Bayesian Estimation of Linear Regression Parameters with Sets of Conjugate Priors

Gero Walter

Department of Statistics
Ludwig-Maximilians-University Munich

May 13th, 2008

Linear Regression – Basics

Data:	Model:
$\underset{(n\times 1)}{(\,z\,,}\quad \underset{(n\times p)}{\mathbf{X}\,)}$	$$\begin{aligned} z &= \mathbf{X}\beta + \varepsilon \\ z_i &= x_i\beta + \varepsilon \\ &= x_{i1}\beta_1 + x_{i2}\beta_2 + \ldots + x_{ip}\beta_p + \varepsilon_i \end{aligned}$$

z_i	obs. i of response (dependent variable,...)
$(x_{i1}, \ldots, x_{ip}) =: x_i$	obs. i of regressors $j = 1, \ldots, p$ (independent variables,...) $\hat{=}$ ith row of $\mathbf{X}$
$(\beta_1, \ldots, \beta_p) =: \beta$	regression coefficients
$(\varepsilon_1, \ldots, \varepsilon_k) =: \varepsilon$	stochastic error term $\sim \mathrm{N}_k(\mathbf{0}, \sigma^2\mathbf{I})$ (σ^2 known)

Linear Regression — Estimation

- Least Squares (LS) method: minimize $\sum_{i=1}^{n}(z_i - x_i\,\beta)^2$:

$$\hat\beta_{LS} = (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T} z\,.$$

- Maximum Likelihood (ML) method: maximize likelihood $z \mid \beta \sim \mathrm{N}_n(\mathbf{X}\beta, \sigma^2\mathbf{I})$ ($\mathbf{X}$ non-stochastic) ➡ $\hat\beta_{LS}$
- Bayesian method: choose prior on β, maximize posterior (take posterior expected value)
  - often: weak prior information
    ➡ "objective Bayesian" paradigm:
    take "noninformative" prior $\beta \propto$ const. ➡ $\hat\beta_{LS}$
  - conjugate prior: convenient choice,
    posterior of same parametrical class as prior

Linear Regression — Standard Conjugate Prior

(see, e.g., O'Hagan (1994))
Called *normal regression model* in Walter et al. (2007):

$$\text{take} \quad \beta \sim \mathrm{N}_p\left(\beta^{(0)}, \sigma^2 \mathbf{\Sigma}^{(0)}\right) \qquad (\mathbf{\Sigma}^{(0)} \text{ p.d.}),$$
$$\text{then} \quad \beta \,|\, z \sim \mathrm{N}_p\left(\beta^{(1)}, \sigma^2 \mathbf{\Sigma}^{(1)}\right),$$

where the updated parameters $\beta^{(1)}$ and $\mathbf{\Sigma}^{(1)}$ are obtained as

$$\beta^{(1)} = \left(\mathbf{X}^\mathsf{T}\mathbf{X} + \mathbf{\Lambda}^{(0)}\right)^{-1}\left(\mathbf{X}^\mathsf{T} z + \mathbf{\Lambda}^{(0)}\beta^{(0)}\right)$$
$$\mathbf{\Sigma}^{(1)} = \left(\mathbf{X}^\mathsf{T}\mathbf{X} + \mathbf{\Lambda}^{(0)}\right)^{-1}, \qquad \text{with } \mathbf{\Lambda}^{(0)} = \mathbf{\Sigma}^{(0)^{-1}}$$

⚠ other conjugate priors possible! ⚠

Updating Sets of Conjugate Priors — Construction of Conjugate Priors

Given sample z of size n whose distribution forms a *linear, canonical exponential familiy* (Bernardo and Smith, 1994), i.e.

$$p(z \mid \psi) \propto \exp\left\{\langle\psi, \tau(z)\rangle - \mathbf{b}(\psi)\right\}$$

➡ Conjugate priors can be constructed as follows:

$$p(\vartheta) \propto \exp\left\{n^{(0)}\left[\langle\psi, y^{(0)}\rangle - \mathbf{b}(\psi)\right]\right\}$$

