SDS 383C: Statistical Modeling I
Fall 2022, Module IX

Abhra Sarkar

Department of Statistics and Data Sciences
The University of Texas at Austin

"All models are wrong, but some are useful."- George E. P. Box

Linear Regression

- Models linear in parameters

$$y_i = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \epsilon_i, \quad \epsilon_i \overset{iid}{\sim} f_\epsilon, \quad i = 1, \ldots, n, \quad \text{with } \mathbb{E}_{f_\epsilon}(\epsilon) = 0.$$

  - $y_i = \beta_0 + x_i\beta_1 + \epsilon_i, \quad \epsilon_i \overset{iid}{\sim} \text{Normal}(0, \sigma^2)$
  - $y_i = \beta_0 + x_i\beta_1 + \epsilon_i, \quad \epsilon_i \overset{iid}{\sim} \text{Laplace}(0, b)$
  - $y_i = \beta_0 + x_i\beta_1 + x_i^2\beta_2 + x_i^3\beta_3 + \epsilon_i, \quad \epsilon_i \overset{iid}{\sim} \text{Normal}(0, \sigma^2)$
  - $y_i = \beta_0 + x_{i,1}\beta_1 + x_{i,2}\beta_2 + \cdots + x_{i,p}\beta_p + \epsilon_i, \quad \epsilon_i \overset{iid}{\sim} \text{Normal}(0, \sigma^2)$

- Matrix-vector notation

$$\mathbf{y}^{n\times 1} = \mathbf{X}^{n\times p}\boldsymbol{\beta}^{n\times 1} + \mathbf{e}^{n\times 1}, \qquad \mathbb{E}(\mathbf{e}) = \mathbf{0}, \quad \text{var}(\mathbf{e}) = \sigma^2\mathbf{I}_n$$

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} x_{1,1} & x_{1,2} & \ldots & x_{1,p} \\ x_{2,1} & x_{1,2} & \ldots & x_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \ldots & x_{n,p} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

- Models linear in parameters

$$y_i = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \epsilon_i, \quad \epsilon_i \overset{iid}{\sim} f_\epsilon, \quad i = 1, \ldots, n, \quad \text{with } \mathbb{E}_{f_\epsilon}(\epsilon) = 0.$$

  - $y_i = \beta_0 + x_i\beta_1 + \epsilon_i, \quad \epsilon_i \overset{iid}{\sim} \text{Normal}(0, \sigma^2)$
  - $y_i = \beta_0 + x_i\beta_1 + \epsilon_i, \quad \epsilon_i \overset{iid}{\sim} \text{Laplace}(0, b)$
  - $y_i = \beta_0 + x_i\beta_1 + x_i^2\beta_2 + x_i^3\beta_3 + \epsilon_i, \quad \epsilon_i \overset{iid}{\sim} \text{Normal}(0, \sigma^2)$
  - $y_i = \beta_0 + x_{i,1}\beta_1 + x_{i,2}\beta_2 + \cdots + x_{i,p}\beta_p + \epsilon_i, \quad \epsilon_i \overset{iid}{\sim} \text{Normal}(0, \sigma^2)$
- Matrix-vector notation

$$\mathbf{y}^{n\times 1} = \mathbf{X}^{n\times p}\boldsymbol{\beta}^{n\times 1} + \mathbf{e}^{n\times 1}, \qquad \mathbb{E}(\mathbf{e}) = \mathbf{0}, \quad \text{var}(\mathbf{e}) = \sigma^2\mathbf{I}_n$$

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} x_{1,1} & x_{1,2} & \ldots & x_{1,p} \\ x_{2,1} & x_{1,2} & \ldots & x_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \ldots & x_{n,p} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

Linear Regression

- Models linear in parameters

$$y_i = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \epsilon_i, \quad \epsilon_i \overset{iid}{\sim} f_\epsilon, \quad i = 1, \dots, n, \quad \text{with } \mathbb{E}_{f_\epsilon}(\epsilon) = 0.$$

  - $y_i = \beta_0 + x_i\beta_1 + \epsilon_i, \quad \epsilon_i \overset{iid}{\sim} \text{Normal}(0, \sigma^2)$
  - $y_i = \beta_0 + x_i\beta_1 + \epsilon_i, \quad \epsilon_i \overset{iid}{\sim} \text{Laplace}(0, b)$
  - $y_i = \beta_0 + x_i\beta_1 + x_i^2\beta_2 + x_i^3\beta_3 + \epsilon_i, \quad \epsilon_i \overset{iid}{\sim} \text{Normal}(0, \sigma^2)$
  - $y_i = \beta_0 + x_{i,1}\beta_1 + x_{i,2}\beta_2 + \cdots + x_{i,p}\beta_p + \epsilon_i, \quad \epsilon_i \overset{iid}{\sim} \text{Normal}(0, \sigma^2)$

- Matrix-vector notation

$$\mathbf{y}^{n\times 1} = \mathbf{X}^{n\times p}\boldsymbol{\beta}^{n\times 1} + \mathbf{e}^{n\times 1}, \qquad \mathbb{E}(\mathbf{e}) = \mathbf{0}, \quad \text{var}(\mathbf{e}) = \sigma^2\mathbf{I}_n$$

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,p} \\ x_{2,1} & x_{1,2} & \dots & x_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \dots & x_{n,p} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

- Models linear in parameters

$$y_i = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \epsilon_i, \quad \epsilon_i \overset{iid}{\sim} f_\epsilon, \quad i = 1, \ldots, n, \quad \text{with } \mathbb{E}_{f_\epsilon}(\epsilon) = 0.$$

  - $y_i = \beta_0 + x_i\beta_1 + \epsilon_i, \quad \epsilon_i \overset{iid}{\sim} \text{Normal}(0, \sigma^2)$
  - $y_i = \beta_0 + x_i\beta_1 + \epsilon_i, \quad \epsilon_i \overset{iid}{\sim} \text{Laplace}(0, b)$
  - $y_i = \beta_0 + x_i\beta_1 + x_i^2\beta_2 + x_i^3\beta_3 + \epsilon_i, \quad \epsilon_i \overset{iid}{\sim} \text{Normal}(0, \sigma^2)$
  - $y_i = \beta_0 + x_{i,1}\beta_1 + x_{i,2}\beta_2 + \cdots + x_{i,p}\beta_p + \epsilon_i, \quad \epsilon_i \overset{iid}{\sim} \text{Normal}(0, \sigma^2)$

- Matrix-vector notation

$$\mathbf{y}^{n\times 1} = \mathbf{X}^{n\times p}\boldsymbol{\beta}^{n\times 1} + \mathbf{e}^{n\times 1}, \qquad \mathbb{E}(\mathbf{e}) = \mathbf{0}, \quad \text{var}(\mathbf{e}) = \sigma^2\mathbf{I}_n$$

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} x_{1,1} & x_{1,2} & \ldots & x_{1,p} \\ x_{2,1} & x_{1,2} & \ldots & x_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \ldots & x_{n,p} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

Linear Regression

- Models linear in parameters

$$y_i = \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta} + \epsilon_i, \quad \epsilon_i \overset{iid}{\sim} f_\epsilon, \quad i = 1, \ldots, n, \quad \text{with } \mathbb{E}_{f_\epsilon}(\epsilon) = 0.$$

  - $y_i = \beta_0 + x_i \beta_1 + \epsilon_i, \quad \epsilon_i \overset{iid}{\sim} \text{Normal}(0, \sigma^2)$
  - $y_i = \beta_0 + x_i \beta_1 + \epsilon_i, \quad \epsilon_i \overset{iid}{\sim} \text{Laplace}(0, b)$
  - $y_i = \beta_0 + x_i \beta_1 + x_i^2 \beta_2 + x_i^3 \beta_3 + \epsilon_i, \quad \epsilon_i \overset{iid}{\sim} \text{Normal}(0, \sigma^2)$
  - $y_i = \beta_0 + x_{i,1} \beta_1 + x_{i,2} \beta_2 + \cdots + x_{i,p} \beta_p + \epsilon_i, \quad \epsilon_i \overset{iid}{\sim} \text{Normal}(0, \sigma^2)$

- Matrix-vector notation

$$\mathbf{y}^{n\times 1} = \mathbf{X}^{n\times p} \boldsymbol{\beta}^{n\times 1} + \mathbf{e}^{n\times 1}, \qquad \mathbb{E}(\mathbf{e}) = \mathbf{0}, \quad \text{var}(\mathbf{e}) = \sigma^2 \mathbf{I}_n$$

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} x_{1,1} & x_{1,2} & \ldots & x_{1,p} \\ x_{2,1} & x_{1,2} & \ldots & x_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \ldots & x_{n,p} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

- Models linear in parameters

$$y_i = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \epsilon_i, \quad \epsilon_i \overset{iid}{\sim} f_\epsilon, \quad i = 1, \dots, n, \quad \text{with } \mathbb{E}_{f_\epsilon}(\epsilon) = 0.$$

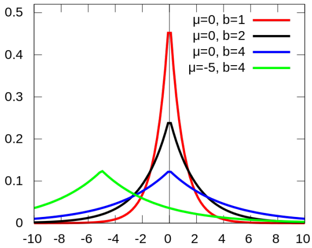
  - $y_i = \beta_0 + x_i\beta_1 + \epsilon_i, \quad \epsilon_i \overset{iid}{\sim} \text{Normal}(0, \sigma^2)$
  - $y_i = \beta_0 + x_i\beta_1 + \epsilon_i, \quad \epsilon_i \overset{iid}{\sim} \text{Laplace}(0, b)$
  - $y_i = \beta_0 + x_i\beta_1 + x_i^2\beta_2 + x_i^3\beta_3 + \epsilon_i, \quad \epsilon_i \overset{iid}{\sim} \text{Normal}(0, \sigma^2)$
  - $y_i = \beta_0 + x_{i,1}\beta_1 + x_{i,2}\beta_2 + \cdots + x_{i,p}\beta_p + \epsilon_i, \quad \epsilon_i \overset{iid}{\sim} \text{Normal}(0, \sigma^2)$

- Matrix-vector notation

$$\mathbf{y}^{n\times 1} = \mathbf{X}^{n\times p}\boldsymbol{\beta}^{n\times 1} + \mathbf{e}^{n\times 1}, \quad \mathbb{E}(\mathbf{e}) = \mathbf{0}, \quad \text{var}(\mathbf{e}) = \sigma^2\mathbf{I}_n$$

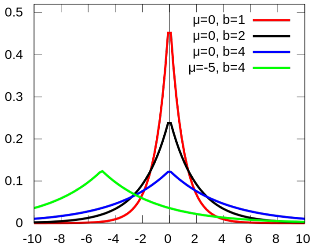
$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,p} \\ x_{2,1} & x_{1,2} & \dots & x_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \dots & x_{n,p} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

- **Laplace Distribution:**
$$y \sim f(y \mid \mu, b) = \frac{1}{2b} \exp\left(-\frac{|y-\mu|}{b}\right).$$
- $\mathbb{E}(y) = \mu$
- $\text{var}(y) = 2b^2$
- $\text{skewness}(y) = 0$
- $\text{excess kurtosis}(y) = 3$



Laplace / Double Exponential Distribution

- **Laplace Distribution:**
  $$y \sim f(y \mid \mu, b) = \frac{1}{2b} \exp\left(-\frac{|y-\mu|}{b}\right).$$
- $\mathbb{E}(y) = \mu$
- $\text{var}(y) = 2b^2$
- $\text{skewness}(y) = 0$
- $\text{excess kurtosis}(y) = 3$



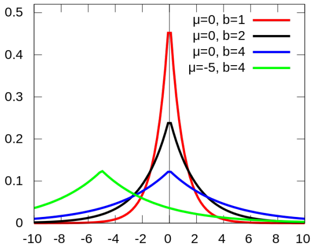
- Laplace as Normal scale mixture

$$\frac{a}{2}\exp(-a\,|z|) = \int_0^\infty \frac{1}{\sqrt{2\pi s^2}} \exp\left(-\frac{z^2}{2s^2}\right) \frac{a^2}{2} \exp\left(-\frac{a^2 s^2}{2}\right) ds^2$$

$$= \int_0^\infty \text{Normal}(z \mid 0, s^2)\ \text{Exp}\left(s^2 \mid \frac{a^2}{2}\right) ds^2$$

Laplace / Double Exponential Distribution

- **Laplace Distribution:**
  $$y \sim f(y \mid \mu, b) = \frac{1}{2b} \exp\left(-\frac{|y-\mu|}{b}\right).$$
- $\mathbb{E}(y) = \mu$
- $\mathrm{var}(y) = 2b^2$
- $\mathrm{skewness}(y) = 0$
- $\text{excess kurtosis}(y) = 3$



- Laplace as Normal scale mixture

$$\frac{a}{2}\exp(-a\,|z|) = \int_0^\infty \frac{1}{\sqrt{2\pi s^2}} \exp\left(-\frac{z^2}{2s^2}\right) \frac{a^2}{2} \exp\left(-\frac{a^2 s^2}{2}\right) ds^2$$
$$= \int_0^\infty \mathrm{Normal}(z \mid 0, s^2)\ \mathrm{Exp}\left(s^2 \mid \frac{a^2}{2}\right) ds^2$$

$\rightarrow$ Useful for Bayesian computation!

