Bayesian Learning

Lecture 5 - Large sample approximations. Classification.

Mattias Villani

Department of Statistics
Stockholm University

mattiasvillani.com @matvil @matvil mattiasvillani

Lecture overview

- **Classification**
- **Normal approximation** of posterior
- **Logistic regression** - demo in R

Bayesian classification

- **Classification: output is a discrete label**.
  - Binary (0-1). Spam/Ham.
  - Multi-class. ($c = 1, 2, ..., C$). Brand choice.
- **Bayesian classification**

$$\underset{c \in \mathcal{C}}{\operatorname{argmax}}\, p(c|\mathbf{x})$$

  where $\mathbf{x} = (x_1, ..., x_p)^\top$ is a covariate/feature vector.
- **Discriminative models** - model $p(c|\mathbf{x})$ directly.
  - Examples: logistic regression, support vector machines.
- **Generative models** - Use Bayes' theorem

$$p(c|\mathbf{x}) \propto p(\mathbf{x}|c)p(c)$$

  with class-conditional distribution $p(\mathbf{x}|c)$ and prior $p(c)$.
  - Examples: discriminant analysis, naive Bayes.

Classification with logistic regression

- Response is assumed to be **binary** ($y = 0$ or 1).
- Example: Spam/Ham. Covariates: \$-symbols, etc.
- **Logistic regression**

$$\Pr(y_i = 1 \mid \boldsymbol{x}_i) = \frac{\exp(\boldsymbol{x}_i^\top \boldsymbol{\beta})}{1 + \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta})}.$$

- **Likelihood**

$$p(\mathbf{y}|\mathbf{X}, \boldsymbol{\beta}) = \prod_{i=1}^{n} \frac{[\exp(\boldsymbol{x}_i^\top \boldsymbol{\beta})]^{y_i}}{1 + \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta})}.$$

- Prior $\beta \sim N(0, \tau^2 I)$. Posterior is non-standard (demo later).
- Alternative: **Probit regression**

$$Pr(y_i = 1|\boldsymbol{x}_i) = \Phi(\boldsymbol{x}_i^\top \boldsymbol{\beta})$$

- **Multi-class** ($c = 1, 2, ..., C$) logistic regression

$$\Pr(y_i = c \mid \boldsymbol{x}_i) = \frac{\exp(\boldsymbol{x}_i^\top \boldsymbol{\beta}_c)}{\sum_{k=1}^{C} \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta}_k)}$$