➡ Updating yields then Posterior:

$$p(\vartheta \mid z) \propto \exp\left\{n^{(1)}\left[\langle\psi, y^{(1)}\rangle - \mathbf{b}(\psi)\right]\right\}, \quad \text{where}$$

$$y^{(1)} = \frac{n^{(0)} y^{(0)} + \tau(z)}{n^{(0)} + n} \qquad \text{and} \qquad n^{(1)} = n^{(0)} + n\,.$$

Updating Sets of Conjugate Priors — Interpretation of Parameters

$y^{(0)}$: "main prior parameter"

- For samples from a $\mathrm{N}(\mu, 1)$, $p(\mu)$ is a $\mathrm{N}(y^{(0)}, \frac{1}{n^{(0)}})$
- For samples from a $\mathrm{M}(\boldsymbol{\theta})$, $p(\boldsymbol{\theta})$ is a $\mathrm{Dir}(n^{(0)}, y^{(0)})$ ($y_j^{(0)} = t_j \hat{=}$ prior probability for class j, $n^{(0)} = s$)

$n^{(0)}$: "prior strength" or "pseudocounts"

With $\tau(z) = \sum_{i=1}^n \tau(z_i)$ and $\tilde{\tau}(z) =: \frac{1}{n} \sum_{i=1}^n \tau(z_i)$:

$$y^{(1)} = \frac{n^{(0)}}{n^{(0)} + n} \cdot y^{(0)} + \frac{n}{n^{(0)} + n} \cdot \tilde{\tau}(z) \,.$$

Updating Sets of Conjugate Priors — Method of Quaeghebeur and de Cooman

Construction of *imprecise* prior: **Vary $y^{(0)}$ in a convex set $\mathcal{Y}^{(0)}$**

➡ Prior credal set can contain also *all convex mixtures* of distributions with $y^{(0)} \in \mathcal{Y}^{(0)}$

➡ Set $\mathcal{Y}^{(1)}$ defining the imprecise posterior is easily derived by

$$\mathcal{Y}^{(1)} = \left\{ \frac{n^{(0)} y^{(0)} + \tau(z)}{n^{(0)} + n} \;\middle|\; y^{(0)} \in \mathcal{Y}^{(0)} \right\} = \frac{n^{(0)}}{n^{(0)} + n} \cdot \mathcal{Y}^{(0)} + \frac{n}{n^{(0)} + n} \cdot \tilde{\tau}(z) \,.$$

Linearity: Vertices of $\mathcal{Y}^{(0)} \longrightarrow$ Vertices of $\mathcal{Y}^{(1)}$

Updating Sets of Conjugate Priors — LUCK-models

(First) Generalization:

Construction of imprecise prior not dependent on *construction of conjugate prior*, the linearity is sufficient.

➥ Defintion of LUCK-models (Linearly Updated Conjugate prior Knowledge): Prior $p(\vartheta)$ and posterior $p(\vartheta \mid z)$ such that

$$p(\vartheta) \propto \exp\left\{ n^{(0)} \left[\langle \psi, y^{(0)} \rangle - \mathbf{b}(\psi) \right] \right\} \quad \text{and}$$
$$p(\vartheta \mid z) \propto \exp\left\{ n^{(1)} \left[\langle \psi, y^{(1)} \rangle - \mathbf{b}(\psi) \right] \right\}, \quad \text{where}$$

$$y^{(1)} = \frac{n^{(0)}}{n^{(0)} + n} \cdot y^{(0)} + \frac{n}{n^{(0)} + n} \cdot \tilde{\tau}(z) \quad \text{and} \quad n^{(1)} = n^{(0)} + n\,.$$

Updating Sets of Conjugate Priors — (i)LUCK-models

➡ Two central properties of LUCK-models:

i) prior & posterior from same class of parametric distributions
ii) updating of one parameter of the prior ($y^{(0)}$) is linear.