Ordinary Least Squares

$$\mathbf{y}^{n\times 1} = \mathbf{X}^{n\times p}\boldsymbol{\beta}^{p\times 1} + \mathbf{e}^{n\times 1}, \qquad \mathbb{E}(\mathbf{e}) = \mathbf{0},\ \text{cov}(\mathbf{e}) = \sigma^2\mathbf{I}_n$$

- Minimize squared error loss:

$$\widehat{\boldsymbol{\beta}} = \arg\min \sum_{i=1}^{n}(y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})^2 = \arg\min(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$$

- $\mathbb{E}(\widehat{\boldsymbol{\beta}}) = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta}$
- $\text{var}(\widehat{\boldsymbol{\beta}}) = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\sigma^2\mathbf{I}_n\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1} = \sigma^2(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$
  - Hat matrix: $\mathbf{H} = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}$
- Fitted values: $\widehat{\mathbf{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y} = \mathbf{H}\mathbf{y}$
- Mean-Variance: $\mathbb{E}(\widehat{\mathbf{y}}) = \mathbf{X}\mathbb{E}(\widehat{\boldsymbol{\beta}}) = \mathbf{X}\boldsymbol{\beta}, \quad \text{var}(\widehat{\mathbf{y}}) = \sigma^2\mathbf{H}$
- Residuals: $\widehat{\mathbf{e}} = (\mathbf{y} - \widehat{\mathbf{y}}) = \mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}} = \{\mathbf{I}_n - \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\}\mathbf{y} = (\mathbf{I}_n - \mathbf{H})\mathbf{y}$
- Mean-Variance: $\mathbb{E}(\widehat{\mathbf{e}}) = (\mathbf{I}_n - \mathbf{H})\mathbf{X}\boldsymbol{\beta} = \mathbf{0}, \quad \text{var}(\widehat{\mathbf{e}}) = \mathbb{E}(\widehat{\mathbf{e}}\widehat{\mathbf{e}}^{\mathrm{T}}) = \sigma^2(\mathbf{I}_n - \mathbf{H})$
- $\text{cov}(\widehat{\mathbf{y}}, \widehat{\mathbf{e}}) = \text{cov}\{\mathbf{H}\mathbf{y}, (\mathbf{I}_n - \mathbf{H})\mathbf{y}\} = \sigma^2\mathbf{H}(\mathbf{I}_n - \mathbf{H}) = \sigma^2(\mathbf{H} - \mathbf{H}^2) = \mathbf{0}$

Ordinary Least Squares

$$\mathbf{y}^{n\times 1} = \mathbf{X}^{n\times p}\boldsymbol{\beta}^{p\times 1} + \mathbf{e}^{n\times 1}, \qquad \mathbb{E}(\mathbf{e}) = \mathbf{0},\ \operatorname{cov}(\mathbf{e}) = \sigma^2\mathbf{I}_n$$

- Minimize squared error loss:

$\widehat{\boldsymbol{\beta}} = \arg\min \sum_{i=1}^{n}(y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})^2 = \arg\min(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$

- $\mathbb{E}(\widehat{\boldsymbol{\beta}}) = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta}$
- $\operatorname{var}(\widehat{\boldsymbol{\beta}}) = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\sigma^2\mathbf{I}_n\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1} = \sigma^2(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$
  - Hat matrix: $\mathbf{H} = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}$
  - Idempotent: $\mathbf{H}^2 = \mathbf{H}$, $(\mathbf{I}_n - \mathbf{H})^2 = (\mathbf{I}_n - \mathbf{H})$
  - $\operatorname{trace}(\mathbf{H}) = \operatorname{trace}\{\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\} = \operatorname{trace}\{\mathbf{X}^{\mathrm{T}}\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\} = p$
  - $\operatorname{trace}(\mathbf{I}_n - \mathbf{H}) = \operatorname{trace}(\mathbf{I}_n) - \operatorname{trace}(\mathbf{H}) = (n - p)$
  - $\mathbf{H}\mathbf{X} = \mathbf{X}$ and $(\mathbf{I}_n - \mathbf{H})\mathbf{X} = \{\mathbf{I}_n - \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\}\mathbf{X} = \mathbf{0}$
- Fitted values: $\widehat{\mathbf{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y} = \mathbf{H}\mathbf{y}$
- Mean-Variance: $\mathbb{E}(\widehat{\mathbf{y}}) = \mathbf{X}\mathbb{E}(\widehat{\boldsymbol{\beta}}) = \mathbf{X}\boldsymbol{\beta}$, $\quad \operatorname{var}(\widehat{\mathbf{y}}) = \sigma^2\mathbf{H}$
- Residuals: $\widehat{\mathbf{e}} = (\mathbf{y} - \widehat{\mathbf{y}}) = \mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}} = \{\mathbf{I}_n - \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\}\mathbf{y} = (\mathbf{I}_n - \mathbf{H})\mathbf{y}$
- Mean-Variance: $\mathbb{E}(\widehat{\mathbf{e}}) = (\mathbf{I}_n - \mathbf{H})\mathbf{X}\boldsymbol{\beta} = \mathbf{0}$, $\quad \operatorname{var}(\widehat{\mathbf{e}}) = \mathbb{E}(\widehat{\mathbf{e}}\widehat{\mathbf{e}}^{\mathrm{T}}) = \sigma^2(\mathbf{I}_n - \mathbf{H})$
- $\operatorname{cov}(\widehat{\mathbf{y}}, \widehat{\mathbf{e}}) = \operatorname{cov}\{\mathbf{H}\mathbf{y}, (\mathbf{I}_n - \mathbf{H})\mathbf{y}\} = \sigma^2\mathbf{H}(\mathbf{I}_n - \mathbf{H}) = \sigma^2(\mathbf{H} - \mathbf{H}^2) = \mathbf{0}$

$$\mathbf{y}^{n\times 1} = \mathbf{X}^{n\times p}\boldsymbol{\beta}^{p\times 1} + \mathbf{e}^{n\times 1}, \qquad \mathbb{E}(\mathbf{e}) = \mathbf{0},\ \operatorname{cov}(\mathbf{e}) = \sigma^2\mathbf{I}_n$$

- Minimize squared error loss:

$$\widehat{\boldsymbol{\beta}} = \arg\min \textstyle\sum_{i=1}^{n}(y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})^2 = \arg\min(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$$

- $\mathbb{E}(\widehat{\boldsymbol{\beta}}) = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta}$
- $\operatorname{var}(\widehat{\boldsymbol{\beta}}) = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\sigma^2\mathbf{I}_n\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1} = \sigma^2(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$
  - Hat matrix: $\mathbf{H} = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}$
  - Idempotent: $\mathbf{H}^2 = \mathbf{H}, \quad (\mathbf{I}_n - \mathbf{H})^2 = (\mathbf{I}_n - \mathbf{H})$
  - $\operatorname{trace}(\mathbf{H}) = \operatorname{trace}\{\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\} = \operatorname{trace}\{\mathbf{X}^{\mathrm{T}}\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\} = p,$
  - $\operatorname{trace}(\mathbf{I}_n - \mathbf{H}) = \operatorname{trace}(\mathbf{I}_n) - \operatorname{trace}(\mathbf{H}) = (n-p)$
  - $\mathbf{H}\mathbf{X} = \mathbf{X}$ and $(\mathbf{I}_n - \mathbf{H})\mathbf{X} = \{\mathbf{I}_n - \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\}\mathbf{X} = \mathbf{0}$
- Fitted values: $\widehat{\mathbf{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y} = \mathbf{H}\mathbf{y}$
- Mean-Variance: $\mathbb{E}(\widehat{\mathbf{y}}) = \mathbf{X}\mathbb{E}(\widehat{\boldsymbol{\beta}}) = \mathbf{X}\boldsymbol{\beta}, \quad \operatorname{var}(\widehat{\mathbf{y}}) = \sigma^2\mathbf{H}$
- Residuals: $\widehat{\mathbf{e}} = (\mathbf{y} - \widehat{\mathbf{y}}) = \mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}} = \{\mathbf{I}_n - \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\}\mathbf{y} = (\mathbf{I}_n - \mathbf{H})\mathbf{y}$
- Mean-Variance: $\mathbb{E}(\widehat{\mathbf{e}}) = (\mathbf{I}_n - \mathbf{H})\mathbf{X}\boldsymbol{\beta} = \mathbf{0}, \quad \operatorname{var}(\widehat{\mathbf{e}}) = \mathbb{E}(\widehat{\mathbf{e}}\widehat{\mathbf{e}}^{\mathrm{T}}) = \sigma^2(\mathbf{I}_n - \mathbf{H})$
- $\operatorname{cov}(\widehat{\mathbf{y}}, \widehat{\mathbf{e}}) = \operatorname{cov}\{\mathbf{H}\mathbf{y}, (\mathbf{I}_n - \mathbf{H})\mathbf{y}\} = \sigma^2\mathbf{H}(\mathbf{I}_n - \mathbf{H}) = \sigma^2(\mathbf{H} - \mathbf{H}^2) = \mathbf{0}$

Ordinary Least Squares

$$\mathbf{y}^{n\times 1} = \mathbf{X}^{n\times p}\boldsymbol{\beta}^{p\times 1} + \mathbf{e}^{n\times 1}, \qquad \mathbb{E}(\mathbf{e}) = \mathbf{0},\ \operatorname{cov}(\mathbf{e}) = \sigma^2\mathbf{I}_n$$

- Minimize squared error loss:

$$\widehat{\boldsymbol{\beta}} = \arg\min \textstyle\sum_{i=1}^n (y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})^2 = \arg\min(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$$