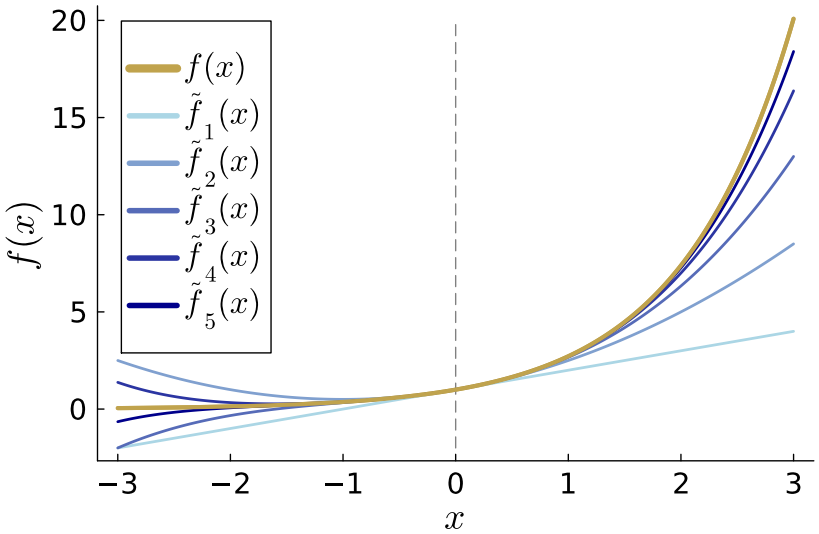
Taylor approximation

- **Taylor approximation** of the function $f(x)$ around $x = a$

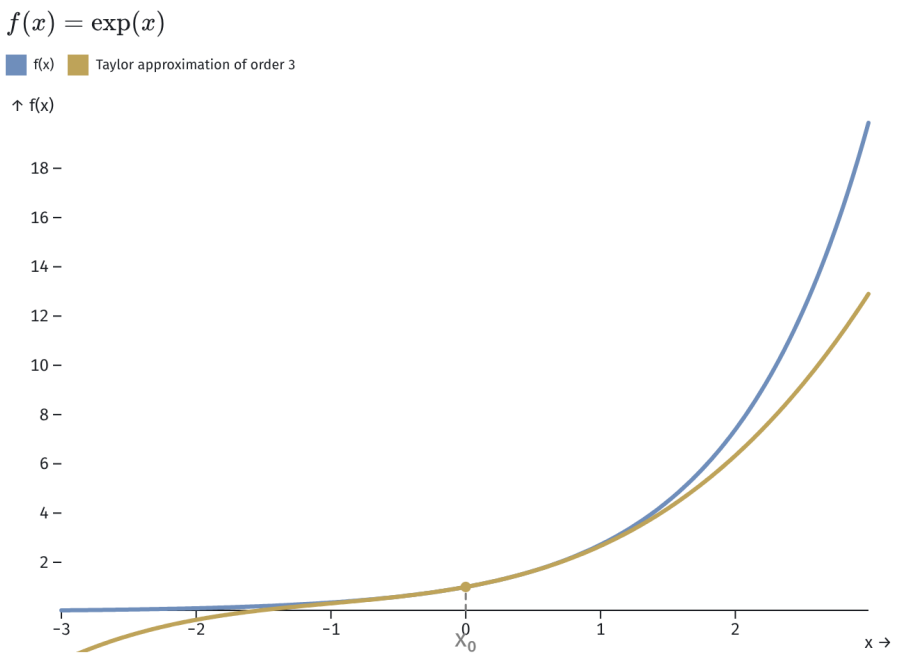
$$f(x) \approx f(a) + \sum_{k=0}^{K} \frac{f^{(k)}(a)}{k!}(x-a)^k$$

- Taylor approximation of $f(x) = \exp(x)$

$$\exp(x) \approx \sum_{k=0}^{K} \frac{x^k}{k!}$$



Interactive - Taylor approximation



Normal approximation of likelihood

- **Taylor expansion of log-likelihood** around the MLE $\theta = \hat{\theta}$:

$$\ln p(\mathbf{x}|\theta) = \ln p(\mathbf{x}|\hat{\theta}) + \frac{\partial \ln p(\mathbf{x}|\theta)}{\partial \theta}|_{\theta=\hat{\theta}}(\theta - \hat{\theta}) + \frac{1}{2!}\frac{\partial^2 \ln p(\mathbf{x}|\theta)}{\partial \theta^2}|_{\theta=\hat{\theta}}(\theta - \hat{\theta})^2 + \dots$$

- Higher order terms (...) negligible in large samples.
- From the definition of the MLE:

$$\frac{\partial \ln p(\mathbf{x}|\theta)}{\partial \theta}|_{\theta=\hat{\theta}} = 0$$

- So, in **large samples**

$$p(\mathbf{x}|\theta) \approx p(\mathbf{x}|\hat{\theta}) \exp\left(-\frac{1}{2} J_{\mathbf{x}}(\hat{\theta})(\theta - \hat{\theta})^2\right)$$

- **Observed information**

$$J_{\mathbf{x}}(\hat{\theta}) = -\frac{\partial^2 \ln p(\mathbf{x}|\theta)}{\partial \theta^2}|_{\theta=\hat{\theta}}$$