LUCK-model with *imprecise* prior (& posterior) produced by method of Quaeghebeur and de Cooman (vary $y^{(0)}$ in a set $\mathcal{Y}^{(0)}$)

↳ **iLUCK-model**

iLUCK-models for Linear Regression — Standard Conjugate Prior

Theorem (Theorem 2 in Walter et al. (2007))

Fixing a value $n^{(0)}$, $\big(p(\beta), p(\beta \mid z)\big)$ *constitutes a* LUCK*-model of size* 1 *with prior parameters*

$$y^{(0)} = \frac{1}{n^{(0)}} \begin{pmatrix} \mathbf{\Lambda}^{(0)} \\ \mathbf{\Lambda}^{(0)} \beta^{(0)} \end{pmatrix} =: \begin{pmatrix} y_a^{(0)} \\ y_b^{(0)} \end{pmatrix}$$

and $n^{(0)}$ *and sample statistic*

$$\tau(z) = \tau(\mathbf{X}, z) = \begin{pmatrix} \mathbf{X}^\mathsf{T}\mathbf{X} \\ \mathbf{X}^\mathsf{T} z \end{pmatrix} =: \begin{pmatrix} \tau_a(\mathbf{X}, z) \\ \tau_b(\mathbf{X}, z) \end{pmatrix} .$$

Called *Imprecise Normal Regression Model* in Walter et al. (2007).

Standard Conjugate Prior — Construction

$$\beta \sim \mathrm{N}_p\left(\beta^{(0)}, \sigma^2 \mathbf{\Sigma}^{(0)}\right)$$

$$\Rightarrow \quad p(\beta) \propto \exp\left\{-\frac{1}{2\sigma^2}\left(\beta - \beta^{(0)}\right)^{\mathsf{T}} \mathbf{\Sigma}^{(0)^{-1}} \left(\beta - \beta^{(0)}\right)\right\}$$

$$\vdots$$

$$\propto \exp\left\{-\frac{1}{2\sigma^2}\beta^{\mathsf{T}}\mathbf{\Lambda}^{(0)}\beta + \frac{2}{2\sigma^2}\beta^{\mathsf{T}}\mathbf{\Lambda}^{(0)}\beta^{(0)}\right\}$$

$$\rightarrow \quad \psi = \left(\left(-\frac{\beta_i\beta_j}{2\sigma^2}\right)_{i,j=1,\dots,p}, \left(-\frac{\beta_i}{\sigma^2}\right)_{i=1,\dots,p}\right)^{\mathsf{T}}$$

$$y^{(0)} = \left(\left(\frac{\lambda_{ij}^{(0)}}{n^{(0)}}\right)_{i,j=1,\dots,p}, \left(\frac{1}{n^{(0)}}\left(\mathbf{\Lambda}^{(0)}\beta^{(0)}\right)_i\right)_{i=1,\dots,p}\right)$$

$$\mathbf{b}(\psi) = 0$$

Standard Conjugate Prior — pro & contra

+ arbitary $\mathbf{\Lambda}^{(0)}$ (p.d.) → very flexible correlation structure

---

- $n^{(0)}$ is 'artificially' introduced
- $\mathbf{b}(\beta) = 0$?!?
- $y^{(0)}$ not interpretable
  & severe 'translation' issues in concrete application

Standard Conjugate Prior — 'Translation' Issues

1. Express prior knowledge on β by a set of $\beta^{(0)}$'s and $\mathbf{\Lambda}^{(0)}$'s.
2. "Translate" this set into set of $y^{(0)}$'s such that resulting set $\mathcal{Y}^{(0)}$ consists only of admissible combinations of parameters (positive definiteness of $\mathbf{\Lambda}^{(0)}$, bounding of $\mathcal{Y}^{(0)}$ as advocated by Quaeghebeur and de Cooman).
3. Update each $y^{(0)}$ in $\mathcal{Y}^{(0)}$ linearly to $y^{(1)}$.
4. "Retranslate" $\mathcal{Y}^{(1)}$ into interpretable set of $\beta^{(1)}$'s and $\mathbf{\Lambda}^{(1)}$'s.