- $\mathbb{E}(\widehat{\boldsymbol{\beta}}) = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta}$
- $\operatorname{var}(\widehat{\boldsymbol{\beta}}) = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\sigma^2\mathbf{I}_n\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1} = \sigma^2(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$
    - Hat matrix: $\mathbf{H} = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}$
    - Idempotent: $\mathbf{H}^2 = \mathbf{H}, \quad (\mathbf{I}_n - \mathbf{H})^2 = (\mathbf{I}_n - \mathbf{H})$
    - $\operatorname{trace}(\mathbf{H}) = \operatorname{trace}\{\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\} = \operatorname{trace}\{\mathbf{X}^{\mathrm{T}}\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\} = p,$
    - $\operatorname{trace}(\mathbf{I}_n - \mathbf{H}) = \operatorname{trace}(\mathbf{I}_n) - \operatorname{trace}(\mathbf{H}) = (n - p)$
    - $\mathbf{H}\mathbf{X} = \mathbf{X}$ and $(\mathbf{I}_n - \mathbf{H})\mathbf{X} = \{\mathbf{I}_n - \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\}\mathbf{X} = \mathbf{0}$
- Fitted values: $\widehat{\mathbf{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y} = \mathbf{H}\mathbf{y}$
- Mean-Variance: $\mathbb{E}(\widehat{\mathbf{y}}) = \mathbf{X}\mathbb{E}(\widehat{\boldsymbol{\beta}}) = \mathbf{X}\boldsymbol{\beta}, \quad \operatorname{var}(\widehat{\mathbf{y}}) = \sigma^2\mathbf{H}$
- Residuals: $\widehat{\mathbf{e}} = (\mathbf{y} - \widehat{\mathbf{y}}) = \mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}} = \{\mathbf{I}_n - \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\}\mathbf{y} = (\mathbf{I}_n - \mathbf{H})\mathbf{y}$
- Mean-Variance: $\mathbb{E}(\widehat{\mathbf{e}}) = (\mathbf{I}_n - \mathbf{H})\mathbf{X}\boldsymbol{\beta} = \mathbf{0}, \quad \operatorname{var}(\widehat{\mathbf{e}}) = \mathbb{E}(\widehat{\mathbf{e}}\widehat{\mathbf{e}}^{\mathrm{T}}) = \sigma^2(\mathbf{I}_n - \mathbf{H})$
- $\operatorname{cov}(\widehat{\mathbf{y}}, \widehat{\mathbf{e}}) = \operatorname{cov}\{\mathbf{H}\mathbf{y}, (\mathbf{I}_n - \mathbf{H})\mathbf{y}\} = \sigma^2\mathbf{H}(\mathbf{I}_n - \mathbf{H}) = \sigma^2(\mathbf{H} - \mathbf{H}^2) = \mathbf{0}$

Ordinary Least Squares

$$\mathbf{y}^{n\times 1} = \mathbf{X}^{n\times p}\boldsymbol{\beta}^{p\times 1} + \mathbf{e}^{n\times 1}, \qquad \mathbb{E}(\mathbf{e}) = \mathbf{0},\ \mathrm{cov}(\mathbf{e}) = \sigma^2\mathbf{I}_n$$

- Minimize squared error loss:

$$\widehat{\boldsymbol{\beta}} = \arg\min \sum_{i=1}^{n}(y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})^2 = \arg\min(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$$

- $\mathbb{E}(\widehat{\boldsymbol{\beta}}) = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta}$
- $\mathrm{var}(\widehat{\boldsymbol{\beta}}) = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\sigma^2\mathbf{I}_n\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1} = \sigma^2(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$
  - Hat matrix: $\mathbf{H} = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}$
  - Idempotent: $\mathbf{H}^2 = \mathbf{H}, \quad (\mathbf{I}_n - \mathbf{H})^2 = (\mathbf{I}_n - \mathbf{H})$
  - $\mathrm{trace}(\mathbf{H}) = \mathrm{trace}\{\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\} = \mathrm{trace}\{\mathbf{X}^{\mathrm{T}}\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\} = p,$
  - $\mathrm{trace}(\mathbf{I}_n - \mathbf{H}) = \mathrm{trace}(\mathbf{I}_n) - \mathrm{trace}(\mathbf{H}) = (n-p)$
  - $\mathbf{H}\mathbf{X} = \mathbf{X}$ and $(\mathbf{I}_n - \mathbf{H})\mathbf{X} = \{\mathbf{I}_n - \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\}\mathbf{X} = \mathbf{0}$
- Fitted values: $\widehat{\mathbf{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y} = \mathbf{H}\mathbf{y}$
- Mean-Variance: $\mathbb{E}(\widehat{\mathbf{y}}) = \mathbf{X}\mathbb{E}(\widehat{\boldsymbol{\beta}}) = \mathbf{X}\boldsymbol{\beta}, \quad \mathrm{var}(\widehat{\mathbf{y}}) = \sigma^2\mathbf{H}$
- Residuals: $\widehat{\mathbf{e}} = (\mathbf{y} - \widehat{\mathbf{y}}) = \mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}} = \{\mathbf{I}_n - \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\}\mathbf{y} = (\mathbf{I}_n - \mathbf{H})\mathbf{y}$
- Mean-Variance: $\mathbb{E}(\widehat{\mathbf{e}}) = (\mathbf{I}_n - \mathbf{H})\mathbf{X}\boldsymbol{\beta} = \mathbf{0}, \quad \mathrm{var}(\widehat{\mathbf{e}}) = \mathbb{E}(\widehat{\mathbf{e}}\widehat{\mathbf{e}}^{\mathrm{T}}) = \sigma^2(\mathbf{I}_n - \mathbf{H})$
- $\mathrm{cov}(\widehat{\mathbf{y}}, \widehat{\mathbf{e}}) = \mathrm{cov}\{\mathbf{H}\mathbf{y}, (\mathbf{I}_n - \mathbf{H})\mathbf{y}\} = \sigma^2\mathbf{H}(\mathbf{I}_n - \mathbf{H}) = \sigma^2(\mathbf{H} - \mathbf{H}^2) = 0$

Ordinary Least Squares

$$\mathbf{y}^{n\times 1} = \mathbf{X}^{n\times p}\boldsymbol{\beta}^{p\times 1} + \mathbf{e}^{n\times 1}, \qquad \mathbb{E}(\mathbf{e}) = \mathbf{0},\ \mathrm{cov}(\mathbf{e}) = \sigma^2\mathbf{I}_n$$

- Minimize squared error loss:

$\widehat{\boldsymbol{\beta}} = \arg\min \sum_{i=1}^{n}(y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})^2 = \arg\min(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$

- $\mathbb{E}(\widehat{\boldsymbol{\beta}}) = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta}$
- $\mathrm{var}(\widehat{\boldsymbol{\beta}}) = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\sigma^2\mathbf{I}_n\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1} = \sigma^2(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$
  - Hat matrix: $\mathbf{H} = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}$
  - Idempotent: $\mathbf{H}^2 = \mathbf{H}, \quad (\mathbf{I}_n - \mathbf{H})^2 = (\mathbf{I}_n - \mathbf{H})$
  - $\mathrm{trace}(\mathbf{H}) = \mathrm{trace}\{\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\} = \mathrm{trace}\{\mathbf{X}^{\mathrm{T}}\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\} = p,$
  - $\mathrm{trace}(\mathbf{I}_n - \mathbf{H}) = \mathrm{trace}(\mathbf{I}_n) - \mathrm{trace}(\mathbf{H}) = (n - p)$
  - $\mathbf{H}\mathbf{X} = \mathbf{X}$ and $(\mathbf{I}_n - \mathbf{H})\mathbf{X} = \{\mathbf{I}_n - \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\}\mathbf{X} = \mathbf{0}$
- Fitted values: $\widehat{\mathbf{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y} = \mathbf{H}\mathbf{y}$
- Mean-Variance: $\mathbb{E}(\widehat{\mathbf{y}}) = \mathbf{X}\mathbb{E}(\widehat{\boldsymbol{\beta}}) = \mathbf{X}\boldsymbol{\beta}, \quad \mathrm{var}(\widehat{\mathbf{y}}) = \sigma^2\mathbf{H}$
- Residuals: $\widehat{\mathbf{e}} = (\mathbf{y} - \widehat{\mathbf{y}}) = \mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}} = \{\mathbf{I}_n - \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\}\mathbf{y} = (\mathbf{I}_n - \mathbf{H})\mathbf{y}$
- Mean-Variance: $\mathbb{E}(\widehat{\mathbf{e}}) = (\mathbf{I}_n - \mathbf{H})\mathbf{X}\boldsymbol{\beta} = \mathbf{0}, \quad \mathrm{var}(\widehat{\mathbf{e}}) = \mathbb{E}(\widehat{\mathbf{e}}\widehat{\mathbf{e}}^{\mathrm{T}}) = \sigma^2(\mathbf{I}_n - \mathbf{H})$
- $\mathrm{cov}(\widehat{\mathbf{y}}, \widehat{\mathbf{e}}) = \mathrm{cov}\{\mathbf{H}\mathbf{y}, (\mathbf{I}_n - \mathbf{H})\mathbf{y}\} = \sigma^2\mathbf{H}(\mathbf{I}_n - \mathbf{H}) = \sigma^2(\mathbf{H} - \mathbf{H}^2) = \mathbf{0}$

Ordinary Least Squares

$$\mathbf{y}^{n\times 1} = \mathbf{X}^{n\times p}\boldsymbol{\beta}^{p\times 1} + \mathbf{e}^{n\times 1}, \qquad \mathbb{E}(\mathbf{e}) = \mathbf{0},\ \text{cov}(\mathbf{e}) = \sigma^2\mathbf{I}_n$$

- Minimize squared error loss:

$$\widehat{\boldsymbol{\beta}} = \arg\min \textstyle\sum_{i=1}^{n}(y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})^2 = \arg\min(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$$

- $\mathbb{E}(\widehat{\boldsymbol{\beta}}) = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta}$
- $\text{var}(\widehat{\boldsymbol{\beta}}) = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\sigma^2\mathbf{I}_n\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1} = \sigma^2(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$
  - Hat matrix: $\mathbf{H} = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}$
  - Idempotent: $\mathbf{H}^2 = \mathbf{H}, \quad (\mathbf{I}_n - \mathbf{H})^2 = (\mathbf{I}_n - \mathbf{H})$
  - $\text{trace}(\mathbf{H}) = \text{trace}\{\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\} = \text{trace}\{\mathbf{X}^{\mathrm{T}}\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\} = p,$
  - $\text{trace}(\mathbf{I}_n - \mathbf{H}) = \text{trace}(\mathbf{I}_n) - \text{trace}(\mathbf{H}) = (n-p)$
  - $\mathbf{H}\mathbf{X} = \mathbf{X}$ and $(\mathbf{I}_n - \mathbf{H})\mathbf{X} = \{\mathbf{I}_n - \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\}\mathbf{X} = \mathbf{0}$
- Fitted values: $\widehat{\mathbf{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y} = \mathbf{H}\mathbf{y}$
- Mean-Variance: $\mathbb{E}(\widehat{\mathbf{y}}) = \mathbf{X}\mathbb{E}(\widehat{\boldsymbol{\beta}}) = \mathbf{X}\boldsymbol{\beta}, \quad \text{var}(\widehat{\mathbf{y}}) = \sigma^2\mathbf{H}$
- Residuals: $\widehat{\mathbf{e}} = (\mathbf{y} - \widehat{\mathbf{y}}) = \mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}} = \{\mathbf{I}_n - \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\}\mathbf{y} = (\mathbf{I}_n - \mathbf{H})\mathbf{y}$
- Mean-Variance: $\mathbb{E}(\widehat{\mathbf{e}}) = (\mathbf{I}_n - \mathbf{H})\mathbf{X}\boldsymbol{\beta} = \mathbf{0}, \quad \text{var}(\widehat{\mathbf{e}}) = \mathbb{E}(\widehat{\mathbf{e}}\widehat{\mathbf{e}}^{\mathrm{T}}) = \sigma^2(\mathbf{I}_n - \mathbf{H})$
- $\text{cov}(\widehat{\mathbf{y}}, \widehat{\mathbf{e}}) = \text{cov}\{\mathbf{H}\mathbf{y}, (\mathbf{I}_n - \mathbf{H})\mathbf{y}\} = \sigma^2\mathbf{H}(\mathbf{I}_n - \mathbf{H}) = \sigma^2(\mathbf{H} - \mathbf{H}^2) = 0$