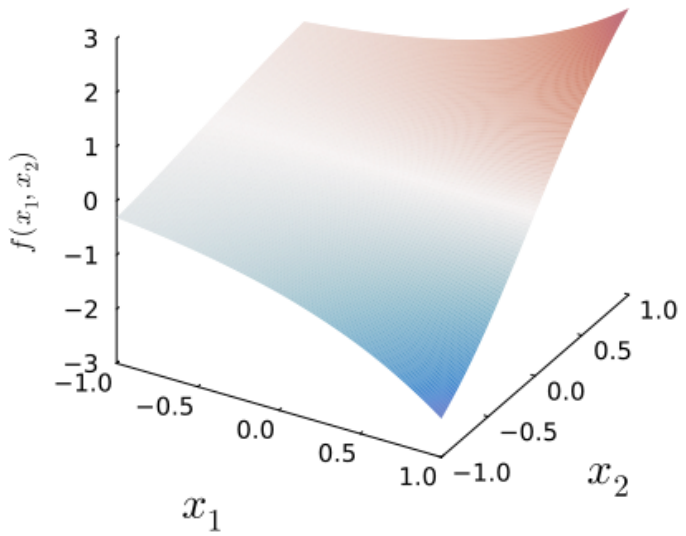
Multi-dimensional Taylor approximation

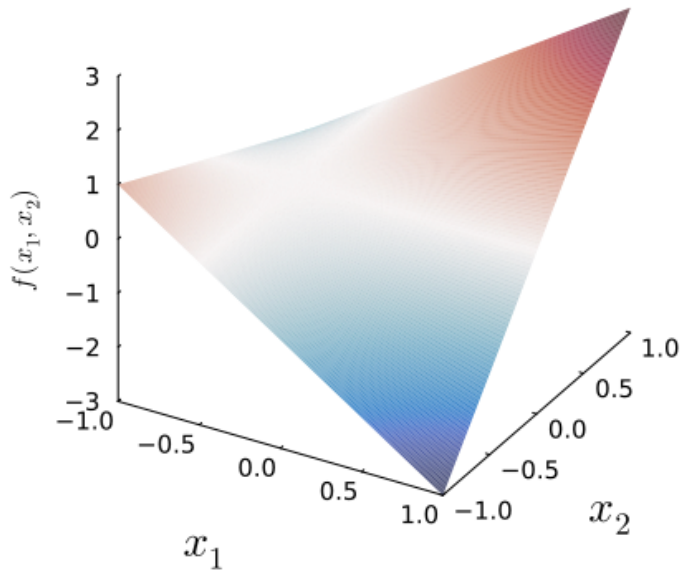
- **Multi-dimensional Taylor approximation** of $f(\boldsymbol{x})$

$$f(\boldsymbol{x}) \approx f(\boldsymbol{a}) + \frac{\partial f(\boldsymbol{x})}{\partial \boldsymbol{x}} (\boldsymbol{x} - \boldsymbol{a}) + \frac{1}{2} (\boldsymbol{x} - \boldsymbol{a})^\top \frac{\partial^2 f(\boldsymbol{x})}{\partial \boldsymbol{x} \partial \boldsymbol{x}^\top} (\boldsymbol{x} - \boldsymbol{a}) + \dots$$

$f(x_1, x_2) = \exp(x_1)\sin(x_2)$

Taylor 2nd order





Multi-dimensional observed information

- Multiparameter **observed information matrix**

$$J_{\mathbf{x}}(\hat{\boldsymbol{\theta}}) = -\frac{\partial^2 \ln p(\mathbf{x}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}\Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}$$

- Example: $\boldsymbol{\theta} = (\theta_1, \theta_2)^\top$

$$\frac{\partial^2 \ln p(\mathbf{x}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} = \begin{pmatrix} \frac{\partial^2 \ln p(\mathbf{x}|\boldsymbol{\theta})}{\partial \theta_1^2} & \frac{\partial^2 \ln p(\mathbf{x}|\boldsymbol{\theta})}{\partial \theta_1 \partial \theta_2} \\ \frac{\partial^2 \ln p(\mathbf{x}|\boldsymbol{\theta})}{\partial \theta_1 \partial \theta_2} & \frac{\partial^2 \ln p(\mathbf{x}|\boldsymbol{\theta})}{\partial \theta_2^2} \end{pmatrix}.$$

