STAT 425: Introduction to Bayesian Analysis

Marina Vannucci

Rice University, USA

Fall 2017
Part 3: Hierarchical and Linear Models

- Hierarchical models
- Linear regression models
- Generalized linear models (logistic and Poisson)
Linear regression models - Outline

- Regression models
- Variable selection
- Hierarchical linear mixed models
Regression Models

- Conjugate priors & prediction
- Example
- Data augmentation
The question of interest is to investigate how a response variable $y$ varies as a function of explanatory variables, $x_1, \ldots, x_p$.

$$y_i = x_i \beta + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2), \quad i = 1, \ldots, n$$

$$E[y_i | \beta, X] = \beta_1 x_{i1} + \ldots + \beta_k x_{ip} \quad \text{and} \quad Var(y_i | \beta, \sigma^2, X) = \sigma^2$$

- $y_i$ is the response
- $x_{i1}, \ldots, x_{ip}$ are the regressors for the $i$-th individual
- $\beta_1, \ldots, \beta_p$ are unknown regression parameters.

In matrix notation, the model can be written as

$$y | \beta, \sigma^2, X \sim \mathcal{N}_n(X \beta, \sigma^2 I)$$

Most fundamental of all statistical models; encompasses ANOVA, regression, ANCOVA, random and mixed effects models.
The parameter vector is $\theta = (\beta_1, \ldots, \beta_p, \sigma^2)$.

The statistical inference problem is to estimate the parameters $\theta$, conditional on $X$ and $y$.

Under a standard noninformative prior distribution, the Bayesian estimates and standard errors coincide with the frequentist results.

Even in the noninformative case, posterior simulations are useful for predictive inference and model checking.
Classical Estimation

- Classical unbiased estimates of the regression parameters $\beta$ and $\sigma^2$ are

$$\hat{\beta} = (X^T X)^{-1} X^T Y; \hat{\sigma}^2 = \frac{1}{n-p}(Y - X^T \hat{\beta})^T (Y - X^T \hat{\beta})$$

also called ordinary least squares estimates (OLS).

- The predicted value of $Y$ is given by

$$Y = X \hat{\beta} = MY \quad \text{where} \quad M = X (X^T X)^{-1} X^T$$

$M$ is called the projector matrix and

$$(Y - X^T \hat{\beta})^T (Y - X^T \hat{\beta}) = Y^T (I - M) Y$$

the model residual

- Assume $n > p$ and $X$ of rank $p$ (columns of $X$ linearly independent)
Standard noninformative prior distribution

A convenient noninformative prior distribution is

\[ p(\beta, \sigma^2) \propto \frac{1}{\sigma^2} \]

which is the Jeffreys’s prior on \( \beta \) and \( \sigma^2 \) (proof: exercise). Improper prior but with valid posteriors. As with the normal distribution, we factor the joint posterior distribution for \( \beta \) and \( \sigma^2 \) as

\[ p(\beta, \sigma^2 | y, X) = p(\beta | \sigma^2, y, X) \, p(\sigma^2 | y, X) \]

where

\[ \beta | \sigma^2, y, X \sim \mathcal{N}(\hat{\beta}, (X^T X)^{-1} \sigma^2), \text{ with } \hat{\beta} = (X^T X)^{-1} X^T y \]

\[ \sigma^2 | y, X \sim \text{Inv-Gamma} \left( \frac{n - k}{2}, \frac{S^2}{2} \right), \text{ with } S^2 = (y - X \hat{\beta})^T (y - X \hat{\beta}) \]
• Striking similarity with classical result.

• The standard non-Bayesian estimates of \( \beta \) and \( \sigma^2 \) are \( \hat{\beta} \) and \( s^2 \), respectively, as defined above, and \( \text{Var}(\hat{\beta}) = (X^T X)^{-1} \sigma^2 \).

• The distribution of \( \sigma^2 \) is also characterized as \( (n - p)S^2 / \sigma^2 \) following a \( \chi^2 \) distribution.

