Logistic Regression vs. Bayesian Logistic Regression in Breast Cancer identification

Morgan Huang, Yuxiang Feng, Issey Sone, Tongyu Wu, Enze Zhao

Mar 23 2025

Table of Contents

Motivation

- Breast Cancer is one of the most common types of cancer, approx. 1 in 8.[1]
- Traditional diagnostic methods like mammography can miss 10-30% of cancers, there is a need for better identification techniques.[2]
- When detected early, the 5-year survival rate is very high (over 90%), but is significantly lowered for late detection.[3]
- Using two approaches to Logistic Regression, a frequentist and Bayesian approach, what kind of performance can we get with our models?

Dataset

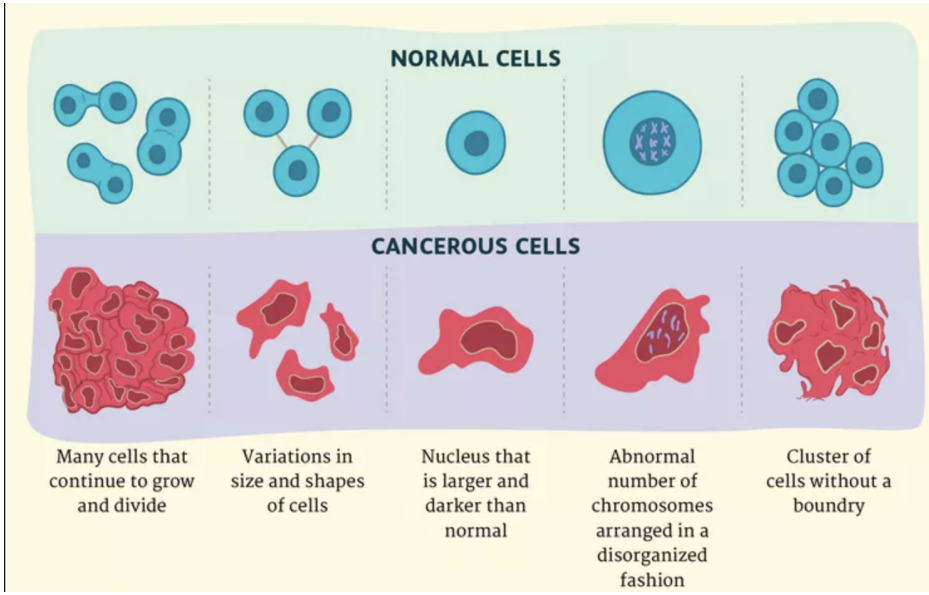
- Wisconsin Breast Cancer Dataset: link.
- Taken from UCI ML repository
- 699 observations
- 444 Benign, 239 Malignant cases (after removing missing values)
- 9 features

Variables and Their Visual Representation

Variables

- Clump Thickness
- Uniformity of Cell Size
- Uniformity of Cell Shape
- Marginal Adhesion
- Single Epithelial Cell Size
- Bare Nuclei
- Bland Chromatin
- Normal Nucleoli
- Mitoses

Note: All features are on a scale of 1–10.



Source: verywellhealth.com / Lynne Eldridge, MD (2023)