Normal posterior approximation

- We can do the same Taylor approximation on log posterior

$$\log p(\boldsymbol{\theta}|\mathbf{x}) = \log p(\mathbf{x}|\boldsymbol{\theta}) + \log p(\boldsymbol{\theta}) - \log p(\boldsymbol{x})$$

- **Approximate normal posterior** in large samples

$$\boldsymbol{\theta}|\mathbf{x} \overset{\text{approx}}{\sim} N\left[\tilde{\boldsymbol{\theta}}, J_{\mathbf{x}}^{-1}(\tilde{\boldsymbol{\theta}})\right]$$

- $\tilde{\boldsymbol{\theta}} = \arg\max p(\boldsymbol{\theta}|\mathbf{x})$ is the posterior mode and
- $J_{\mathbf{x}}^{-1}(\tilde{\boldsymbol{\theta}})$ is now with respect to posterior $\log p(\boldsymbol{\theta}|\mathbf{x})$.
- Likelihood will dominate the prior in large samples so
  - $\tilde{\boldsymbol{\theta}} \approx \hat{\boldsymbol{\theta}}$
  - $J_{\mathbf{x}}^{-1}(\tilde{\boldsymbol{\theta}})$ will be close to the **observed information**.
- Important: sufficient with proportional form

$$\log p(\boldsymbol{\theta}|\mathbf{x}) = \log p(\mathbf{x}|\boldsymbol{\theta}) + \log p(\boldsymbol{\theta})$$

Normal posterior approximation

Normal posterior approximation

The posterior can in large samples be approximated by

$$\boldsymbol{\theta}|\mathbf{y} \overset{\text{a}}{\sim} \text{N}\left(\tilde{\boldsymbol{\theta}}, J^{-1}_{\boldsymbol{\theta},\mathbf{y}}(\tilde{\boldsymbol{\theta}})\right)$$

where $\tilde{\boldsymbol{\theta}}$ is the posterior mode and

$$J_{\tilde{\boldsymbol{\theta}},\mathbf{y}} = -\frac{\partial^2 \ln p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}|_{\boldsymbol{\theta}=\tilde{\boldsymbol{\theta}}}$$

is the $d \times d$ observed information matrix at $\tilde{\boldsymbol{\theta}}$.

Example: gamma posterior

- **Poisson model**: $\theta|y_1, ..., y_n \sim \text{Gamma}(\alpha + \sum_{i=1}^n y_i, \beta + n)$

$$\log p(\theta|y_1, ..., y_n) \propto (\alpha + \sum_{i=1}^n y_i - 1) \log \theta - \theta(\beta + n)$$

- First derivative of log density

$$\frac{\partial \ln p(\theta|\mathbf{y})}{\partial \theta} = \frac{\alpha + \sum_{i=1}^n y_i - 1}{\theta} - (\beta + n)$$

$$\tilde{\theta} = \frac{\alpha + \sum_{i=1}^n y_i - 1}{\beta + n}$$

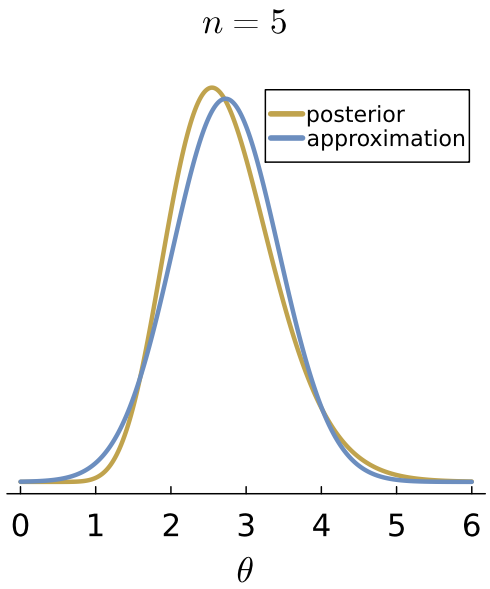
- Second derivative at mode $\tilde{\theta}$