2. highly complex for arbitrary p

↪ analytical results derived for $p = 2$ (& further simplifications).

→ properties of resulting model very plausible.

Standard Conjugate Prior — Data Example

AIRGENE: EU financed panel study

air pollutants $\stackrel{?}{\longrightarrow}$ inflammation markers in myocardial infarction survivors

but:

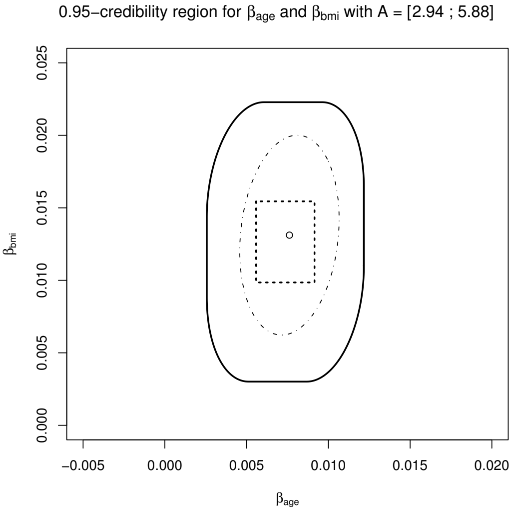
inflammation markers $\longleftrightarrow$ BMI (Body-Mass-Index) and age

➡ must be taken into account to adjust air pollutants $\longrightarrow$ inflammation markers.

Model:

$$\log(\texttt{fib})_i = [\underline{\beta}_0, \overline{\beta}_0] + \texttt{age}_i \cdot [\underline{\beta}_{\texttt{age}}, \overline{\beta}_{\texttt{age}}] + \texttt{bmi}_i \cdot [\underline{\beta}_{\texttt{bmi}}, \overline{\beta}_{\texttt{bmi}}] + \varepsilon_i$$

Standard Conjugate Prior — Data Example



Very low 'trust' in prior information corresponding to 1 – 2 observations

iLUCK-models for Linear Regression — Another Conjugate Prior

Constructed along the method described in (Bernardo et al, 1994):

'Standardize' Data with known σ^2: $z \longrightarrow \frac{z}{\sigma}$ and $\mathbf{X} \longrightarrow \frac{\mathbf{X}}{\sigma}$

$$\Rightarrow \quad f(z \mid \beta) \propto \exp\left\{\frac{1}{2}\sum_{i=1}^{n}(z_i - x_i^\mathsf{T}\beta)^2\right\}$$

$$= \exp\left\{\frac{1}{2}(z - \mathbf{X}\beta)^\mathsf{T}(z - \mathbf{X}\beta)\right\}$$

$$\vdots$$

$$\propto \exp\Big\{\underbrace{\beta^\mathsf{T}\mathbf{X}^\mathsf{T}z}_{=\langle\psi,\tau(z)\rangle}\ \underbrace{-\frac{1}{2}\beta^\mathsf{T}\mathbf{X}^\mathsf{T}\mathbf{X}\beta}_{n\mathbf{b}(\psi)}\Big\}$$

$$\Rightarrow \quad \psi = \beta\,,$$

$$\tau(z) = \mathbf{X}^\mathsf{T}z\,, \qquad \mathbf{b}(\psi) = \frac{1}{2n}\beta^\mathsf{T}\mathbf{X}^\mathsf{T}\mathbf{X}\beta$$

Other Conjugate Prior — Construction

construction of prior: $p(\vartheta) \propto \exp\left\{ n^{(0)} \left[\langle \psi, y^{(0)} \rangle - \mathbf{b}(\psi) \right] \right\}$

from sample model: $p(\beta) \propto \exp\left\{ n^{(0)} \left[y^{(0)\mathsf{T}} \beta - \frac{1}{2n} \beta^{\mathsf{T}} \mathbf{X}^{\mathsf{T}} \mathbf{X} \beta \right] \right\}$