• The marginal posterior distribution of \( \beta | y, X \), averaging over \( \sigma^2 \), is multivariate \( t \) with \( n - p \) degrees of freedom.

• The posterior distribution is proper as long as \( n > p \) and the rank of \( X \) equals \( p \), i.e., the columns of \( X \) must be linearly independent.

• For posterior inference sample draws on \( \sigma^2 \) first, then \( \beta \), to sample from joint posterior of \( \beta \) and \( \sigma^2 \) (or do Gibbs sampler).
Posterior predictive distribution for new data

- Suppose we apply the regression model to a new set of data, for which we have observed the matrix $\tilde{X}$ of explanatory variables, and we wish to predict the outcome $\tilde{y}$.
- If $\beta$ and $\sigma^2$ were known exactly, the vector $\tilde{y}$ would have a normal distribution with mean $\tilde{X}\beta$ and variance matrix $\sigma^2 I$. Instead, our current knowledge of $\beta$ and $\sigma^2$ is summarized by our posterior distribution.
- **Posterior predictive simulation:** First draw $(\beta, \sigma^2)$ from their joint posterior distribution, then draw $\tilde{y} \sim N(\tilde{X}\hat{\beta}, \sigma^2 I)$.
- **Analytic form of the posterior predictive distribution:** $p(\tilde{y}|y)$ is multivariate $t$ with location $\tilde{X}\hat{\beta}$, square scale matrix $s^2(I + \tilde{X}(X^TX)^{-1}\tilde{X})$, and $n - p$ degrees of freedom.
Suppose one simulates many samples $\tilde{y}_1, \ldots, \tilde{y}_n$ from the posterior predictive distribution conditional on the same covariate vectors, $x_1, \ldots, x_n$ used to simulate the data.

To judge if a particular response value $y_i$ is consistent with the fitted model, one looks at the position of $y_i$ relative to the histogram of simulated values of $\tilde{y}_i$ from the corresponding predictive distribution.

If $y_i$ is in the tail of the distribution, that indicates that this observation is a potential outlier.
A conjugate class of priors is specified as

\[
\begin{align*}
\beta | \sigma^2 & \sim \mathcal{N}_p(\beta_0, \sigma^2 V) \\
\sigma^2 & \sim \text{IG}(\nu_0/2, \sigma_0^2/2),
\end{align*}
\]

where $\beta_0$ is a $p$-dimensional vector and $V$ is a $p \times p$ positive definite symmetric matrix. We call this the Normal-Inverse-Gamma (NIG) prior and denote it as $NIG(\beta_0, V, \nu_0/2, \sigma_0^2/2)$, which defines the joint probability distribution of the vector $\beta$ and the scalar $\sigma^2$. 
This leads to the following posterior distribution

\[
\beta | \sigma^2, X, Y \sim \mathcal{N}(m, V)
\]

\[
\sigma^2 | X, Y \sim IG(a^*, b^*)
\]

where

\[
m = (V_0^{-1} + X^T X)^{-1} (V_0^{-1} \beta_0 + X^T Y)
\]

\[
V = (V_0^{-1} + X^T X)^{-1}
\]

\[
a^* = (\nu_0 + n)/2
\]

\[
b^* = \frac{1}{2} \sigma_0^2 + \frac{1}{2} [\beta_0^T V_0^{-1} \beta_0 + Y^T Y - m V^{-1} m]
\]

Furthermore, the marginal posterior distribution of \( \beta \) can be obtained by integrating out \( \sigma^2 \) from the NIG joint posterior, resulting in the marginal posterior distribution being a multivariate t-density with degrees of freedom \( 2a^* \), mean \( m \) and covariance \( (b^*/a^*)V \).
Remarks