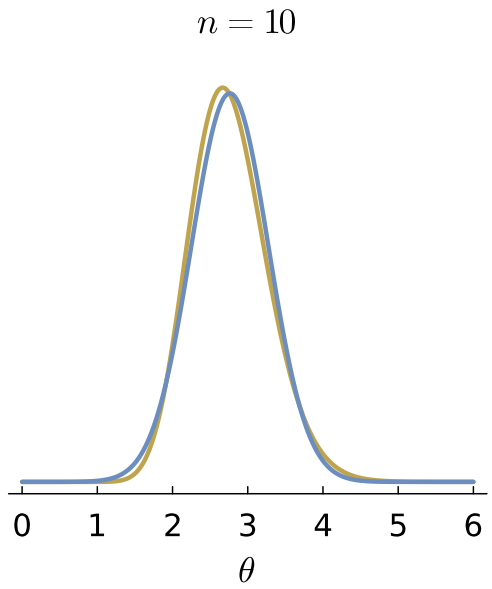
$$\frac{\partial^2 \ln p(\theta|\mathbf{y})}{\partial \theta^2}|_{\theta=\tilde{\theta}} = -\frac{\alpha + \sum_{i=1}^n y_i - 1}{\left(\frac{\alpha+\sum_{i=1}^n y_i-1}{\beta+n}\right)^2} = -\frac{(\beta + n)^2}{\alpha + \sum_{i=1}^n y_i - 1}$$

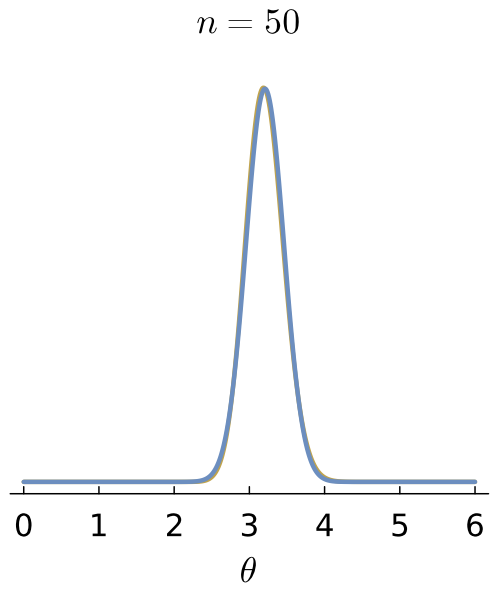
- **Normal approximation**

$$N\left[\frac{\alpha + \sum_{i=1}^n y_i - 1}{\beta + n}, \frac{\alpha + \sum_{i=1}^n y_i - 1}{(\beta + n)^2}\right]$$

Example: gamma posterior for eBay bidders data







Normal approximation of posterior

- $\boldsymbol{\theta}|\mathbf{y} \overset{\text{approx}}{\sim} N\left[\tilde{\boldsymbol{\theta}}, J_{\mathbf{y}}^{-1}(\tilde{\boldsymbol{\theta}})\right]$ works also when θ is a vector.
- How to compute $\tilde{\boldsymbol{\theta}}$ and $J_{\mathbf{y}}(\tilde{\boldsymbol{\theta}})$?
- Standard **optimization routines** may be used. (`optim.r`).
  - **Input**: expression proportional to $\log p(\theta|\mathbf{y})$. Initial values.
  - **Output**: $\log p(\tilde{\boldsymbol{\theta}}|\mathbf{y})$, $\tilde{\boldsymbol{\theta}}$ and Hessian matrix $(-J_{\mathbf{y}}(\tilde{\boldsymbol{\theta}}))$.
- **Automatic differentation** - efficient derivatives on computer.
- **Re-parametrization** may improve normal approximation:
  - If $\theta \geq 0$ use $\phi = \log(\theta)$.
  - If $0 \leq \theta \leq 1$, use $\phi = \log\left(\frac{\theta}{1-\theta}\right)$.