Ordinary Least Squares

$$\mathbf{y}^{n\times 1} = \mathbf{X}^{n\times p}\boldsymbol{\beta}^{p\times 1} + \mathbf{e}^{n\times 1}, \qquad \mathbb{E}(\mathbf{e}) = \mathbf{0},\ \mathrm{cov}(\mathbf{e}) = \sigma^2\mathbf{I}_n$$

- Minimize squared error loss:

$\widehat{\boldsymbol{\beta}} = \arg\min \sum_{i=1}^{n}(y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})^2 = \arg\min(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$

- $\mathbb{E}(\widehat{\boldsymbol{\beta}}) = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta}$
- $\mathrm{var}(\widehat{\boldsymbol{\beta}}) = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\sigma^2\mathbf{I}_n\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1} = \sigma^2(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$
  - Hat matrix: $\mathbf{H} = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}$
  - Idempotent: $\mathbf{H}^2 = \mathbf{H}, \quad (\mathbf{I}_n - \mathbf{H})^2 = (\mathbf{I}_n - \mathbf{H})$
  - $\mathrm{trace}(\mathbf{H}) = \mathrm{trace}\{\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\} = \mathrm{trace}\{\mathbf{X}^{\mathrm{T}}\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\} = p,$
  - $\mathrm{trace}(\mathbf{I}_n - \mathbf{H}) = \mathrm{trace}(\mathbf{I}_n) - \mathrm{trace}(\mathbf{H}) = (n - p)$
  - $\mathbf{HX} = \mathbf{X}$ and $(\mathbf{I}_n - \mathbf{H})\mathbf{X} = \{\mathbf{I}_n - \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\}\mathbf{X} = \mathbf{0}$
- Fitted values: $\widehat{\mathbf{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y} = \mathbf{H}\mathbf{y}$
- Mean-Variance: $\mathbb{E}(\widehat{\mathbf{y}}) = \mathbf{X}\mathbb{E}(\widehat{\boldsymbol{\beta}}) = \mathbf{X}\boldsymbol{\beta}, \quad \mathrm{var}(\widehat{\mathbf{y}}) = \sigma^2\mathbf{H}$
- Residuals: $\widehat{\mathbf{e}} = (\mathbf{y} - \widehat{\mathbf{y}}) = \mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}} = \{\mathbf{I}_n - \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\}\mathbf{y} = (\mathbf{I}_n - \mathbf{H})\mathbf{y}$
- Mean-Variance: $\mathbb{E}(\widehat{\mathbf{e}}) = (\mathbf{I}_n - \mathbf{H})\mathbf{X}\boldsymbol{\beta} = \mathbf{0}, \quad \mathrm{var}(\widehat{\mathbf{e}}) = \mathbb{E}(\widehat{\mathbf{e}}\widehat{\mathbf{e}}^{\mathrm{T}}) = \sigma^2(\mathbf{I}_n - \mathbf{H})$
- $\mathrm{cov}(\widehat{\mathbf{y}}, \widehat{\mathbf{e}}) = \mathrm{cov}\{\mathbf{H}\mathbf{y}, (\mathbf{I}_n - \mathbf{H})\mathbf{y}\} = \sigma^2\mathbf{H}(\mathbf{I}_n - \mathbf{H}) = \sigma^2(\mathbf{H} - \mathbf{H}^2) = \mathbf{0}$

$$\mathbf{y}^{n\times 1} = \mathbf{X}^{n\times p}\boldsymbol{\beta}^{p\times 1} + \mathbf{e}^{n\times 1}, \quad \mathbb{E}(\mathbf{e}) = \mathbf{0},\ \mathrm{cov}(\mathbf{e}) = \sigma^2\mathbf{I}_n$$

- Minimize squared error loss:

$$\widehat{\boldsymbol{\beta}} = \arg\min \textstyle\sum_{i=1}^{n}(y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})^2 = \arg\min(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$$

- $\mathbb{E}(\widehat{\boldsymbol{\beta}}) = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta}$
- $\mathrm{var}(\widehat{\boldsymbol{\beta}}) = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\sigma^2\mathbf{I}_n\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1} = \sigma^2(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$
    - Hat matrix: $\mathbf{H} = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}$
    - Idempotent: $\mathbf{H}^2 = \mathbf{H}, \quad (\mathbf{I}_n - \mathbf{H})^2 = (\mathbf{I}_n - \mathbf{H})$
    - $\mathrm{trace}(\mathbf{H}) = \mathrm{trace}\{\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\} = \mathrm{trace}\{\mathbf{X}^{\mathrm{T}}\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\} = p,$
    - $\mathrm{trace}(\mathbf{I}_n - \mathbf{H}) = \mathrm{trace}(\mathbf{I}_n) - \mathrm{trace}(\mathbf{H}) = (n - p)$
    - $\mathbf{HX} = \mathbf{X}$ and $(\mathbf{I}_n - \mathbf{H})\mathbf{X} = \{\mathbf{I}_n - \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\}\mathbf{X} = \mathbf{0}$
- Fitted values: $\widehat{\mathbf{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y} = \mathbf{H}\mathbf{y}$
- Mean-Variance: $\mathbb{E}(\widehat{\mathbf{y}}) = \mathbf{X}\mathbb{E}(\widehat{\boldsymbol{\beta}}) = \mathbf{X}\boldsymbol{\beta}, \quad \mathrm{var}(\widehat{\mathbf{y}}) = \sigma^2\mathbf{H}$
- Residuals: $\widehat{\mathbf{e}} = (\mathbf{y} - \widehat{\mathbf{y}}) = \mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}} = \{\mathbf{I}_n - \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\}\mathbf{y} = (\mathbf{I}_n - \mathbf{H})\mathbf{y}$
- Mean-Variance: $\mathbb{E}(\widehat{\mathbf{e}}) = (\mathbf{I}_n - \mathbf{H})\mathbf{X}\boldsymbol{\beta} = \mathbf{0}, \quad \mathrm{var}(\widehat{\mathbf{e}}) = \mathbb{E}(\widehat{\mathbf{e}}\widehat{\mathbf{e}}^{\mathrm{T}}) = \sigma^2(\mathbf{I}_n - \mathbf{H})$
- $\mathrm{cov}(\widehat{\mathbf{y}}, \widehat{\mathbf{e}}) = \mathrm{cov}\{\mathbf{H}\mathbf{y}, (\mathbf{I}_n - \mathbf{H})\mathbf{y}\} = \sigma^2\mathbf{H}(\mathbf{I}_n - \mathbf{H}) = \sigma^2(\mathbf{H} - \mathbf{H}^2) = \mathbf{0}$

Ordinary Least Squares

$$\mathbf{y}^{n\times 1} = \mathbf{X}^{n\times p}\boldsymbol{\beta}^{p\times 1} + \mathbf{e}^{n\times 1}, \qquad \mathbb{E}(\mathbf{e}) = \mathbf{0},\ \mathrm{cov}(\mathbf{e}) = \sigma^2\mathbf{I}_n$$

- Minimize squared error loss:

$$\widehat{\boldsymbol{\beta}} = \arg\min \textstyle\sum_{i=1}^{n}(y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})^2 = \arg\min(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$$

- $\mathbb{E}(\widehat{\boldsymbol{\beta}}) = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta}$
- $\mathrm{var}(\widehat{\boldsymbol{\beta}}) = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\sigma^2\mathbf{I}_n\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1} = \sigma^2(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$
  - Hat matrix: $\mathbf{H} = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}$
  - Idempotent: $\mathbf{H}^2 = \mathbf{H}, \quad (\mathbf{I}_n - \mathbf{H})^2 = (\mathbf{I}_n - \mathbf{H})$
  - $\mathrm{trace}(\mathbf{H}) = \mathrm{trace}\{\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\} = \mathrm{trace}\{\mathbf{X}^{\mathrm{T}}\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\} = p,$
  - $\mathrm{trace}(\mathbf{I}_n - \mathbf{H}) = \mathrm{trace}(\mathbf{I}_n) - \mathrm{trace}(\mathbf{H}) = (n - p)$
  - $\mathbf{HX} = \mathbf{X}$ and $(\mathbf{I}_n - \mathbf{H})\mathbf{X} = \{\mathbf{I}_n - \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\}\mathbf{X} = \mathbf{0}$
- Fitted values: $\widehat{\mathbf{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y} = \mathbf{H}\mathbf{y}$
- Mean-Variance: $\mathbb{E}(\widehat{\mathbf{y}}) = \mathbf{X}\mathbb{E}(\widehat{\boldsymbol{\beta}}) = \mathbf{X}\boldsymbol{\beta}, \quad \mathrm{var}(\widehat{\mathbf{y}}) = \sigma^2\mathbf{H}$
- Residuals: $\widehat{\mathbf{e}} = (\mathbf{y} - \widehat{\mathbf{y}}) = \mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}} = \{\mathbf{I}_n - \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\}\mathbf{y} = (\mathbf{I}_n - \mathbf{H})\mathbf{y}$
- Mean-Variance: $\mathbb{E}(\widehat{\mathbf{e}}) = (\mathbf{I}_n - \mathbf{H})\mathbf{X}\boldsymbol{\beta} = \mathbf{0}, \quad \mathrm{var}(\widehat{\mathbf{e}}) = \mathbb{E}(\widehat{\mathbf{e}}\widehat{\mathbf{e}}^{\mathrm{T}}) = \sigma^2(\mathbf{I}_n - \mathbf{H})$
- $\mathrm{cov}(\widehat{\mathbf{y}}, \widehat{\mathbf{e}}) = \mathrm{cov}\{\mathbf{H}\mathbf{y}, (\mathbf{I}_n - \mathbf{H})\mathbf{y}\} = \sigma^2\mathbf{H}(\mathbf{I}_n - \mathbf{H}) = \sigma^2(\mathbf{H} - \mathbf{H}^2) = \mathbf{0}$

Ordinary Least Squares

$$\mathbf{y}^{n\times 1} = \mathbf{X}^{n\times p}\boldsymbol{\beta}^{p\times 1} + \mathbf{e}^{n\times 1}, \qquad \mathbb{E}(\mathbf{e}) = \mathbf{0},\ \mathrm{cov}(\mathbf{e}) = \sigma^2\mathbf{I}_n$$

- Minimize squared error loss:

$$\widehat{\boldsymbol{\beta}} = \arg\min \textstyle\sum_{i=1}^{n}(y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})^2 = \arg\min(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$$