- Bayesian prediction same as before
- The non-informative priors can be obtained by letting \( V_0^{-1} \to 0 \) (i.e. the null matrix, essentially no prior information) and \( \nu_0/2 \to -p \) and \( \sigma_0^2 \to 0 \).
- Informative vs noninformative
- A middle ground solution between informative and non-informative priors can be obtained by setting, \( V_0 = g(X^T X)^{-1} \) and \( p(\sigma^2) \sim \sigma^{-2} \) which is commonly referred to as the Zellner’s g-prior: a (conditional) Gaussian prior on \( \beta \) and an improper prior on \( \sigma^2 \). Note that it appears the prior is data-dependent but is not since the whole model is conditional on \( X \). The constant \( g \) here is interpreted as a measure of the amount of information available in the prior relative to the sample. For example, setting \( 1/c = 0.5 \), gives the prior the same weight as 50% of the sample.
- Specification of \( V_0 \) as diagonal/block diagonal/completely unstructured
Zellner’s $g$-prior

\[
\begin{align*}
\beta | \sigma^2 & \sim \mathcal{N}_k(\beta_0, g\sigma^2(X^T X)^{-1}) \\
\sigma^2 & \sim \text{Inv-Gamma} \left( \frac{\nu_0}{2}, \frac{\sigma_0^2}{2} \right)
\end{align*}
\]

- The constant $g$ reflects the amount of information in the data relative to the prior; if one believes strongly in the prior guess, one would choose a small value for $g$.
- A nice feature of the $g$-prior is that the posterior distribution has a relatively simple functional form:

\[
p(\beta, \sigma^2 | y, X) = p(\beta | y, \sigma^2)p(\sigma^2 | y)
\]

where ...
where \( \beta | y, X, \sigma^2 \sim \mathcal{N}_p(m, V) \)

\[
m = \frac{g}{g + 1} \left( \frac{\beta_0}{g} + (X^T X)^{-1} X^T y \right)
\]

\[
V = \frac{g}{g + 1} \sigma^2 (X^T X)^{-1}
\]

and \( \sigma^2 | y, X, \sim \text{Inv-Gamma} \left( \frac{\nu_0 + n}{2}, \frac{1}{2} \nu_0 \sigma_0^2 + S_g^2 \right) \), where

\[
S_g^2 = y^T \left( I - \frac{g}{g + 1} X (X^T X)^{-1} X^T \right) y
\]
Using semi-conjugate prior distribution

\[ \beta \sim \mathcal{N}_p(\beta_0, \Sigma_0) \]

\[ \sigma^2 \sim \text{Inv-Gamma} \left( \frac{\nu_0}{2}, \frac{\sigma_0^2}{2} \right) \]

The joint posterior distribution \( p(\beta, \sigma^2 | y, X) \) can be approximated via Gibbs sampling.
Given current values \( \{ \beta(t), \sigma^2(t) \} \), new values can be generated by:

1. sampling \( \beta^{(t+1)} \) from its full conditional \( \beta|y, X, \sigma^2(t) \sim \mathcal{N}_p(m(t), V(t)) \), where
   \[
   V(t) = \left( \Sigma_0^{-1} + \frac{1}{\sigma^2(t)} X^T X \right)^{-1} \quad m(t) = V(t) \left( \Sigma_0^{-1} \beta_0 + \frac{1}{\sigma^2(t)} X^T y \right)
   \]

2. sampling \( \sigma^2(t+1) \) from its full conditional
   \( \sigma^2|y, X, \beta^{(t)} \sim \text{Inv-Gamma} \left( \frac{\nu_0 + n}{2}, \frac{\nu_0 \sigma^2_0 + S^2(t)}{2} \right) \), where
   \[
   S^2(t) = (y - X \beta^{(t+1)})^T (y - X \beta^{(t+1)})
   \]
Recap on Linear Regression

An approach for modeling the linear relationship between a dependent variable and one or more explanatory variables

- Simple linear regression
  \[ Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad \epsilon_i \sim \text{N}(0, \sigma^2) \]

- Multiple linear regression
  \[ Y_i = \beta^T x_i + \epsilon_i = \beta_0 + \beta_1 x_{i,1} + \ldots + \beta_p x_{i,p} + \epsilon_i \]