Histograms

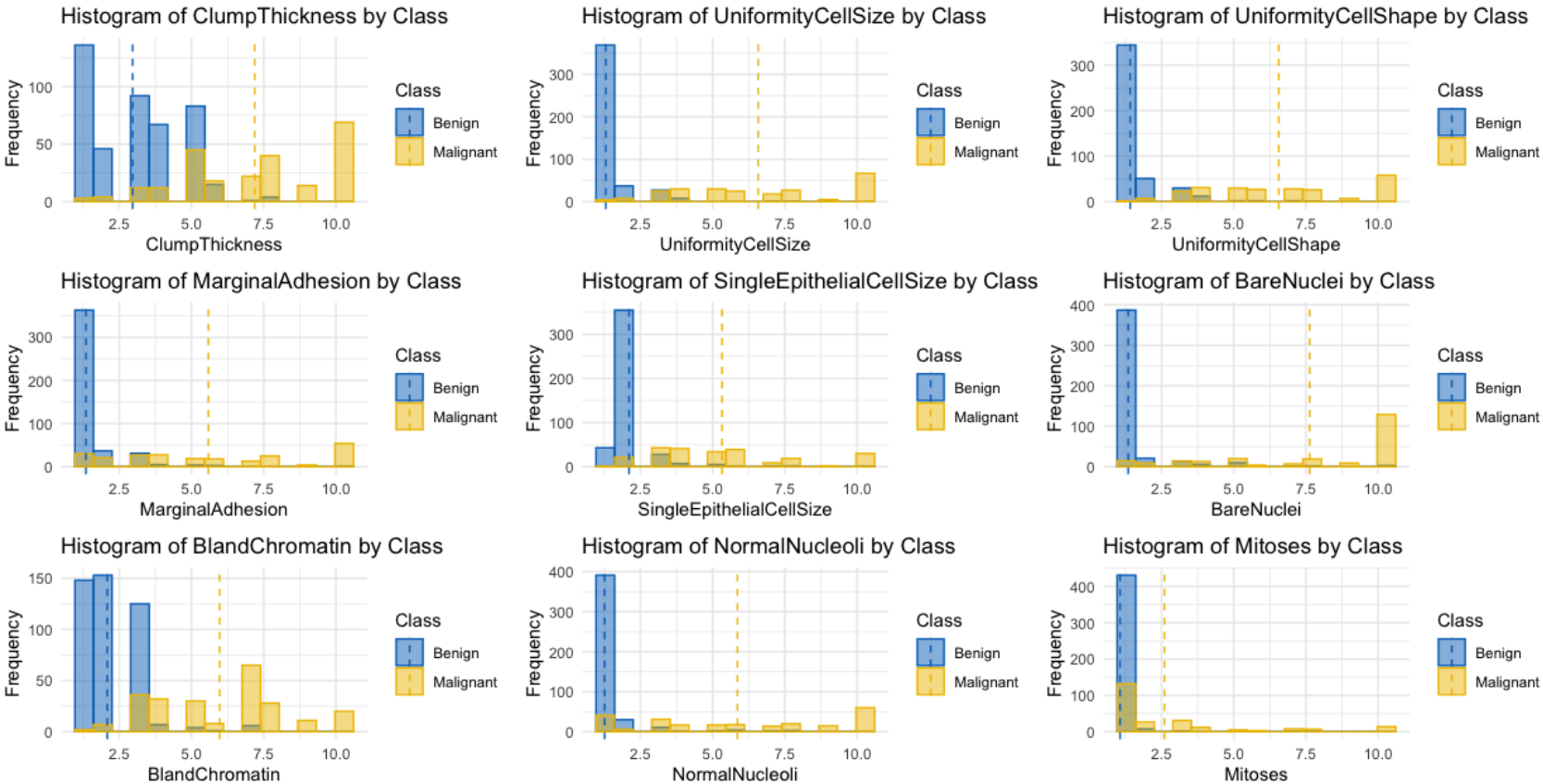


Figure 1: Histograms' of the 9 predictor variables in our dataset

Box plots

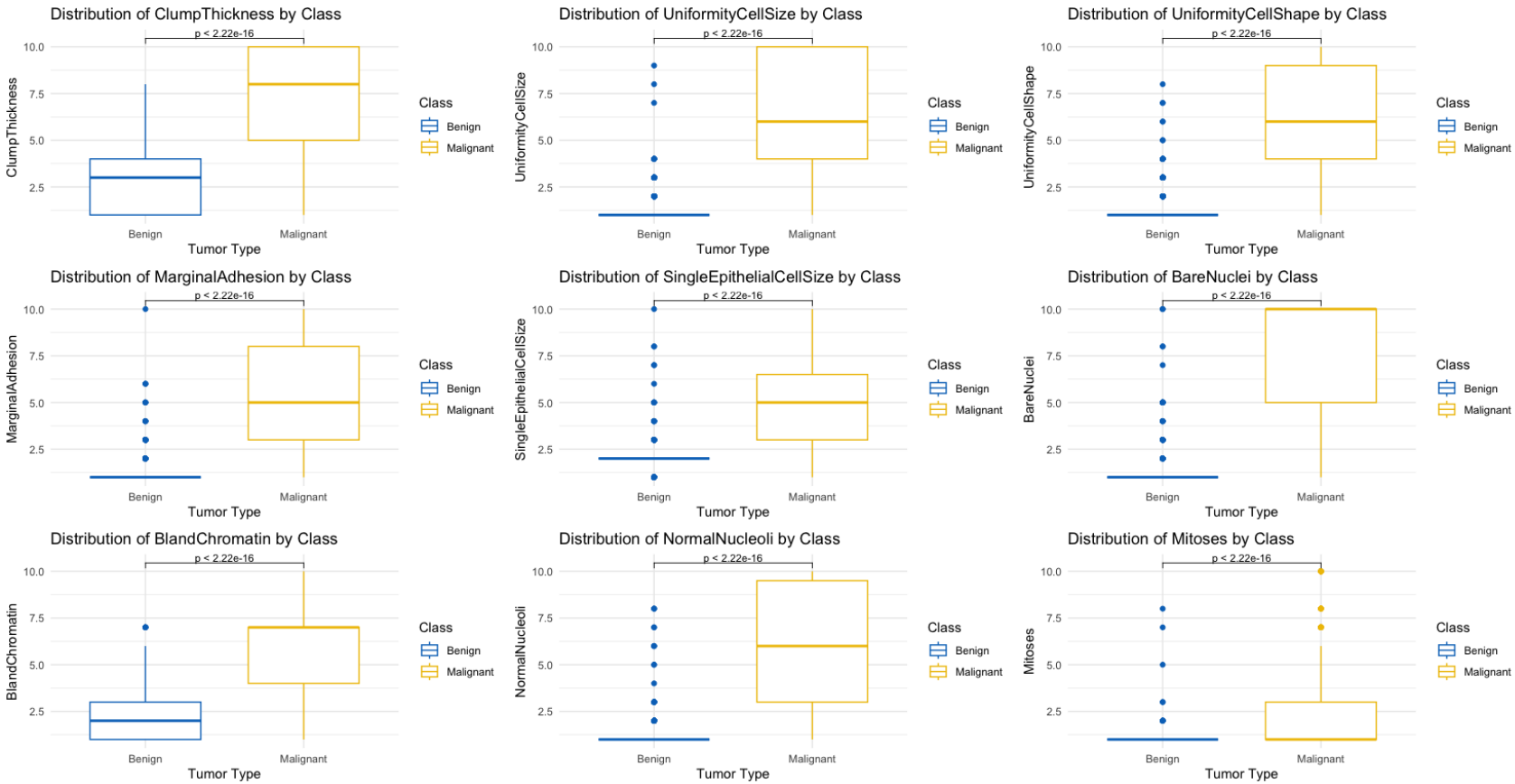


Figure 2: Box plots' of the 9 predictor variables in our dataset

Heatmap

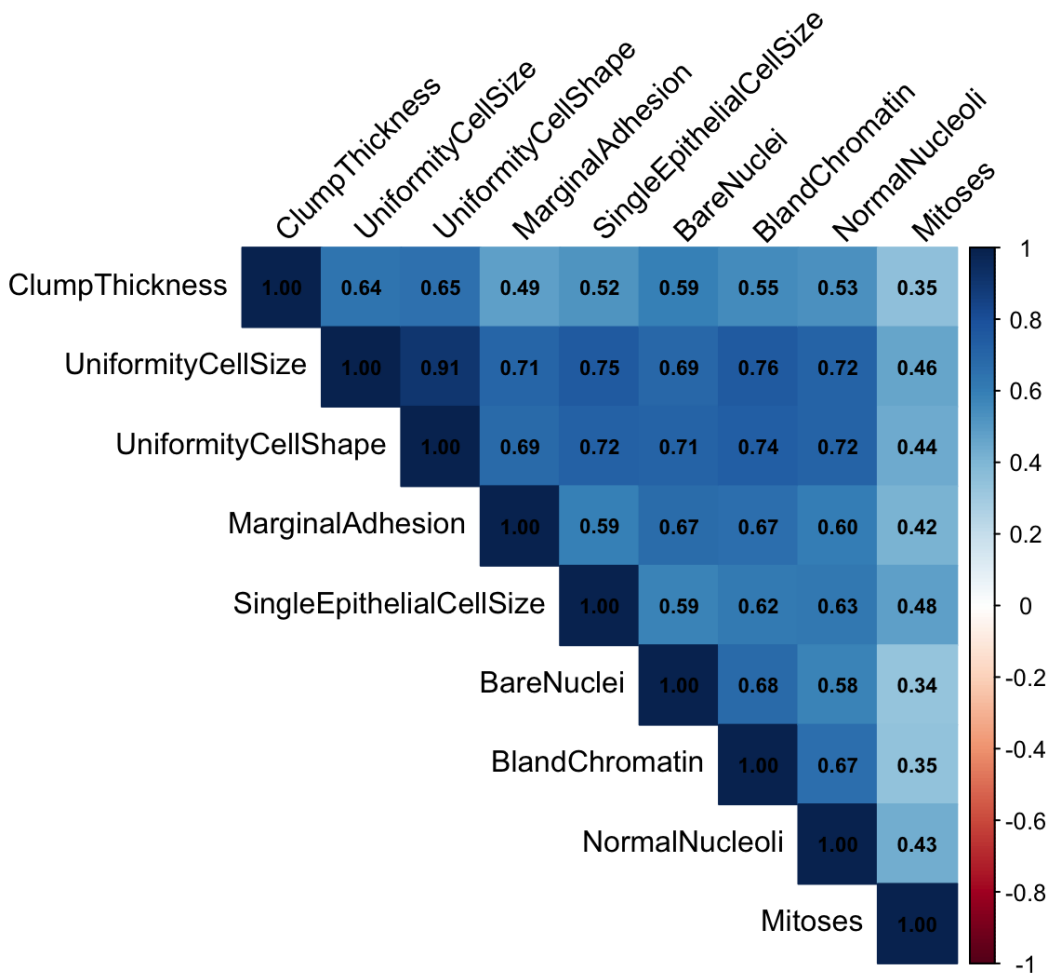


Figure 3: Heat map of the 9 predictor variables

Logistic Regression

- We will focus more on inference rather than prediction
- Models probability of $\pi(x)$, the probability of a patient having malignant tumor, as follows:

$$\text{logit}(\pi(x)) = \log(\frac{\pi(x)}{1-\pi(x)}) = \alpha + \beta X$$

- Solving for $\pi(x)$, we get $\pi(x) = \frac{\exp(\alpha+\beta X)}{1+\exp(\alpha+\beta X)}$
- Research Question: How does clump thickness affect breast cancer diagnosis? What predictors are sufficient in detecting malignant tumors?

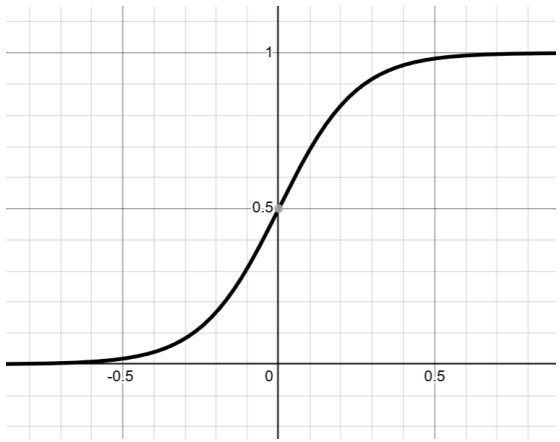


Figure 4: Logit function

Statistical Tests

- We would like to test to see if one or many predictors are statistically relevant in our model
- For just one predictor, β_j, we use a Wald test, testing for $H_0 : \beta_j = 0$ against $H_a : \beta_j \neq 0$, with test statistic:

$$z = \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)} \sim N(0, 1)$$

- For multiple predictors $\beta_i, \ldots, \beta_p$, we use LRT, testing for $H_0 : \beta_i = \cdots = \beta_j = 0$ against H_a : at least one $\beta_j \neq 0$:

$$D = -2(L_0 - L_a) \sim \chi^2_{(df)}$$

where L_0 and L_a are log likelihoods under H_0 and H_a, and degrees of freedom is difference between # of df under H_a and # of df under H_0

Clump Thickness

- We test if the level of Clump thickness is statistically significant in our model, where $\hat{\beta}_1 = 1.11$. We thus get: $z = 3.896$, with p-value less than 0.001
- Constructing a 95% confidence interval for β_1, we get:

$$(1.088, 1.132)$$

- We are 95% confident that the odds of having a malignant tumor will multiply by a factor between $\exp(1.088) = 2.97$ and $\exp(1.132) = 3.1$ when we increase the level of Clump thickness, controlling for all other predictors