- $\mathbb{E}(\widehat{\boldsymbol{\beta}}) = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta}$
- $\mathrm{var}(\widehat{\boldsymbol{\beta}}) = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\sigma^2\mathbf{I}_n\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1} = \sigma^2(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$
    - Hat matrix: $\mathbf{H} = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}$
    - Idempotent: $\mathbf{H}^2 = \mathbf{H}, \quad (\mathbf{I}_n - \mathbf{H})^2 = (\mathbf{I}_n - \mathbf{H})$
    - $\mathrm{trace}(\mathbf{H}) = \mathrm{trace}\{\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\} = \mathrm{trace}\{\mathbf{X}^{\mathrm{T}}\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\} = p,$
    - $\mathrm{trace}(\mathbf{I}_n - \mathbf{H}) = \mathrm{trace}(\mathbf{I}_n) - \mathrm{trace}(\mathbf{H}) = (n-p)$
    - $\mathbf{HX} = \mathbf{X}$ and $(\mathbf{I}_n - \mathbf{H})\mathbf{X} = \{\mathbf{I}_n - \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\}\mathbf{X} = \mathbf{0}$
- Fitted values: $\widehat{\mathbf{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y} = \mathbf{H}\mathbf{y}$
- Mean-Variance: $\mathbb{E}(\widehat{\mathbf{y}}) = \mathbf{X}\mathbb{E}(\widehat{\boldsymbol{\beta}}) = \mathbf{X}\boldsymbol{\beta}, \quad \mathrm{var}(\widehat{\mathbf{y}}) = \sigma^2\mathbf{H}$
- Residuals: $\widehat{\mathbf{e}} = (\mathbf{y} - \widehat{\mathbf{y}}) = \mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}} = \{\mathbf{I}_n - \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\}\mathbf{y} = (\mathbf{I}_n - \mathbf{H})\mathbf{y}$
- Mean-Variance: $\mathbb{E}(\widehat{\mathbf{e}}) = (\mathbf{I}_n - \mathbf{H})\mathbf{X}\boldsymbol{\beta} = \mathbf{0}, \quad \mathrm{var}(\widehat{\mathbf{e}}) = \mathbb{E}(\widehat{\mathbf{e}}\widehat{\mathbf{e}}^{\mathrm{T}}) = \sigma^2(\mathbf{I}_n - \mathbf{H})$
- $\mathrm{cov}(\widehat{\mathbf{y}}, \widehat{\mathbf{e}}) = \mathrm{cov}\{\mathbf{H}\mathbf{y}, (\mathbf{I}_n - \mathbf{H})\mathbf{y}\} = \sigma^2\mathbf{H}(\mathbf{I}_n - \mathbf{H}) = \sigma^2(\mathbf{H} - \mathbf{H}^2) = \mathbf{0}$

Ordinary Least Squares

$$\mathbf{y}^{n\times 1} = \mathbf{X}^{n\times p}\boldsymbol{\beta}^{p\times 1} + \mathbf{e}^{n\times 1}, \qquad \mathbb{E}(\mathbf{e}) = \mathbf{0},\ \mathrm{cov}(\mathbf{e}) = \sigma^2\mathbf{I}_n$$

- Minimize squared error loss:

$\widehat{\boldsymbol{\beta}} = \arg\min \sum_{i=1}^{n}(y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})^2 = \arg\min(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$

- $\mathbb{E}(\widehat{\boldsymbol{\beta}}) = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta}$
- $\mathrm{var}(\widehat{\boldsymbol{\beta}}) = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\sigma^2\mathbf{I}_n\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1} = \sigma^2(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$
  - Hat matrix: $\mathbf{H} = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}$
  - Idempotent: $\mathbf{H}^2 = \mathbf{H}, \quad (\mathbf{I}_n - \mathbf{H})^2 = (\mathbf{I}_n - \mathbf{H})$
  - $\mathrm{trace}(\mathbf{H}) = \mathrm{trace}\{\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\} = \mathrm{trace}\{\mathbf{X}^{\mathrm{T}}\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\} = p,$
  - $\mathrm{trace}(\mathbf{I}_n - \mathbf{H}) = \mathrm{trace}(\mathbf{I}_n) - \mathrm{trace}(\mathbf{H}) = (n - p)$
  - $\mathbf{H}\mathbf{X} = \mathbf{X}$ and $(\mathbf{I}_n - \mathbf{H})\mathbf{X} = \{\mathbf{I}_n - \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\}\mathbf{X} = \mathbf{0}$
- Fitted values: $\widehat{\mathbf{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y} = \mathbf{H}\mathbf{y}$
- Mean-Variance: $\mathbb{E}(\widehat{\mathbf{y}}) = \mathbf{X}\mathbb{E}(\widehat{\boldsymbol{\beta}}) = \mathbf{X}\boldsymbol{\beta}, \quad \mathrm{var}(\widehat{\mathbf{y}}) = \sigma^2\mathbf{H}$
- Residuals: $\widehat{\mathbf{e}} = (\mathbf{y} - \widehat{\mathbf{y}}) = \mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}} = \{\mathbf{I}_n - \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\}\mathbf{y} = (\mathbf{I}_n - \mathbf{H})\mathbf{y}$
- Mean-Variance: $\mathbb{E}(\widehat{\mathbf{e}}) = (\mathbf{I}_n - \mathbf{H})\mathbf{X}\boldsymbol{\beta} = \mathbf{0}, \quad \mathrm{var}(\widehat{\mathbf{e}}) = \mathbb{E}(\widehat{\mathbf{e}}\widehat{\mathbf{e}}^{\mathrm{T}}) = \sigma^2(\mathbf{I}_n - \mathbf{H})$
- $\mathrm{cov}(\widehat{\mathbf{y}}, \widehat{\mathbf{e}}) = \mathrm{cov}\{\mathbf{H}\mathbf{y}, (\mathbf{I}_n - \mathbf{H})\mathbf{y}\} = \sigma^2\mathbf{H}(\mathbf{I}_n - \mathbf{H}) = \sigma^2(\mathbf{H} - \mathbf{H}^2) = 0$

Ordinary Least Squares

$$\mathbf{y}^{n\times 1} = \mathbf{X}^{n\times p}\boldsymbol{\beta}^{p\times 1} + \mathbf{e}^{n\times 1}, \qquad \mathbb{E}(\mathbf{e}) = \mathbf{0},\ \mathrm{cov}(\mathbf{e}) = \sigma^2\mathbf{I}_n$$

- Minimize squared error loss:

$\widehat{\boldsymbol{\beta}} = \arg\min \sum_{i=1}^{n}(y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})^2 = \arg\min(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$

- $\mathbb{E}(\widehat{\boldsymbol{\beta}}) = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta}$
- $\mathrm{var}(\widehat{\boldsymbol{\beta}}) = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\sigma^2\mathbf{I}_n\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1} = \sigma^2(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$
    - Hat matrix: $\mathbf{H} = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}$
    - Idempotent: $\mathbf{H}^2 = \mathbf{H}, \quad (\mathbf{I}_n - \mathbf{H})^2 = (\mathbf{I}_n - \mathbf{H})$
    - $\mathrm{trace}(\mathbf{H}) = \mathrm{trace}\{\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\} = \mathrm{trace}\{\mathbf{X}^{\mathrm{T}}\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\} = p,$
    - $\mathrm{trace}(\mathbf{I}_n - \mathbf{H}) = \mathrm{trace}(\mathbf{I}_n) - \mathrm{trace}(\mathbf{H}) = (n-p)$
    - $\mathbf{HX} = \mathbf{X}$ and $(\mathbf{I}_n - \mathbf{H})\mathbf{X} = \{\mathbf{I}_n - \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\}\mathbf{X} = \mathbf{0}$
- Fitted values: $\widehat{\mathbf{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y} = \mathbf{H}\mathbf{y}$
- Mean-Variance: $\mathbb{E}(\widehat{\mathbf{y}}) = \mathbf{X}\mathbb{E}(\widehat{\boldsymbol{\beta}}) = \mathbf{X}\boldsymbol{\beta}, \quad \mathrm{var}(\widehat{\mathbf{y}}) = \sigma^2\mathbf{H}$
- Residuals: $\widehat{\mathbf{e}} = (\mathbf{y} - \widehat{\mathbf{y}}) = \mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}} = \{\mathbf{I}_n - \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\}\mathbf{y} = (\mathbf{I}_n - \mathbf{H})\mathbf{y}$
- Mean-Variance: $\mathbb{E}(\widehat{\mathbf{e}}) = (\mathbf{I}_n - \mathbf{H})\mathbf{X}\boldsymbol{\beta} = \mathbf{0}, \quad \mathrm{var}(\widehat{\mathbf{e}}) = \mathbb{E}(\widehat{\mathbf{e}}\widehat{\mathbf{e}}^{\mathrm{T}}) = \sigma^2(\mathbf{I}_n - \mathbf{H})$
- $\mathrm{cov}(\widehat{\mathbf{y}}, \widehat{\mathbf{e}}) = \mathrm{cov}\{\mathbf{H}\mathbf{y}, (\mathbf{I}_n - \mathbf{H})\mathbf{y}\} = \sigma^2\mathbf{H}(\mathbf{I}_n - \mathbf{H}) = \sigma^2(\mathbf{H} - \mathbf{H}^2) = \mathbf{0}$

Ordinary Least Squares

$$\mathbf{y}^{n\times 1} = \mathbf{X}^{n\times p}\boldsymbol{\beta}^{p\times 1} + \mathbf{e}^{n\times 1}, \qquad \mathbb{E}(\mathbf{e}) = \mathbf{0},\ \mathrm{cov}(\mathbf{e}) = \sigma^2\mathbf{I}_n$$

- Minimize squared error loss:

$$\widehat{\boldsymbol{\beta}} = \arg\min \textstyle\sum_{i=1}^{n}(y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})^2 = \arg\min(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$$

- Fitted values: $\widehat{\mathbf{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y} = \mathbf{H}\mathbf{y}$
- Mean-Variance: $\mathbb{E}(\widehat{\mathbf{y}}) = \mathbf{H}\mathbb{E}(\mathbf{y}) = \mathbf{H}\mathbf{X}\boldsymbol{\beta} = \mathbf{X}\boldsymbol{\beta}, \quad \mathrm{var}(\widehat{\mathbf{y}}) = \sigma^2\mathbf{H}$
- Residuals: $\widehat{\mathbf{e}} = (\mathbf{y} - \widehat{\mathbf{y}}) = \mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}} = \{\mathbf{I}_n - \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\}\mathbf{y} = (\mathbf{I}_n - \mathbf{H})\mathbf{y}$
- Mean-Variance: $\mathbb{E}(\widehat{\mathbf{e}}) = (\mathbf{I}_n - \mathbf{H})\mathbf{X}\boldsymbol{\beta} = \mathbf{0}, \quad \mathrm{var}(\widehat{\mathbf{e}}) = \mathbb{E}(\widehat{\mathbf{e}}\widehat{\mathbf{e}}^{\mathrm{T}}) = \sigma^2(\mathbf{I}_n - \mathbf{H})$
- $\mathrm{cov}(\widehat{\mathbf{y}}, \widehat{\mathbf{e}}) = \mathrm{cov}\{\mathbf{H}\mathbf{y}, (\mathbf{I}_n - \mathbf{H})\mathbf{y}\} = \sigma^2\mathbf{H}(\mathbf{I}_n - \mathbf{H}) = \sigma^2(\mathbf{H} - \mathbf{H}^2) = \mathbf{0}$

Ordinary Least Squares

$$\mathbf{y}^{n\times 1} = \mathbf{X}^{n\times p}\boldsymbol{\beta}^{p\times 1} + \mathbf{e}^{n\times 1}, \qquad \mathbb{E}(\mathbf{e}) = \mathbf{0},\ \mathrm{cov}(\mathbf{e}) = \sigma^2\mathbf{I}_n$$

- Minimize squared error loss:

$$\widehat{\boldsymbol{\beta}} = \arg\min \textstyle\sum_{i=1}^{n}(y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})^2 = \arg\min(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$$