- Multivariate linear regression
  \[ Y_{i,1} = x_i^T \beta_1 + \epsilon_{i,1} = \beta_{1,0} + \beta_{1,1} x_{i,1} + \ldots + \beta_{1,p} x_{i,p} + \epsilon_{i,1} \]
  \[ \vdots \]
  \[ Y_{i,m} = x_i^T \beta_m + \epsilon_{i,m} = \beta_{m,0} + \beta_{m,1} x_{i,1} + \ldots + \beta_{m,p} x_{i,p} + \epsilon_{i,m} \]

An extremely powerful data analysis tool

- Parameter estimation
- Prediction
Multiple Linear Regression

Linear regression model when $m = 1$ and $p \geq 2$

- **Model 1**

  $$Y_i = \beta_0 + \beta_1 x_{i,1} + \ldots + \beta_p x_{i,p} + \epsilon_i, \quad i = 1, \ldots, n$$

- **Model 2**

  $$Y_i = \mathbf{\beta}^T \mathbf{x}_i + \epsilon_i, \quad i = 1, \ldots, n,$$
  where $\mathbf{\beta} = (\beta_0, \ldots, \beta_p)^T$ and $\mathbf{x}_i = (1, x_{i,1}, \ldots, x_{i,p})^T$

- **Model 3**

  $$\mathbf{Y} = \mathbf{X}\mathbf{\beta} + \mathbf{\epsilon}, \quad \mathbf{\epsilon} \sim \text{MN}(\mathbf{0}, \sigma^2 \mathbf{I})$$
  where $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$, $\mathbf{X} = (\mathbf{x}_1^T, \ldots, \mathbf{x}_n^T)^T$, and $\mathbf{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^T$
Bayesian Estimation

\[ Y \sim \text{MN}(X\beta, \sigma^2 I_p) \]

- Data: \( y \) and \( X \)
- Parameters: \( \beta \) and \( \sigma^2 \)
- Likelihood:
  \[
  f(y|X, \beta, \sigma^2) = (2\pi\sigma^2)^{-\frac{n}{2}} \exp \left( -\frac{1}{2\sigma^2}(y - X\beta)^T(y - X\beta) \right)
  \]
- Prior: \( \pi(\beta, \sigma^2) \)
- Posterior: \( \pi(\beta, \sigma^2|y, X) \)
- Predictive: \( \pi(Y_f|y, X, x_f) = \int f(Y_f|x_f, \beta, \sigma^2)\pi(\beta, \sigma^2|y, X)d\beta d\sigma^2 \)
Bayesian Models with Jeffreys Prior

- Prior: $\pi(\beta, \sigma^2) = \pi(\sigma^2)$

  \[ \pi(\beta, \sigma^2) \propto \sigma^{-2} \]

- Inference in closed form:

  \[ \beta | y, X, \sigma^2 \sim MN \left( (X^T X)^{-1} X^T y, (X^T X / \sigma^2)^{-1} \right) \]

  \[ \sigma^2 | y, X \sim IG \left( \frac{n - p}{2}, \frac{(y - X \hat{\beta})^T (y - X \hat{\beta})}{2} \right) \]

- Alternatively, Gibbs sampler to approximate the posterior

  \[ \sigma^2 | y, X, \beta \sim IG \left( \frac{n - p}{2}, \frac{(y - X \beta)^T (y - X \beta)}{2} \right) \]
Bayesian Models with Conjugate Prior

- Prior: \( \pi(\beta, \sigma^2) = \pi(\beta|\sigma^2)\pi(\sigma^2) \)
  \[
  \beta|\sigma^2 \sim \text{MN}(\beta_0, \sigma^2 V_0) \\
  \sigma^2 \sim \text{IG}\left(\frac{\nu_0}{2}, \frac{\nu_0\sigma_0^2}{2}\right)
  \]

- Conditionals:
  \[
  \beta|y, X, \sigma^2 \sim \text{MN}(m, \sigma^2 V) \\
  \sigma^2|y, X \sim \text{IG}(a, b),
  \]
  where \( V = (V_0^{-1} + X^TX)^{-1}, m = V(V_0^{-1}\beta_0 + X^Ty), a = \frac{\nu_0 + n}{2}, \)
  and \( b = \frac{\nu_0\sigma_0^2}{2} + \frac{\beta_0^TV_0^{-1}\beta_0 + y^Ty - mV^{-1}m}{2} \)