Normal approximation of posterior

- **Heavy tailed approximation**: $\boldsymbol{\theta}|\mathbf{y} \overset{\text{approx}}{\sim} t_v\left[\tilde{\theta}, J_{\mathbf{y}}^{-1}(\tilde{\boldsymbol{\theta}})\right]$ for suitable degrees of freedom v.
- Even if the posterior of θ is approx normal, **interesting functions** of $g(\theta)$ may not be (e.g. predictions).
- But approximate posterior of $g(\theta)$ can be obtained by **simulating** from $N\left[\tilde{\theta}, J_{\mathbf{y}}^{-1}(\tilde{\theta})\right]$.

Bayesian logistic regression

- **Logistic regression**

$$\Pr(y_i = 1 \mid \boldsymbol{x}_i) = \frac{\exp(\boldsymbol{x}_i^\top \boldsymbol{\beta})}{1 + \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta})}.$$

- **Odds**

$$\text{Odds}(\boldsymbol{x}) \equiv \frac{\Pr(\text{y} = 1|\boldsymbol{x})}{\Pr(\text{y} = 0|\boldsymbol{x})} = \exp(\boldsymbol{x}^\top \boldsymbol{\beta}).$$

- **Odds ratio**

$$OR_j = \frac{Odds(x_1, \ldots, x_j + 1, \ldots, x_p)}{Odds(x_1, \ldots, x_j, \ldots, x_p)} = \exp(\beta_j)$$

- **Likelihood**

$$p(\mathbf{y}|\mathbf{X}, \boldsymbol{\beta}) = \prod_{i=1}^{n} \frac{[\exp(\boldsymbol{x}_i^\top \boldsymbol{\beta})]^{y_i}}{1 + \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta})}.$$

- **Normal approximation**:

$$\boldsymbol{\beta}|\boldsymbol{y} \sim N\left(\tilde{\boldsymbol{\beta}}, J_{\mathbf{y}}^{-1}(\tilde{\boldsymbol{\beta}})\right).$$

Logistic regression - who survived the Titanic?