Density of a multivariate normal with mean $f(n^{(0)}, y^{(0)})$ and inverse covariance matrix $\mathbf{S}(n^{(0)}, y^{(0)})$:

$$p(\beta) \propto \exp\left\{ -\frac{1}{2} \Big(\beta - f(\,,\,)\Big)^{\mathsf{T}} \mathbf{S}(\,,\,) \Big(\beta - f(\,,\,)\Big) \right\}$$

$$\propto \exp\left\{ -\frac{1}{2} f(\,,\,)^{\mathsf{T}} \mathbf{S}(\,,\,) \beta + \frac{1}{2} \beta^{\mathsf{T}} \mathbf{S}(\,,\,) \beta \right\}$$

$\rightarrow$ $$\mathbf{S}(n^{(0)}, y^{(0)}) = \mathbf{S}(n^{(0)}) = \frac{n^{(0)}}{n} \mathbf{X}^{\mathsf{T}} \mathbf{X},$$

$$f(n^{(0)}, y^{(0)}) = f(y^{(0)}) = n \, (\mathbf{X}^{\mathsf{T}} \mathbf{X})^{-1} y^{(0)}$$

Other Conjugate Prior — How to Proceed

1. fix lower and upper bounds for $f(y^{(0)})$ based on prior knowledege on β; $n^{(0)}$ must be chosen fix ($\rightarrow$ Friday) and determines the prior covariance matrix for β: $\mathbb{V}(\beta) = \frac{n}{n^{(0)}}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}$;
2. 'translate' bounds for $f(y^{(0)})$ into bounds for $y^{(0)}$ by $y^{(0)} = \frac{1}{n}(\mathbf{X}^\mathsf{T}\mathbf{X})f(y^{(0)})$;
3. perform the linear update step on $n^{(0)}$ and the bounds for $y^{(0)}$ to obtain $n^{(1)}$ and bounds for $y^{(1)}$;
4. 'retranslate' the bounds for $y^{(1)}$ into interpretable bounds for $f(y^{(1)})$.

As all transformations are linear and no p.d.-safeguarding necessary, iLUCK-model calculus ($f(\mathcal{Y}^{(0)}) \longrightarrow f(\mathcal{Y}^{(1)})$) is easy!

Other Conjugate Prior — Updating

$$\begin{aligned}\mathbb{E}[\beta \mid z] &= f(y^{(1)}) \\ &= n(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\left(\frac{n^{(0)}}{n^{(0)}+n}\, y^{(0)} + \frac{n}{n^{(0)}+n}\cdot\frac{1}{n}(\mathbf{X}^\mathsf{T} z)\right) \\ &= \frac{n^{(0)}}{n^{(0)}+n}\cdot f(y^{(0)}) + \frac{n}{n^{(0)}+n}\cdot \underbrace{(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T} z}_{\hat\beta_{LS}},\end{aligned}$$

$$\begin{aligned}\mathbb{V}(\beta \mid z) &= \frac{n}{n^{(1)}}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1} \\ &= \frac{n}{n^{(0)}+n}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\,.\end{aligned}$$

Other Conjugate Prior — pro & contra

+ $y^{(0)}$ not interpretable, but easy transformation
+ easy updating as weighted average of prior guess and $\hat{\beta}_{LS}$ intuitively appealing

---

- no flexible correlation structure for β
- $\mathbb{V}(\beta)$ and $\mathbb{V}(\beta \mid z)$ not interval-valued (fixed $n^{(0)}$!)

Concluding Remarks

Presented models for generalized Bayesian estimation of regression coefficients:

- either flexible covariance structure and difficult calculations
- or fixed covariance structure and easy calculations

Second generalizion: *generalized* iLUCK-models:
$n^{(0)}$ varying in set $\mathcal{N}^{(0)}$ additionally

➥ see my contribution on Friday