- Alternatively, Gibbs sampler to approximate the posterior

- Hyperparameter setting:
  \[
  \beta_0 = \hat{\beta}_{\text{ols}}, V_0 = g(X^TX)^{-1}, \nu_0 = 1, \sigma_0^2 = \hat{\sigma}_{\text{ols}}^2.
  \]
Bayesian Models with g-prior

- Prior: \( \pi(\beta, \sigma^2) = \pi(\beta|\sigma^2)\pi(\sigma^2) \)

  \[ \beta|\sigma^2 \sim \text{MN}(0, g(XX^T/\sigma^2)^{-1}) \]

  \[ \sigma^2 \sim \text{IG}\left(\frac{\nu_0}{2}, \frac{\nu_0\sigma_0^2}{2}\right) \]

- Conditionals:

  \[ \beta|y, X, \sigma^2 \sim \text{MN}\left(\frac{g}{g+1}(XX^T)^{-1}X^Ty, \frac{g}{g+1}(XX^T/\sigma^2)^{-1}\right) \]

  \[ \sigma^2|y, X \sim \text{IG}\left(\frac{\nu_0+n}{2}, \frac{\nu_0\sigma_0^2+y^T(I - \frac{g}{g+1}(XX^T)^{-1}X^T)y}{2}\right) \]

- Alternatively, Gibbs sampler to approximate the posterior

- Hyperparameter setting:

  \[ g = n, \nu_0 = 1, \sigma_0^2 = \hat{\sigma}_{ols}^2. \]
Bayesian Models with Semi-conjugate Prior

- Prior: \( \pi(\beta, \sigma^2) = \pi(\beta)\pi(\sigma^2) \)

\[ \beta \sim MN(\beta_0, \Sigma_0) \]
\[ \sigma^2 \sim IG\left(\frac{\nu_0}{2}, \frac{\nu_0\sigma_0^2}{2}\right) \]

- Conditionals:

\[ \beta | y, X, \sigma^2 \sim MN\left(\left(\Sigma_0^{-1} + X^T X / \sigma^2\right)^{-1} \left(\Sigma_0^{-1} \beta_0 + X^T y / \sigma^2\right), \left(\Sigma_0^{-1} + X^T X / \sigma^2\right)^{-1}\right) \]
\[ \sigma^2 | y, X, \beta \sim IG\left(\frac{\nu_0 + n}{2}, \frac{\nu_0\sigma_0^2 + (y - X\beta)^T (y - X\beta)}{2}\right) \]

- Gibbs sampler to approximate the posterior

- Hyperparameter setting:

\[ \beta_0 = \hat{\beta}_{\text{ols}}, \Sigma_0 = n(\mathbf{X}^T \mathbf{X} / \hat{\sigma}_{\text{ols}}^2)^{-1}, \nu_0 = 1, \sigma_0^2 = \hat{\sigma}_{\text{ols}}^2. \]
Example: Bird species

Measurements on breeding pairs of land-bird species were collected from 16 islands around Britain over the course of several decades. The dataset birdextint.txt contains the following variables for each species:

- **TIME**: the average time of extinction of the species on the island where it appeared
- **NESTING**: the average number of nesting pairs
- **SIZE**: the size of the species (0=small or 1=large)
- **STATUS**: the migratory status of the species (0=migrant or 1=resident)

The objective is to fit a model that relates the time of extinction of the bird species to the covariates.
bird = read.table("birdextinct.txt",header=T,sep="\t")
attach(bird)
hist(TIME)

The distribution of the outcome variable, TIME, is strongly right-skewed. Let’s transform it to the log-scale:

LOGTIME = log(TIME)
hist(LOGTIME)
Let us look at the relationship between LOGTIME and the three predictor variables.