Sufficient Predictors

- Given our short biological description of the predictors, we test if the following predictors are sufficient for our analysis
- Predictors to keep in reduced model: Clump Thickness, Marginal Adhesion, SECS, Bare Nuclei, and Mitosis
- Using LRT we compare reduced model with full model:

```
Likelihood ratio test

Model 1: Class ~ (Clump_thickness + Uniformity_of_cell_shape + Uniformity_of_cell_size +
    Marginal_adhesion + Single_epithelial_cell_size + Bare_nuclei +
    Bland_chromatin + Normal_nucleoli + Mitoses)^2
Model 2: Class ~ (Clump_thickness + Single_epithelial_cell_size + Marginal_adhesion +
    Bare_nuclei + Mitoses)^2
  #Df  LogLik  Df  Chisq Pr(>Chisq)
1  46 -22.060
2  16 -50.653 -30 57.185    0.00199 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 5: R output

- Test statistic D $= 57.185$, with $p < 0.01$
- Conclude that those predictors are not sufficient enough to create a diagnosis

Model Selection

- Goal: Obtain a model simple enough to interpret, and complex enough that it fits well with the data
- We can use many LRT tests to compare different models, but to also consider the number of parameters, we use Akaike Information Criteria(AIC), defined as:

$$AIC = -2(\log L(\hat{\theta}|X) - k)$$

  where k is the number of parameters in the model
- Describes the KL-distance between distribution of fitted values and expected values, accounting for k

Stepwise AIC

- We use backwards elimination for our stepwise AIC algorithm
- Start with all possible predictors and interaction terms(up to power of 2), and eliminate parameters such that the AIC is minimized
- Terminate when AIC cannot be minimized any further
- Final model with AIC = 108.99:
  - Removed: Uniformity of Cell Size, Single Epithelial Cell Size