- Fitted values: $\widehat{\mathbf{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y} = \mathbf{H}\mathbf{y}$
- Mean-Variance: $\mathbb{E}(\widehat{\mathbf{y}}) = \mathbf{H}\mathbb{E}(\mathbf{y}) = \mathbf{H}\mathbf{X}\boldsymbol{\beta} = \mathbf{X}\boldsymbol{\beta}, \quad \mathrm{var}(\widehat{\mathbf{y}}) = \sigma^2\mathbf{H}$
- Residuals: $\widehat{\mathbf{e}} = (\mathbf{y} - \widehat{\mathbf{y}}) = \mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}} = \{\mathbf{I}_n - \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\}\mathbf{y} = (\mathbf{I}_n - \mathbf{H})\mathbf{y}$
- Mean-Variance: $\mathbb{E}(\widehat{\mathbf{e}}) = (\mathbf{I}_n - \mathbf{H})\mathbf{X}\boldsymbol{\beta} = \mathbf{0}, \quad \mathrm{var}(\widehat{\mathbf{e}}) = \mathbb{E}(\widehat{\mathbf{e}}\widehat{\mathbf{e}}^{\mathrm{T}}) = \sigma^2(\mathbf{I}_n - \mathbf{H})$
- $\mathrm{cov}(\widehat{\mathbf{y}}, \widehat{\mathbf{e}}) = \mathrm{cov}\{\mathbf{H}\mathbf{y}, (\mathbf{I}_n - \mathbf{H})\mathbf{y}\} = \sigma^2\mathbf{H}(\mathbf{I}_n - \mathbf{H}) = \sigma^2(\mathbf{H} - \mathbf{H}^2) = \mathbf{0}$

- Let $\widehat{\sigma}^2 = (\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}})^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}})/n$
- $\mathbb{E}(\widehat{\sigma}^2) = \mathbb{E}\left\{(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}})^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}})/n\right\} = \mathbb{E}(\widehat{\mathbf{e}}^{\mathrm{T}}\widehat{\mathbf{e}})/n = \mathbb{E}\{\mathrm{trace}(\widehat{\mathbf{e}}^{\mathrm{T}}\widehat{\mathbf{e}})\}/n$

  $= \mathbb{E}\{\mathrm{trace}(\widehat{\mathbf{e}}\widehat{\mathbf{e}}^{\mathrm{T}})\}/n = \mathrm{trace}\{\mathbb{E}(\widehat{\mathbf{e}}\widehat{\mathbf{e}}^{\mathrm{T}})\}/n = \mathrm{trace}\{\sigma^2(\mathbf{I}_n - \mathbf{H})\}/n = \sigma^2(n-p)/n$
- $\mathbb{E}(s^2) = \mathbb{E}\{n\widehat{\sigma}^2/(n-p)\} = \mathbb{E}\left\{(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}})^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}})/(n-p)\right\} = \sigma^2$

Ordinary Least Squares

$$\mathbf{y}^{n\times 1} = \mathbf{X}^{n\times p}\boldsymbol{\beta}^{p\times 1} + \mathbf{e}^{n\times 1}, \qquad \mathbb{E}(\mathbf{e}) = \mathbf{0},\ \text{cov}(\mathbf{e}) = \sigma^2\mathbf{I}_n$$

- Minimize squared error loss:

$\widehat{\boldsymbol{\beta}} = \arg\min \sum_{i=1}^{n}(y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})^2 = \arg\min(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$

- Fitted values: $\widehat{\mathbf{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y} = \mathbf{H}\mathbf{y}$
- Mean-Variance: $\mathbb{E}(\widehat{\mathbf{y}}) = \mathbf{H}\mathbb{E}(\mathbf{y}) = \mathbf{H}\mathbf{X}\boldsymbol{\beta} = \mathbf{X}\boldsymbol{\beta}, \quad \text{var}(\widehat{\mathbf{y}}) = \sigma^2\mathbf{H}$
- Residuals: $\widehat{\mathbf{e}} = (\mathbf{y} - \widehat{\mathbf{y}}) = \mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}} = \{\mathbf{I}_n - \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\}\mathbf{y} = (\mathbf{I}_n - \mathbf{H})\mathbf{y}$
- Mean-Variance: $\mathbb{E}(\widehat{\mathbf{e}}) = (\mathbf{I}_n - \mathbf{H})\mathbf{X}\boldsymbol{\beta} = \mathbf{0}, \quad \text{var}(\widehat{\mathbf{e}}) = \mathbb{E}(\widehat{\mathbf{e}}\widehat{\mathbf{e}}^{\mathrm{T}}) = \sigma^2(\mathbf{I}_n - \mathbf{H})$
- $\text{cov}(\widehat{\mathbf{y}}, \widehat{\mathbf{e}}) = \text{cov}\{\mathbf{H}\mathbf{y}, (\mathbf{I}_n - \mathbf{H})\mathbf{y}\} = \sigma^2\mathbf{H}(\mathbf{I}_n - \mathbf{H}) = \sigma^2(\mathbf{H} - \mathbf{H}^2) = \mathbf{0}$

- Let $\widehat{\sigma}^2 = (\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}})^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}})/n$
- $\mathbb{E}(\widehat{\sigma}^2) = \mathbb{E}\left\{(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}})^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}})/n\right\} = \mathbb{E}(\widehat{\mathbf{e}}^{\mathrm{T}}\widehat{\mathbf{e}})/n = \mathbb{E}\{\text{trace}(\widehat{\mathbf{e}}^{\mathrm{T}}\widehat{\mathbf{e}})\}/n$

  $= \mathbb{E}\{\text{trace}(\widehat{\mathbf{e}}\widehat{\mathbf{e}}^{\mathrm{T}})\}/n = \text{trace}\{\mathbb{E}(\widehat{\mathbf{e}}\widehat{\mathbf{e}}^{\mathrm{T}})\}/n = \text{trace}\{\sigma^2(\mathbf{I}_n - \mathbf{H})\}/n = \sigma^2(n-p)/n$
- $\mathbb{E}(s^2) = \mathbb{E}\{n\widehat{\sigma}^2/(n-p)\} = \mathbb{E}\left\{(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}})^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}})/(n-p)\right\} = \sigma^2$

Ordinary Least Squares

$$\mathbf{y}^{n\times 1} = \mathbf{X}^{n\times p}\boldsymbol{\beta}^{p\times 1} + \mathbf{e}^{n\times 1}, \qquad \mathbb{E}(\mathbf{e}) = \mathbf{0},\ \operatorname{cov}(\mathbf{e}) = \sigma^2\mathbf{I}_n$$

- Minimize squared error loss:

$\widehat{\boldsymbol{\beta}} = \arg\min \sum_{i=1}^{n}(y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})^2 = \arg\min(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$

- Fitted values: $\widehat{\mathbf{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y} = \mathbf{H}\mathbf{y}$
- Mean-Variance: $\mathbb{E}(\widehat{\mathbf{y}}) = \mathbf{H}\mathbb{E}(\mathbf{y}) = \mathbf{H}\mathbf{X}\boldsymbol{\beta} = \mathbf{X}\boldsymbol{\beta}, \quad \operatorname{var}(\widehat{\mathbf{y}}) = \sigma^2\mathbf{H}$
- Residuals: $\widehat{\mathbf{e}} = (\mathbf{y} - \widehat{\mathbf{y}}) = \mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}} = \{\mathbf{I}_n - \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\}\mathbf{y} = (\mathbf{I}_n - \mathbf{H})\mathbf{y}$
- Mean-Variance: $\mathbb{E}(\widehat{\mathbf{e}}) = (\mathbf{I}_n - \mathbf{H})\mathbf{X}\boldsymbol{\beta} = \mathbf{0}, \quad \operatorname{var}(\widehat{\mathbf{e}}) = \mathbb{E}(\widehat{\mathbf{e}}\widehat{\mathbf{e}}^{\mathrm{T}}) = \sigma^2(\mathbf{I}_n - \mathbf{H})$
- $\operatorname{cov}(\widehat{\mathbf{y}}, \widehat{\mathbf{e}}) = \operatorname{cov}\{\mathbf{H}\mathbf{y}, (\mathbf{I}_n - \mathbf{H})\mathbf{y}\} = \sigma^2\mathbf{H}(\mathbf{I}_n - \mathbf{H}) = \sigma^2(\mathbf{H} - \mathbf{H}^2) = \mathbf{0}$

- Let $\widehat{\sigma}^2 = (\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}})^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}})/n$
- $\mathbb{E}(\widehat{\sigma}^2) = \mathbb{E}\left\{(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}})^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}})/n\right\} = \mathbb{E}(\widehat{\mathbf{e}}^{\mathrm{T}}\widehat{\mathbf{e}})/n = \mathbb{E}\{\operatorname{trace}(\widehat{\mathbf{e}}^{\mathrm{T}}\widehat{\mathbf{e}})\}/n$

  $= \mathbb{E}\{\operatorname{trace}(\widehat{\mathbf{e}}\widehat{\mathbf{e}}^{\mathrm{T}})\}/n = \operatorname{trace}\{\mathbb{E}(\widehat{\mathbf{e}}\widehat{\mathbf{e}}^{\mathrm{T}})\}/n = \operatorname{trace}\{\sigma^2(\mathbf{I}_n - \mathbf{H})\}/n = \sigma^2(n-p)/n$
- $\mathbb{E}(s^2) = \mathbb{E}\{n\widehat{\sigma}^2/(n-p)\} = \mathbb{E}\left\{(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}})^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}})/(n-p)\right\} = \sigma^2$

Ordinary Least Squares

$$\mathbf{y}^{n\times 1} = \mathbf{X}^{n\times p}\boldsymbol{\beta}^{p\times 1} + \mathbf{e}^{n\times 1}, \qquad \mathbb{E}(\mathbf{e}) = \mathbf{0},\ \mathrm{cov}(\mathbf{e}) = \sigma^2\mathbf{I}_n$$

$$\text{OLS:} \quad \widehat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$$

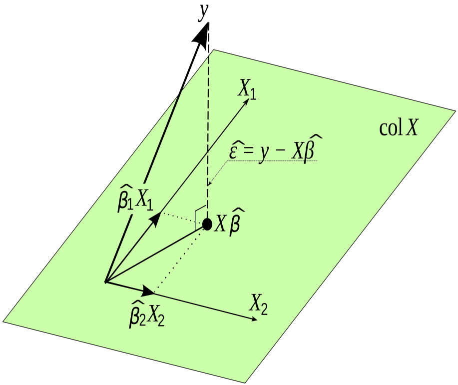
- Fitted values: $\widehat{\mathbf{y}} = \mathbf{H}\mathbf{y}$, and residuals: $\widehat{\mathbf{e}} = (\mathbf{y} - \widehat{\mathbf{y}}) = (\mathbf{I}_n - \mathbf{H})\mathbf{y}$

$$y = \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

$$\mathbf{y}^{n\times 1} = \mathbf{X}^{n\times p}\boldsymbol{\beta}^{p\times 1} + \mathbf{e}^{n\times 1}, \quad \mathbb{E}(\mathbf{e}) = \mathbf{0},\ \mathrm{cov}(\mathbf{e}) = \sigma^2 \mathbf{I}_n$$

$$\text{OLS:} \quad \widehat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$$

- Fitted values: $\widehat{\mathbf{y}} = \mathbf{H}\mathbf{y}$, and residuals: $\widehat{\mathbf{e}} = (\mathbf{y} - \widehat{\mathbf{y}}) = (\mathbf{I}_n - \mathbf{H})\mathbf{y}$



$$y = \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

Maximum Likelihood

$$\mathbf{y}^{n\times 1} = \mathbf{X}^{n\times p}\boldsymbol{\beta}^{p\times 1} + \mathbf{e}^{n\times 1}$$