- There is a positive linear relationship between the average number of nesting pairs and the time to extinction.
- There are a few outliers that stand out for the general pattern.
- For the categorical variables, we use the function `jitter` in R to see overlapping points.
The regression model is given by:

$$\text{LOGTIME}_i = \beta_0 + \beta_1 \text{NESTING}_i + \beta_2 \text{SIZE}_i + \beta_3 \text{STATUS}_i + \varepsilon_i$$

where the two categorical variables are represented by binary indicator variables.

Let us first fit the standard least-squares model via the function `lm` in R:

```r
fit = lm(LOGTIME ~ NESTING + as.factor(SIZE) + as.factor(STATUS), data=bird)
summary(fit)
```

Coefficients:

|                     | Estimate | Std. Error | t value | Pr(>|t|) |
|---------------------|----------|------------|---------|----------|
| (Intercept)         | 0.43087  | 0.20706    | 2.081   | 0.041870 * |
| NESTING             | 0.26501  | 0.03679    | 7.203   | 1.33e-09 *** |
| as.factor(SIZE)1    | -0.65220 | 0.16667    | -3.913  | 0.000242 *** |
| as.factor(STATUS)1  | 0.50417  | 0.18263    | 2.761   | 0.007712 ** |
Let us now perform Bayesian inference. We can sample \((\sigma^2, \beta)\) from the joint posterior distribution by

- first drawing \(\sigma^2\) from the inverse-gamma\(((n - k)/2, S^2/2)\) density
- then simulating the vector \(\beta\) from the multivariate normal density with mean \(\hat{\beta}\) and variance-covariance matrix \((X^T X)^{-1}\sigma^2\).

```r
y = LOGTIME
n = length(y)
x = as.matrix(cbind(rep(1, n), bird[, 3:5]))
Vb = solve(t(x) %*% x)
betahat = Vb %*% t(x) %*% y
S2 = t(y - x %*% betahat) %*% (y - x %*% betahat)
library(LearnBayes)
T = 10000; k = dim(x)[2]
sigma2 = rigamma(T, (n - k)/2, S2/2)
beta = matrix(NA, T, k)
for(i in 1:T) {
    beta[i,] = rmnorm(1, betahat, sigma2[i] * Vb)
}
```

Marina Vannucci (Rice University, USA) Bayesian Analysis (Part 3) Fall 2017 30 / 40
Let us look at the distributions of the simulated posterior draws:

- **Nesting**
  - $\beta_1$

- **Size**
  - $\beta_2$

- **Status**
  - $\beta_3$

- **Variance**
  - $\sigma^2$
Below are some Monte Carlo summary statistics for each parameter:

```r
apply(beta, 2, mean)
[1] 0.4307409 0.2651380 -0.6522922 0.5046479
```

```r
apply(beta, 2, quantile, c(0.025, 0.5, 0.975))
[,1] [,2] [,3] [,4
2.5% 0.01606187 0.1913462 -0.9872335 0.1387419
50% 0.43047972 0.2651018 -0.6513544 0.5053110
97.5% 0.84611036 0.3388107 -0.3179631 0.8671088
```

```r
quantile(sigma2, c(.025, 0.5, 0.975))
  2.5% 50% 97.5%
0.3042492 0.4304378 0.6357527
```

Since we used a non-informative prior, the posterior summaries are equivalent to the ordinary regression estimates.
Next, suppose we are interested in estimating the mean log extinction time for four new sets of observations with the following covariates:

<table>
<thead>
<tr>
<th>Covariate set</th>
<th>Nesting pairs</th>
<th>Size</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>4</td>
<td>small</td>
<td>migrant</td>
</tr>
<tr>
<td>B</td>
<td>4</td>
<td>small</td>
<td>resident</td>
</tr>
<tr>
<td>C</td>
<td>4</td>
<td>large</td>
<td>migrant</td>
</tr>
<tr>
<td>D</td>
<td>4</td>
<td>large</td>
<td>resident</td>
</tr>
</tbody>
</table>
For the classical approach, we can use the R function `predict`:

```r
predict(fit, data.frame(NESTING=rep(4,4),
                        SIZE=c(rep(0,2),rep(1,2)),
                        STATUS=rep(c(0,1),2)), interval="prediction")
```