```
Clump_thickness:Marginal_adhesion
Uniformity_of_cell_shape:Marginal_adhesion
Uniformity_of_cell_shape:Bare_nuclei
Marginal_adhesion:Bare_nuclei
Bare_nuclei:Mitoses
Bland_chromatin:Normal_nucleoli
Bland_chromatin:Mitoses
```

Figure 6: Interaction Terms Included

- An improvement from AIC = 136.12, when including all possible main effects and interactions

Bayesian Logistic Regression

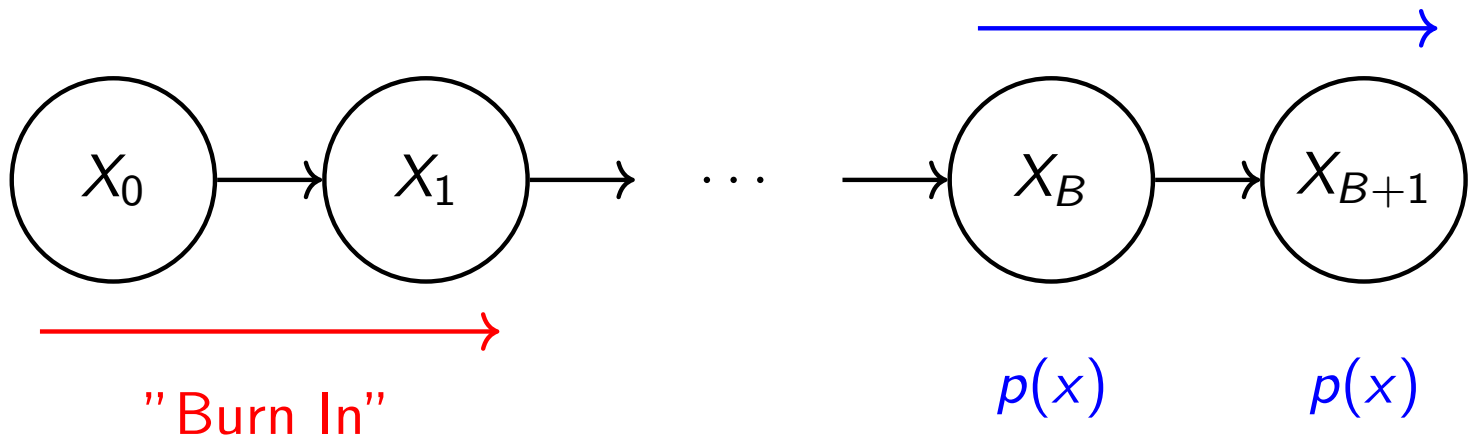
- Probabilistic approach to logistic regression using prior beliefs of the features and updates with the observations to form a posterior distribution
- $P(y|X,\beta) = \prod_{i=1}^{n} \sigma(\beta^T x_i)^{y_i}(1-\sigma(\beta^T x_i))^{1-y_i}$, where σ is the sigmoid function
- The posterior follows:

$$P(\beta|X,y) \propto P(y|X,\beta) \cdot P(\beta)$$

- Gives a full posterior distribution of predictors
- Incorporate Prior belief based on existing studies
- Regularization built in with the priors and is more robust to imbalanced classes and smaller datasets

MCMC (Monte Carlo Markov Chain)

- Class of algorithms to draw samples from a probability distribution, usually ones too complex
- Construct Markov Chain which converge to a target distribution
- given a sample $x_t \sim p(x)$, uses that sample to produce a new sample $x_{t+1} \sim p(x)$



MCMC visualized

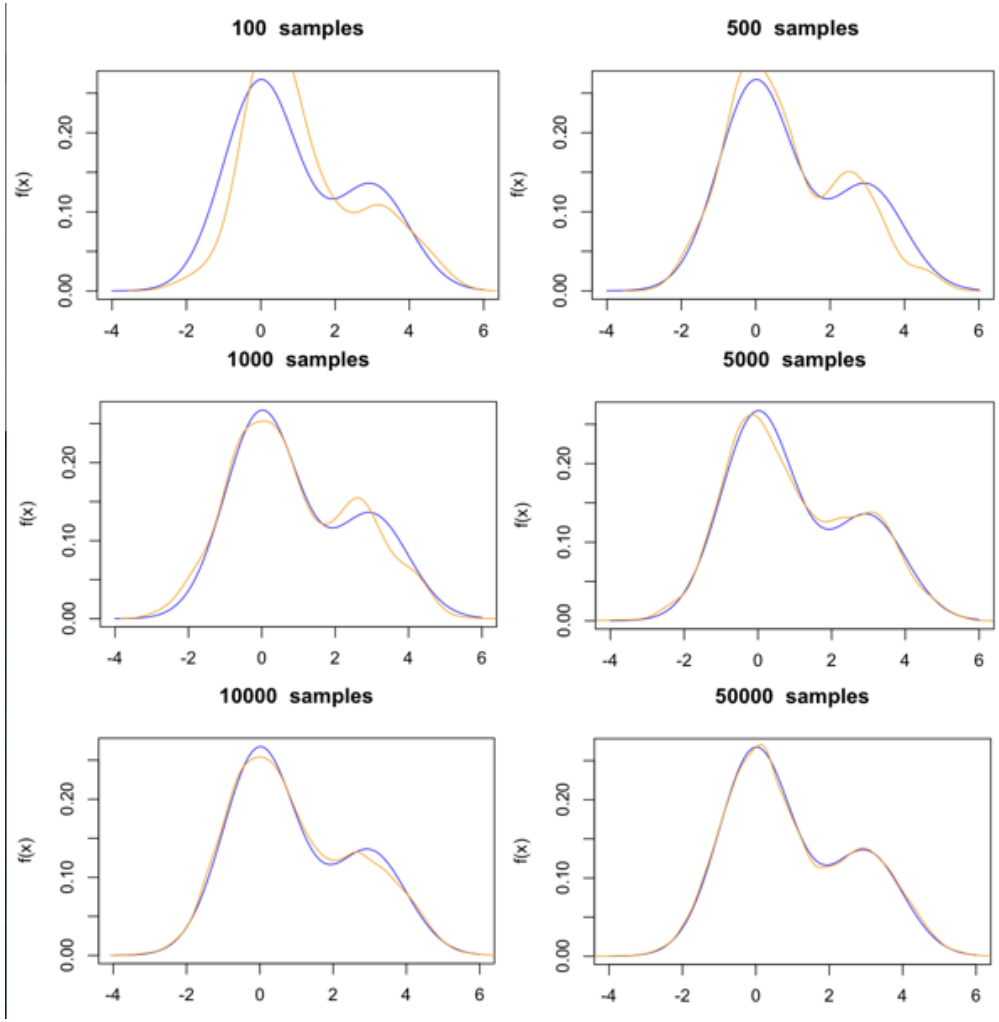


Figure 7: MCMC visualized with different number of samples

Source: By Chdrappi - Using R; FOSS statistical software, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=25674906

Metropolis Hastings

1. $p(x) = \frac{f(x)}{NC}$, NC = Normalizing Constant
2. Start with drawing distributions centered around the previous sample eg.
$$g(x_{t+1}|x_t) = N(x_t, \sigma^2)$$
3. Based on this new candidate, you either accept or reject with probability $A(x_t \rightarrow x_{t+1})$
