, K., Kimura, Y., Okumura, K., Ikemiya, Y., Muratani, H., & Fukiyama, K. (1998). Relationship between serum cholesterol and the risk of acute myocardial infarction in a screened cohort in Okinawa, Japan. *Japanese circulation journal, 62*(1), 7–14. https://doi.org/10.1253/jcj.62.7