<table>
<thead>
<tr>
<th>fit</th>
<th>lwr</th>
<th>upr</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.4909277</td>
<td>0.13467449</td>
<td>2.847181</td>
</tr>
<tr>
<td>1.9950931</td>
<td>0.66191816</td>
<td>3.328268</td>
</tr>
<tr>
<td>0.8387294</td>
<td>-0.51147467</td>
<td>2.188934</td>
</tr>
<tr>
<td>1.3428949</td>
<td>0.01251361</td>
<td>2.673276</td>
</tr>
</tbody>
</table>
For Bayesian inference, we can simulate draws from the predictive distribution and look at the histograms.
cbind(apply(ytilde, 2, mean), t(apply(ytilde, 2, quantile, c(0.025, 0.975))))

<table>
<thead>
<tr>
<th></th>
<th>2.5%</th>
<th>97.5%</th>
</tr>
</thead>
<tbody>
<tr>
<td>[1,]</td>
<td>1.4926618</td>
<td>0.14143269</td>
</tr>
<tr>
<td>[2,]</td>
<td>1.9932358</td>
<td>0.66109645</td>
</tr>
<tr>
<td>[3,]</td>
<td>0.8362441</td>
<td>-0.51500608</td>
</tr>
<tr>
<td>[4,]</td>
<td>1.3452031</td>
<td>0.01216690</td>
</tr>
</tbody>
</table>
Let us check if the observations are consistent with the fitted model.

- Let $y_i^*$ denote the density of a future log extinction time for a bird with covariate vector $x_i$.
- We can simulate draws of the posterior predictive distributions for all $y_1^*, \ldots, y_{62}^*$.

```r
ystar = matrix(NA, T, n)
for(i in 1:T) {
    ystar[i,] = rmnorm(1, x[i,] %*% beta[i,], sigma2[i]*diag(n))
}

cbind(y, apply(ystar, 2, mean), t(apply(ystar, 2, quantile, c(0.025, 0.975))))
```
We summarize each predictive distribution by the 95% credible interval and graph these using the `matplot` command.

- The observed log times $y_1, \ldots, y_n$ are placed as solid dots in the plot.
- Points falling outside the corresponding 95% interval are possible outliers.
Let us revisit the bird extinction example with a semi-conjugate prior. Suppose we take $\beta_0 = (0, 0, 0, 0)$, $\Sigma_0 = I$, $\nu_0 = 0.2$, $\sigma_0^2 = 0.2$.

```r
b0 = rep(0,k); sigma0 = diag(k); n0=0.2; s0=0.2
T = 10000; nburn = 5000
beta = matrix(NA, T, k); sigma2 = rep(NA, T)
beta[1,] = rep(0,k); sigma2[1] = 1
for(t in 2:T) {
  V = solve(solve(sigma0)+1/sigma2[(t-1)]*t(x)%*%x)
  m = V%*%(solve(sigma0)%*%b0+1/sigma2[(t-1)]*t(x)%*%y)
  beta[t,] = rmnorm(1, m, V)
  S2 = t(y-x%*%beta[t,]) %*% (y-x%*%beta[t,])
  sigma2[t] = rigamma(1, (n0+n)/2, (n0*s0 + S2)/2)
}
```
apply(beta, 2, mean)
[1]  0.4133903  0.2669139 -0.6282731  0.4977767

apply(beta, 2, quantile, c(0.025, 0.5, 0.975))
2.5%  0.007325563  0.1958596 -0.9511456  0.1392730
50%   0.412625341  0.2668820 -0.6270184  0.4985722
97.5%  0.802677165  0.3395624 -0.3071078  0.8464564

quantile(sigma2, c(.025, 0.5, 0.975))
  2.5%      50%      97.5%
0.3015902  0.4289842  0.6267141