4. $A(a \rightarrow b) = \min(1, \frac{f(b)}{f(a)} \frac{g(a|b)}{g(b|a)})$
5. sample $u \sim U[0, 1]$
6. if $A(a \rightarrow b) > u$ Then we accept our new state and we draw from a distribution centered around the new point x_{t+1}
7. if $A(a \rightarrow b) \leq u$ Then we don't accept our new state and we draw from the same distribution centered around x_t

Metropolis Hastings visualized

- Suppose g is symmetric, ie. $\frac{g(a|b)}{g(b|a)} = 1$
- $A(a \to b) = \min(1, \frac{p(b)}{p(a)})$, remember f(x) differs from p(x) only by a normalizing constant

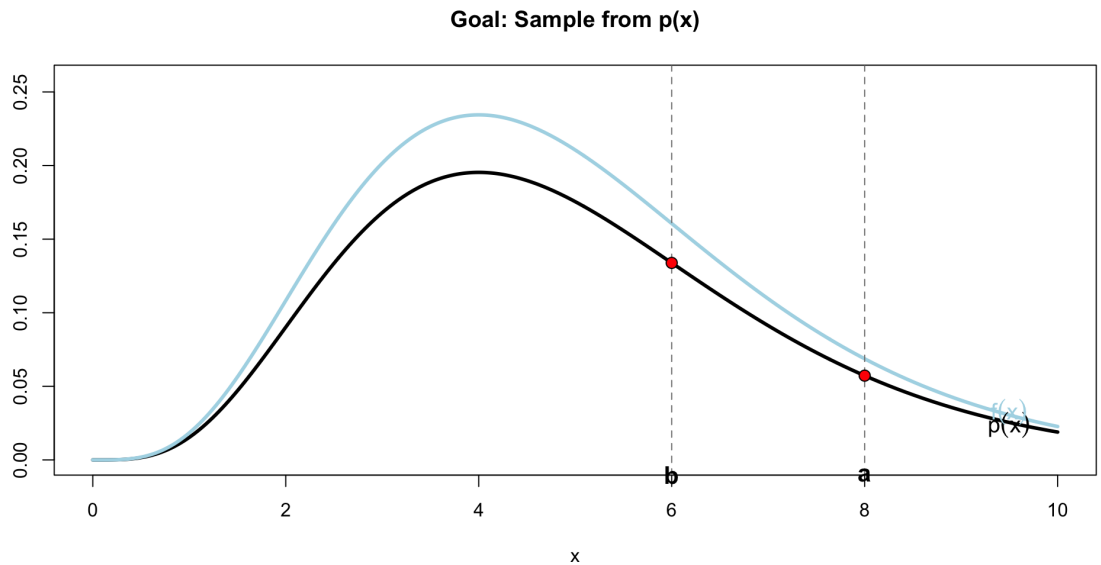


Figure 8: First case of Metropolis Hastings

$A(a \to b) = 1$ since p(b) > p(a) in this scenario

Metropolis Hastings visualized pt2

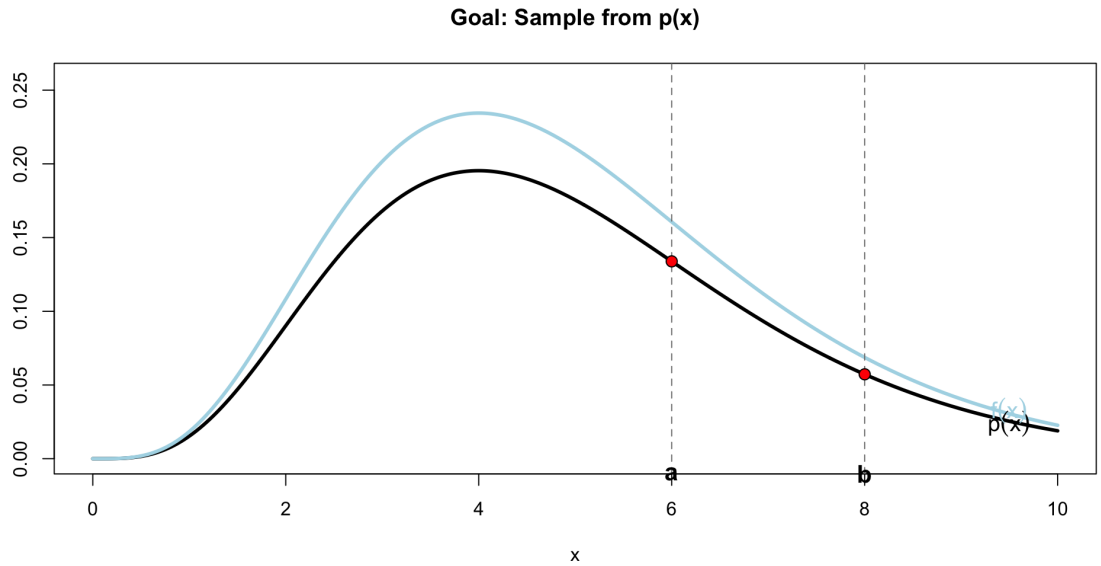


Figure 9: Second case of Metropolis Hastings

$$A(a \to b) = \frac{p(b)}{p(a)} \text{ since p(b)} < \text{p(a) in this scenario}$$

- Whether we go to b here depends on what we sample from the Uniform distribution

Bayesian Setup

We assigned a Cauchy Prior for the predictors for these reasons:

- Andrew Gelman (et al., 2008) suggested a Cauchy(0, 2.5) as a weakly informative prior for logistic regression
- Cauchy distribution is heavy-tailed, so it doesn't heavily penalize larger coefficients.
- "Most effects are small, but I allow for occasional large coefficients."

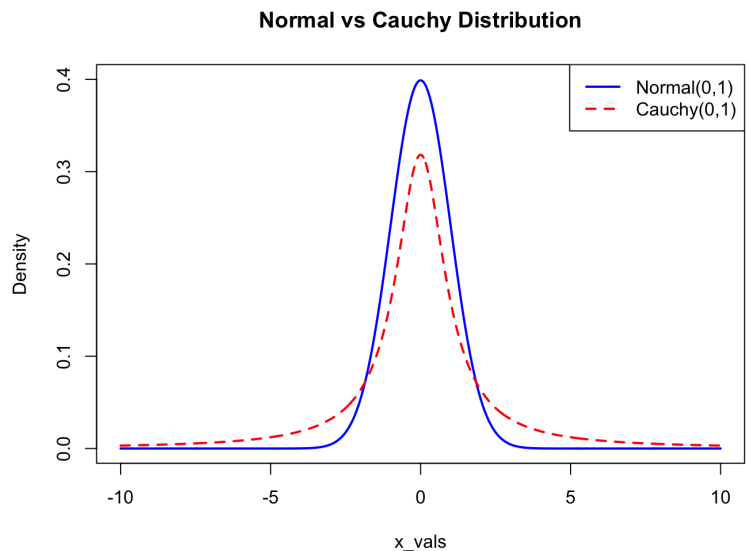


Figure 10: Normal vs. Cauchy distribution

Distribution of predictors

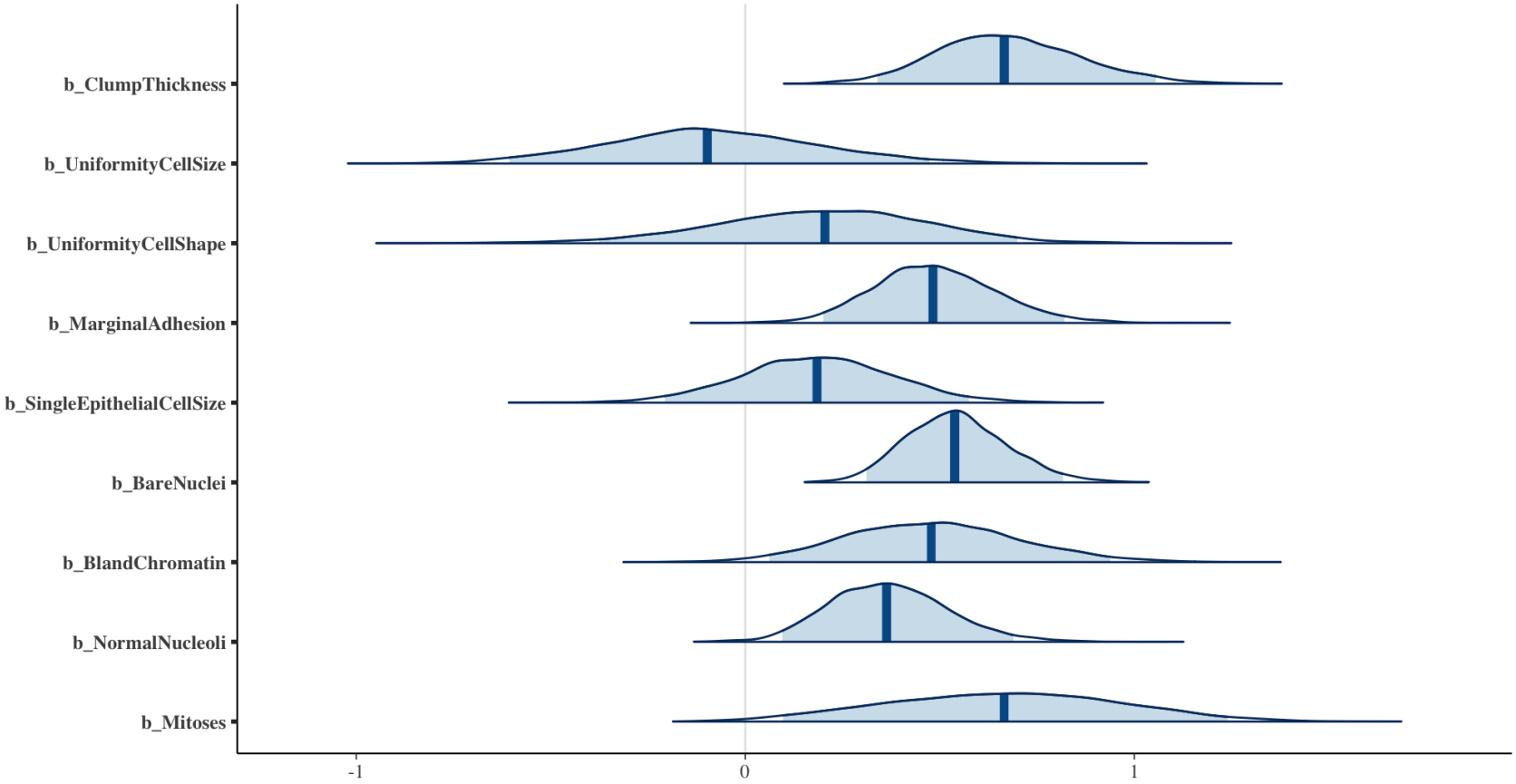


Figure 11: Distribution of predictors of Bayesian model

Hypothesis Testing via. Posterior Probabilities

- Instead of p-values, we compute probabilities:

$$P(\beta_j > 0|\text{data}) > 0$$

- Ex. if clump thickness is significant $P(\beta_1 > 0|\text{data})$
- $P(\beta_1 > 0|\text{data}) = 1$