- Minimize squared error loss:

$$\widehat{\boldsymbol{\beta}}_{OLS} = \arg\min \textstyle\sum_{i=1}^{n}(y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})^2 = \arg\min(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})$$

- $$\min_{\boldsymbol{\beta}}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) \equiv \min_{\boldsymbol{\beta}}\left\{\frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\right\}$$
$$\equiv \max_{\boldsymbol{\beta}}\left\{-\frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\right\}$$

- Maximize a normal log-likelihood function:

$$\max_{\boldsymbol{\beta},\sigma^2}\mathcal{L}(\boldsymbol{\beta},\sigma^2) \equiv \min_{\boldsymbol{\beta},\sigma^2}\left\{\frac{n}{2}\log\sigma^2 + \frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\right\}$$

- MLE: $$\frac{\partial\mathcal{L}(\boldsymbol{\beta},\sigma^2)}{\partial\boldsymbol{\beta}} = \frac{1}{\sigma^2}(\mathbf{X}^{\mathrm{T}}\mathbf{y} - \mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{\beta}) = \mathbf{0}$$

and $$\frac{\partial\mathcal{L}(\boldsymbol{\beta},\sigma^2)}{\partial\sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) = 0$$

$$\Rightarrow \widehat{\boldsymbol{\beta}}_{OLS} = \widehat{\boldsymbol{\beta}}_{MLE} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}, \qquad \widehat{\sigma}^2_{MLE} = (\mathbf{y}-\mathbf{X}\widehat{\boldsymbol{\beta}})^{\mathrm{T}}(\mathbf{y}-\mathbf{X}\widehat{\boldsymbol{\beta}})/n$$

Maximum Likelihood

$$\mathbf{y}^{n\times 1} = \mathbf{X}^{n\times p}\boldsymbol{\beta}^{p\times 1} + \mathbf{e}^{n\times 1}, \qquad \mathbf{e} \sim \mathrm{MVN}_n(\mathbf{0}, \sigma^2\mathbf{I}_n)$$

- Minimize squared error loss:

$$\widehat{\boldsymbol{\beta}}_{OLS} = \arg\min \textstyle\sum_{i=1}^{n}(y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})^2 = \arg\min(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})$$

- $$\min_{\boldsymbol{\beta}}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) \equiv \min_{\boldsymbol{\beta}}\left\{\frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\right\}$$

$$\equiv \max_{\boldsymbol{\beta}}\left\{-\frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\right\}$$

- Maximize a normal log-likelihood function:

$$\max_{\boldsymbol{\beta},\sigma^2}\mathcal{L}(\boldsymbol{\beta},\sigma^2) \equiv \min_{\boldsymbol{\beta},\sigma^2}\left\{\frac{n}{2}\log\sigma^2 + \frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\right\}$$

- MLE: $$\frac{\partial\mathcal{L}(\boldsymbol{\beta},\sigma^2)}{\partial\boldsymbol{\beta}} = \frac{1}{\sigma^2}(\mathbf{X}^{\mathrm{T}}\mathbf{y} - \mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{\beta}) = \mathbf{0}$$

and $$\frac{\partial\mathcal{L}(\boldsymbol{\beta},\sigma^2)}{\partial\sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) = 0$$

$$\Rightarrow \widehat{\boldsymbol{\beta}}_{OLS} = \widehat{\boldsymbol{\beta}}_{MLE} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}, \qquad \widehat{\sigma}^2_{MLE} = (\mathbf{y}-\mathbf{X}\widehat{\boldsymbol{\beta}})^{\mathrm{T}}(\mathbf{y}-\mathbf{X}\widehat{\boldsymbol{\beta}})/n$$

Maximum Likelihood

$$\mathbf{y}^{n\times 1} = \mathbf{X}^{n\times p}\boldsymbol{\beta}^{p\times 1} + \mathbf{e}^{n\times 1}, \qquad \mathbf{e} \sim \mathrm{MVN}_n(\mathbf{0}, \sigma^2\mathbf{I}_n)$$

- Minimize squared error loss:

$$\widehat{\boldsymbol{\beta}}_{OLS} = \arg\min \textstyle\sum_{i=1}^{n}(y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})^2 = \arg\min(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

- $$\begin{aligned}\min_{\boldsymbol{\beta}}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) &\equiv \min_{\boldsymbol{\beta}}\left\{\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right\} \\ &\equiv \max_{\boldsymbol{\beta}}\left\{-\frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right\}\end{aligned}$$

- Maximize a normal log-likelihood function:

$$\max_{\boldsymbol{\beta},\sigma^2}\mathcal{L}(\boldsymbol{\beta},\sigma^2) \equiv \min_{\boldsymbol{\beta},\sigma^2}\left\{\frac{n}{2}\log\sigma^2 + \frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right\}$$

- MLE: $$\frac{\partial\mathcal{L}(\boldsymbol{\beta},\sigma^2)}{\partial\boldsymbol{\beta}} = \frac{1}{\sigma^2}(\mathbf{X}^{\mathrm{T}}\mathbf{y} - \mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{\beta}) = \mathbf{0}$$

  and $$\frac{\partial\mathcal{L}(\boldsymbol{\beta},\sigma^2)}{\partial\sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = 0$$

  $$\Rightarrow \widehat{\boldsymbol{\beta}}_{OLS} = \widehat{\boldsymbol{\beta}}_{MLE} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}, \qquad \widehat{\sigma}^2_{MLE} = (\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}})^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}})/n$$

Maximum Likelihood

$$\mathbf{y}^{n\times 1} = \mathbf{X}^{n\times p}\boldsymbol{\beta}^{p\times 1} + \mathbf{e}^{n\times 1}, \qquad \mathbf{e} \sim \mathrm{MVN}_n(\mathbf{0}, \sigma^2\mathbf{I}_n)$$

$$\text{MLE:} \quad \widehat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}, \qquad \widehat{\sigma}^2 = \tfrac{1}{n}(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}})^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}})$$

- Log-likelihood: $\mathcal{L}(\boldsymbol{\beta}, \sigma^2) = \left\{ -\frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \right\}$
- Gradients: $$\frac{\partial\mathcal{L}(\boldsymbol{\beta}, \sigma^2)}{\partial\boldsymbol{\beta}} = \frac{1}{\sigma^2}(\mathbf{X}^{\mathrm{T}}\mathbf{y} - \mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{\beta})$$
  $$\frac{\partial\mathcal{L}(\boldsymbol{\beta}, \sigma^2)}{\partial\sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

Maximum Likelihood

$$\mathbf{y}^{n\times 1} = \mathbf{X}^{n\times p}\boldsymbol{\beta}^{p\times 1} + \mathbf{e}^{n\times 1}, \qquad \mathbf{e} \sim \mathrm{MVN}_n(\mathbf{0}, \sigma^2\mathbf{I}_n)$$

$$\text{MLE:} \quad \widehat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}, \qquad \widehat{\sigma}^2 = \tfrac{1}{n}(\mathbf{y}-\mathbf{X}\widehat{\boldsymbol{\beta}})^{\mathrm{T}}(\mathbf{y}-\mathbf{X}\widehat{\boldsymbol{\beta}})$$

- Log-likelihood: $\mathcal{L}(\boldsymbol{\beta},\sigma^2) = \left\{-\frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\right\}$
- Gradients:
$$\frac{\partial\mathcal{L}(\boldsymbol{\beta},\sigma^2)}{\partial\boldsymbol{\beta}} = \frac{1}{\sigma^2}(\mathbf{X}^{\mathrm{T}}\mathbf{y} - \mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{\beta})$$
$$\frac{\partial\mathcal{L}(\boldsymbol{\beta},\sigma^2)}{\partial\sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})$$
- Hessian:
$$-\frac{\partial^2\mathcal{L}(\boldsymbol{\beta},\sigma^2)}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^{\mathrm{T}}} = \frac{1}{\sigma^2}\mathbf{X}^{\mathrm{T}}\mathbf{X}$$
$$-\frac{\partial\mathcal{L}(\boldsymbol{\beta},\sigma^2)}{\partial(\sigma^2)^2} = -\frac{n}{2\sigma^4} + \frac{1}{\sigma^6}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})$$
$$-\frac{\partial\mathcal{L}(\boldsymbol{\beta},\sigma^2)}{\partial\boldsymbol{\beta}\partial\sigma^2} = \frac{-1}{\sigma^4}(\mathbf{X}^{\mathrm{T}}\mathbf{y} - \mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{\beta})$$
- Fisher Information: $\mathbf{I}(\boldsymbol{\beta},\sigma^2) = \begin{bmatrix} \frac{1}{\sigma^2}\mathbf{X}^{\mathrm{T}}\mathbf{X} & 0 \\ 0 & \frac{n}{2\sigma^4} \end{bmatrix}$

Maximum Likelihood

$$\mathbf{y}^{n\times 1} = \mathbf{X}^{n\times p}\boldsymbol{\beta}^{p\times 1} + \mathbf{e}^{n\times 1}, \qquad \mathbf{e} \sim \mathrm{MVN}_n(\mathbf{0}, \sigma^2\mathbf{I}_n)$$

$$\text{MLE:} \quad \widehat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}, \qquad \widehat{\sigma}^2 = \tfrac{1}{n}(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}})^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}})$$

- Log-likelihood: $\mathcal{L}(\boldsymbol{\beta}, \sigma^2) = \left\{-\frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right\}$
- Gradients: $$\frac{\partial\mathcal{L}(\boldsymbol{\beta}, \sigma^2)}{\partial\boldsymbol{\beta}} = \frac{1}{\sigma^2}(\mathbf{X}^{\mathrm{T}}\mathbf{y} - \mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{\beta})$$ $$\frac{\partial\mathcal{L}(\boldsymbol{\beta}, \sigma^2)}{\partial\sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$
- Hessian: $$-\frac{\partial^2\mathcal{L}(\boldsymbol{\beta}, \sigma^2)}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^{\mathrm{T}}} = \frac{1}{\sigma^2}\mathbf{X}^{\mathrm{T}}\mathbf{X}$$ $$-\frac{\partial\mathcal{L}(\boldsymbol{\beta}, \sigma^2)}{\partial(\sigma^2)^2} = -\frac{n}{2\sigma^4} + \frac{1}{\sigma^6}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$ $$-\frac{\partial\mathcal{L}(\boldsymbol{\beta}, \sigma^2)}{\partial\boldsymbol{\beta}\partial\sigma^2} = \frac{-1}{\sigma^4}(\mathbf{X}^{\mathrm{T}}\mathbf{y} - \mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{\beta})$$
- Fisher Information: $\mathbf{I}(\boldsymbol{\beta}, \sigma^2) = \begin{bmatrix} \frac{1}{\sigma^2}\mathbf{X}^{\mathrm{T}}\mathbf{X} & 0 \\ 0 & \frac{n}{2\sigma^4} \end{bmatrix}$

Maximum Likelihood

$$\mathbf{y}^{n\times 1} = \mathbf{X}^{n\times p}\boldsymbol{\beta}^{p\times 1} + \mathbf{e}^{n\times 1}, \qquad \mathbf{e} \sim \mathrm{MVN}_n(\mathbf{0}, \sigma^2\mathbf{I}_n)$$

$$\text{MLE:} \quad \widehat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}, \qquad \widehat{\sigma}^2 = \tfrac{1}{n}(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}})^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}})$$