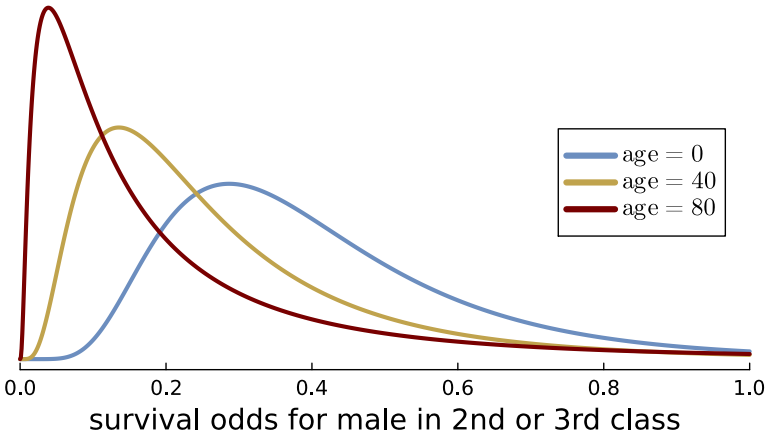
- Prior

$$\boldsymbol{\beta} \sim N(\boldsymbol{\mu}, \boldsymbol{\Omega})$$

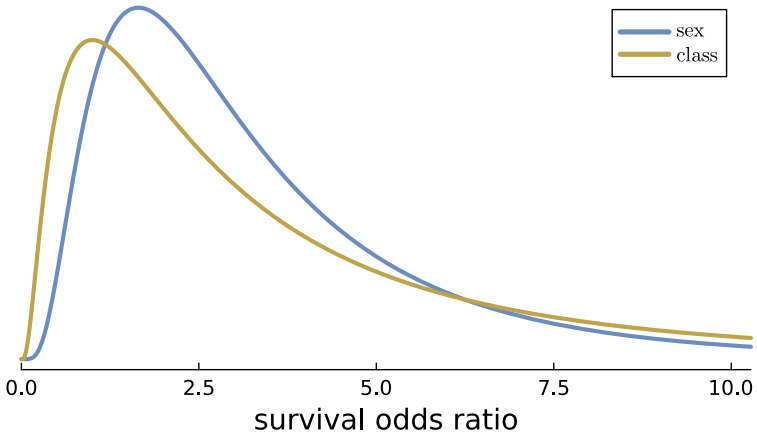
with

$$\boldsymbol{\mu} = (-1, -1/80, 1, 1)^\top \qquad \boldsymbol{\Omega} = \begin{pmatrix} 0.25 & 0 & 0 & 0 \\ 0 & 1/(80^2) & 0 & 0 \\ 0 & 0 & 0.5 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

age



sex and class

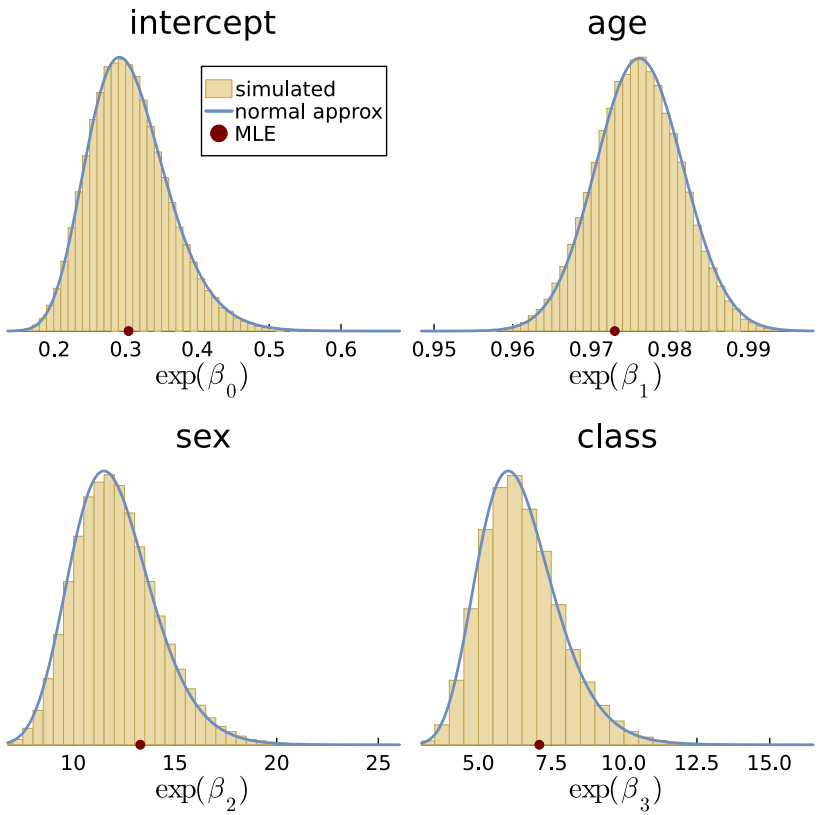


Logistic regression - who survived the Titanic?

- **Normal posterior approximation**

$$\boldsymbol{\beta}|\boldsymbol{y} \sim N\left(\tilde{\boldsymbol{\beta}}, J_{\mathbf{y}}^{-1}(\tilde{\boldsymbol{\beta}})\right).$$

- Means that the posterior of each β_j is univariate normal.
- Marginal posterior for each $\exp(\beta_j)$ **is LogNormal**.



Logistic regression - who survived the Titanic?

- Comparison with non-informative prior $\boldsymbol{\beta} \sim N(\mathbf{0}, 10^2 \boldsymbol{I}_p)$.