- Overwhelming evidence that the coefficient for Clump Thickness is positive, ie. the coefficient is significant
- Ex. if UniformityCellShape is significant $P(\beta_3 > 0|\text{data})$
- $P(\beta_3 > 0|\text{data}) = 0.77$
- There is a 77% chance that an increase in UniformityCellShape increases the odds of a Malignant tumor
- Not a significant coefficient

Results

On an 80-20 split with threshold of 0.5, we get the following results from the two models

- Logistic Regression:
  - 97.62% training accuracy, 96.3% testing accuracy
  - Sensitivity: 93.62%, Specificity: 97.73%
- Bayesian Regression:
  - 97.08% training accuracy, 98.52% testing accuracy
  - Sensitivity: 100%, Specificity: 97.73%

In medical settings, it is better to have higher specificity, since false classifying malignant tumor is costly

- Can change threshold to increase specificity

Post Posterior Check

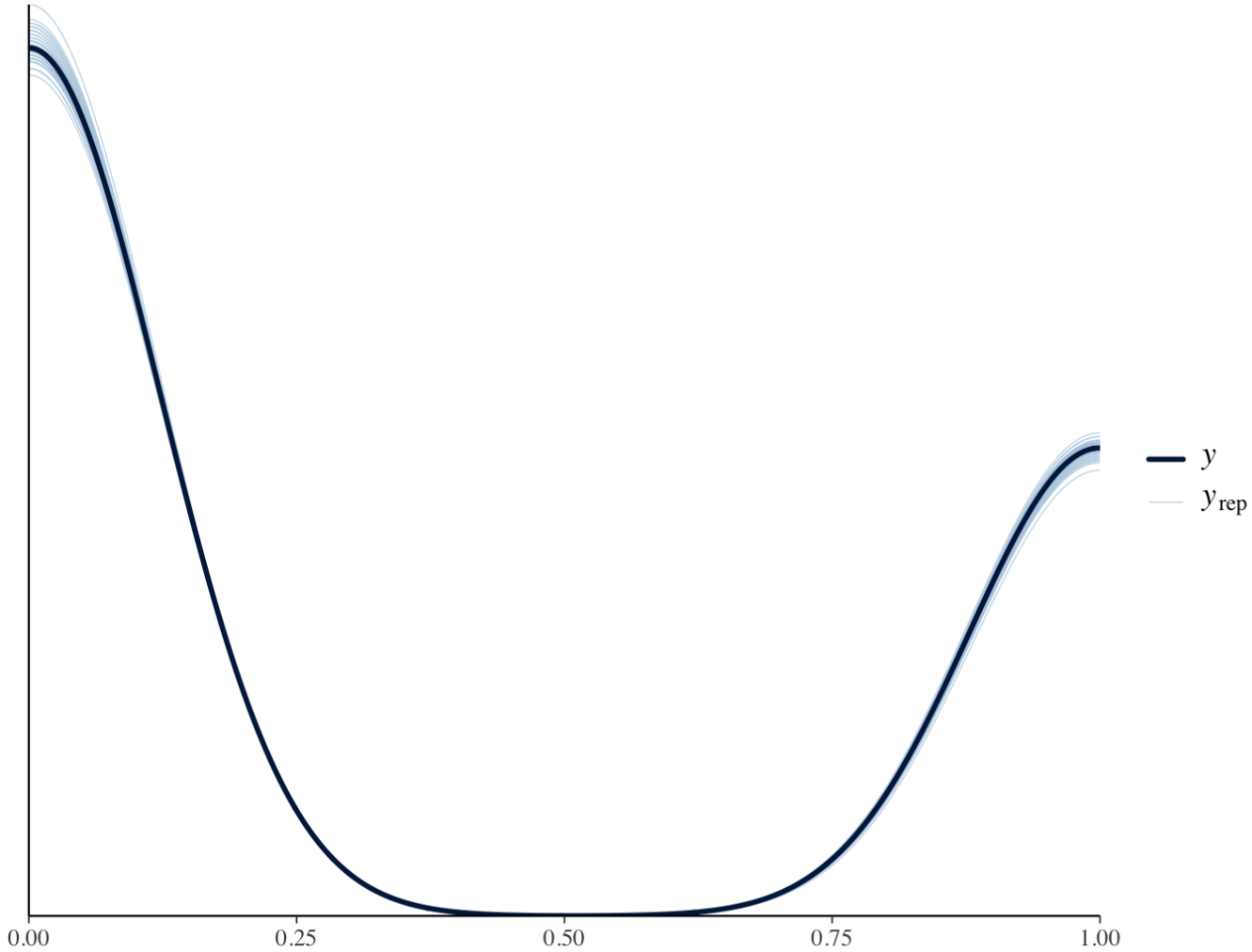


Figure 12: Post Posterior Check 50 draws

Significant predictors

Bayesian LR

- ClumpThickness
- MarginalAdhesion
- BareNuclei
- BlandChromatin
- NormalNucleoi
- Mitoses

Ordinary LR

- ClumpThickness
- UniformityOfCellShape
- MarginalAdhesion
- BareNuclei
- BlandChromatin
- NormalNucleoi

Bayesian vs. Frequentist model

Bayesian LR

- Slower due to MCMC sampling
- Use posterior distribution to understand coefficients and their uncertainty
- More robust (less prone to overfitting)

Regular LR

- Faster, especially with large datasets
- More interpretable and works better with classical inference (p-values, wald tests)
- More prone to overfitting, especially since we added interaction terms

Future directions

- Trying different prior distributions (Jeffreys' Prior, or tighter priors)
- Using K-fold cross validation to optimize the decision threshold for classification
- Dimensionality reduction(PCA), or adding a regularization termi to prevent multicollinearity and also prevent overfitting in the regular LR case
- Analyze how the models' perform across subgroups (age, race, etc.) (if data allows)
- Survival analysis to model progression risk over time instead of classification (if data allows)

References

1. Canada, Public Health Agency of. "Government of Canada." Canada.Ca, / Gouvernement du Canada, 30 Dec. 2024, www.canada.ca/en/public-health/services/chronic-diseases/cancer/breast-cancer.html.
2. "Limitations of Mammograms: How Accurate Are Mammograms?" How Accurate Are Mammograms? — American Cancer Society, www.cancer.org/cancer/types/breast-cancer/screening-tests-and-early-detection/mammograms/limitations-of-mammograms.html. Accessed 17 Mar. 2025.
3. "Survival Rates for Breast Cancer." American Cancer Society, www.cancer.org/cancer/types/breast-cancer/understanding-a-breast-cancer-diagnosis/breast-cancer-survival-rates.html. Accessed 17 Mar. 2025.
4. Eldridge, Lynne. Cancer Cells vs. Normal Cells: How Are They Different? 21 Apr. 2023. verywellHealth, Dotdash Meredith, https://www.verywellhealth.com/thmb/dMIwdIGJt8aJ25ScrDbYciu3J3U=/750x0/illo_normal-cells-cancer-cells-596cdd256f53ba00111a65bb.png. Accessed 23 Mar. 2025.
5. Nahhas, Ramzi W. Introduction to Regression Methods for Public Health Using R. CHAPMAN & HALL CRC, 2025.
6. Chdrappi. Metropolis algorithm convergence example. 1 Apr. 2013. Using R; FOSS Statistical Software, https://commons.wikimedia.org/wiki/File:Metropolis_algorithm_convergence_example.png. Accessed 23 Mar. 2025.
7. Gelman, Andrew, et al. "A weakly informative default prior distribution for logistic and other regression models." The Annals of Applied Statistics, vol. 2, no. 4, 1 Dec. 2008, https://doi.org/10.1214/08-aoas191.