- Log-likelihood: $\mathcal{L}(\boldsymbol{\beta}, \sigma^2) = \left\{ -\dfrac{n}{2}\log\sigma^2 - \dfrac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \right\}$
- Fisher information: $\mathbf{I}(\boldsymbol{\beta}, \sigma^2) = \begin{bmatrix} \frac{1}{\sigma^2}\mathbf{X}^{\mathrm{T}}\mathbf{X} & 0 \\ 0 & \frac{n}{2\sigma^4} \end{bmatrix}$

Maximum Likelihood

$$\mathbf{y}^{n\times 1} = \mathbf{X}^{n\times p}\boldsymbol{\beta}^{p\times 1} + \mathbf{e}^{n\times 1}, \qquad \mathbf{e} \sim \mathrm{MVN}_n(\mathbf{0}, \sigma^2 \mathbf{I}_n)$$

$$\text{MLE:} \quad \widehat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}, \qquad \widehat{\sigma}^2 = \tfrac{1}{n}(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}})^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}})$$

- Log-likelihood: $\mathcal{L}(\boldsymbol{\beta}, \sigma^2) = \left\{ -\frac{n}{2}\log \sigma^2 - \frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \right\}$
- Fisher information: $\mathbf{I}(\boldsymbol{\beta}, \sigma^2) = \begin{bmatrix} \frac{1}{\sigma^2}\mathbf{X}^{\mathrm{T}}\mathbf{X} & 0 \\ 0 & \frac{n}{2\sigma^4} \end{bmatrix}$
- Asymptotic distribution:
$\begin{pmatrix} \widehat{\boldsymbol{\beta}} \\ \widehat{\sigma}^2 \end{pmatrix} \approx \mathrm{MVN}_{p+1}\left[\begin{pmatrix} \boldsymbol{\beta} \\ \sigma^2 \end{pmatrix}, \ \mathbf{I}(\boldsymbol{\beta}, \sigma^2)^{-1} = \begin{pmatrix} \sigma^2(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1} & 0 \\ 0 & \frac{2\sigma^4}{n} \end{pmatrix} \right]$

Ridge Regression

$$\mathbf{y}^{n\times 1} = \mathbf{X}^{n\times p}\boldsymbol{\beta}^{p\times 1} + \mathbf{e}^{n\times 1}, \qquad \mathbf{e} \sim \mathrm{MVN}_n(\mathbf{0}, \sigma^2\mathbf{I}_n)$$

$$\text{OLS/MLE:} \quad \widehat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$$

- What if $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ is ill conditioned?
- Minimize penalized squared error loss:

$$\widehat{\beta}_{RR} = \arg\min\{(\mathbf{y} - \mathbf{X}\beta)^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\beta) + \lambda\beta^{\mathrm{T}}\beta\}$$

- Ridge: $\widehat{\beta} = (\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda\mathbf{I}_p)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}, \qquad \widehat{\sigma}^2 = (\mathbf{y} - \mathbf{X}\widehat{\beta})^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\widehat{\beta})/n$
- Mean-variance: $\mathbb{E}(\widehat{\beta}) = (\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda\mathbf{I}_p)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{X}\beta \neq \beta,$

$$\mathrm{var}(\widehat{\beta}) = \sigma^2(\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda\mathbf{I}_p)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda\mathbf{I}_p)^{-1}$$

- Typically, cross-validation is used to determine λ.

Ridge Regression

$$\mathbf{y}^{n\times 1} = \mathbf{X}^{n\times p}\boldsymbol{\beta}^{p\times 1} + \mathbf{e}^{n\times 1}, \qquad \mathbf{e} \sim \mathrm{MVN}_n(\mathbf{0}, \sigma^2\mathbf{I}_n)$$

$$\text{OLS/MLE:} \quad \widehat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$$

- What if $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ is ill conditioned?
- Minimize penalized squared error loss:

$$\widehat{\boldsymbol{\beta}}_{RR} = \arg\min\{(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) + \lambda\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\beta}\}$$

- Ridge: $\widehat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda\mathbf{I}_p)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}, \qquad \widehat{\sigma}^2 = (\mathbf{y}-\mathbf{X}\widehat{\boldsymbol{\beta}})^{\mathrm{T}}(\mathbf{y}-\mathbf{X}\widehat{\boldsymbol{\beta}})/n$
- Mean-variance: $\mathbb{E}(\widehat{\boldsymbol{\beta}}) = (\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda\mathbf{I}_p)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{\beta} \neq \boldsymbol{\beta}$,

$$\mathrm{var}(\widehat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda\mathbf{I}_p)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda\mathbf{I}_p)^{-1}$$

- Typically, cross-validation is used to determine λ.

Ridge Regression

$$\mathbf{y}^{n\times 1} = \mathbf{X}^{n\times p}\boldsymbol{\beta}^{p\times 1} + \mathbf{e}^{n\times 1}, \qquad \mathbf{e} \sim \mathrm{MVN}_n(\mathbf{0}, \sigma^2\mathbf{I}_n)$$

$$\text{OLS/MLE:} \quad \widehat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$$

- What if $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ is ill conditioned?
- Minimize penalized squared error loss:

$$\widehat{\boldsymbol{\beta}}_{RR} = \arg\min\{(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) + \lambda\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\beta}\}$$

- Ridge: $\widehat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda\mathbf{I}_p)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}, \qquad \widehat{\sigma}^2 = (\mathbf{y}-\mathbf{X}\widehat{\boldsymbol{\beta}})^{\mathrm{T}}(\mathbf{y}-\mathbf{X}\widehat{\boldsymbol{\beta}})/n$
- Mean-variance: $\mathbb{E}(\widehat{\boldsymbol{\beta}}) = (\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda\mathbf{I}_p)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{\beta} \neq \boldsymbol{\beta}$,

$$\mathrm{var}(\widehat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda\mathbf{I}_p)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda\mathbf{I}_p)^{-1}$$

- Typically, cross-validation is used to determine λ.

Ridge Regression

$$\mathbf{y}^{n\times 1} = \mathbf{X}^{n\times p}\boldsymbol{\beta}^{p\times 1} + \mathbf{e}^{n\times 1}, \qquad \mathbf{e} \sim \mathrm{MVN}_n(\mathbf{0}, \sigma^2\mathbf{I}_n)$$

$$\text{OLS/MLE:} \quad \widehat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$$

- What if $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ is ill conditioned?
- Minimize penalized squared error loss:

$$\widehat{\boldsymbol{\beta}}_{RR} = \arg\min\{(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \lambda\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\beta}\}$$

- Ridge: $\widehat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda\mathbf{I}_p)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}, \qquad \widehat{\sigma}^2 = (\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}})^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}})/n$
- Mean-variance: $\mathbb{E}(\widehat{\boldsymbol{\beta}}) = (\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda\mathbf{I}_p)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{\beta} \neq \boldsymbol{\beta}$,

$$\mathrm{var}(\widehat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda\mathbf{I}_p)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda\mathbf{I}_p)^{-1}$$

- Typically, cross-validation is used to determine λ.

Ridge Regression

$$\mathbf{y}^{n\times 1} = \mathbf{X}^{n\times p}\boldsymbol{\beta}^{p\times 1} + \mathbf{e}^{n\times 1}, \qquad \mathbf{e} \sim \mathrm{MVN}_n(\mathbf{0}, \sigma^2\mathbf{I}_n)$$

$$\text{OLS/MLE:} \quad \widehat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$$

- What if $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ is ill conditioned?
- Minimize penalized squared error loss:

$$\widehat{\boldsymbol{\beta}}_{RR} = \arg\min\{(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \lambda\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\beta}\}$$

- Ridge: $\widehat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda\mathbf{I}_p)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}, \qquad \widehat{\sigma}^2 = (\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}})^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}})/n$
- Mean-variance: $\mathbb{E}(\widehat{\boldsymbol{\beta}}) = (\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda\mathbf{I}_p)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{\beta} \neq \boldsymbol{\beta}$,

$$\mathrm{var}(\widehat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda\mathbf{I}_p)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda\mathbf{I}_p)^{-1}$$

- Typically, cross-validation is used to determine λ.

Least Absolute Shrinkage and Selection Operator (LASSO)

$$\mathbf{y}^{n\times 1} = \mathbf{X}^{n\times p}\boldsymbol{\beta}^{p\times 1} + \mathbf{e}^{n\times 1}, \qquad \mathbf{e} \sim \mathrm{MVN}_n(\mathbf{0}, \sigma^2\mathbf{I}_n)$$

$$\text{OLS/MLE:} \quad \widehat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$$

$$\text{Ridge:} \quad \widehat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$$

- What if $\boldsymbol{\beta}$ is sparse?
- Minimize penalized squared error loss:

$$\widehat{\boldsymbol{\beta}}_{LASSO} = \arg\min\{(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \lambda\,\|\boldsymbol{\beta}\|_1\}$$

- Can be solved via the coordinate-wise gradient descent algorithm.
- Typically cross-validation is used to determine λ.

Least Absolute Shrinkage and Selection Operator (LASSO)

$$\mathbf{y}^{n\times 1} = \mathbf{X}^{n\times p}\boldsymbol{\beta}^{p\times 1} + \mathbf{e}^{n\times 1}, \qquad \mathbf{e} \sim \mathrm{MVN}_n(\mathbf{0}, \sigma^2\mathbf{I}_n)$$

$$\text{OLS/MLE:} \quad \widehat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$$

$$\text{Ridge:} \quad \widehat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$$

- What if $\boldsymbol{\beta}$ is sparse?
- Minimize penalized squared error loss:

$$\widehat{\boldsymbol{\beta}}_{LASSO} = \arg\min\{(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \lambda\left\|\boldsymbol{\beta}\right\|_1\}$$

- Can be solved via the coordinate-wise gradient descent algorithm.
- Typically cross-validation is used to determine λ.

Least Absolute Shrinkage and Selection Operator (LASSO)

$$\mathbf{y}^{n\times 1} = \mathbf{X}^{n\times p}\boldsymbol{\beta}^{p\times 1} + \mathbf{e}^{n\times 1}, \qquad \mathbf{e} \sim \mathrm{MVN}_n(\mathbf{0}, \sigma^2\mathbf{I}_n)$$

$$\text{OLS/MLE:} \quad \widehat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$$

$$\text{Ridge:} \quad \widehat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$$

- What if $\boldsymbol{\beta}$ is sparse?
- Minimize penalized squared error loss:

$$\widehat{\boldsymbol{\beta}}_{LASSO} = \arg\min\{(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \lambda \left\|\boldsymbol{\beta}\right\|_1\}$$

- Can be solved via the coordinate-wise gradient descent algorithm.
- Typically cross-validation is used to determine λ.

Least Absolute Shrinkage and Selection Operator (LASSO)

$$\mathbf{y}^{n\times 1} = \mathbf{X}^{n\times p}\boldsymbol{\beta}^{p\times 1} + \mathbf{e}^{n\times 1}, \qquad \mathbf{e} \sim \mathrm{MVN}_n(\mathbf{0}, \sigma^2\mathbf{I}_n)$$
$$\text{OLS/MLE:} \quad \widehat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$$
$$\text{Ridge:} \quad \widehat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$$

- What if $\boldsymbol{\beta}$ is sparse?
- Minimize penalized squared error loss:

$$\widehat{\boldsymbol{\beta}}_{LASSO} = \arg\min\{(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \lambda \|\boldsymbol{\beta}\|_1\}$$

- Can be solved via the coordinate-wise gradient descent algorithm.
- Typically cross-validation is used to determine λ.

Elastic Net: $L_1 + L_2$ Penalization

$$\mathbf{y}^{n\times 1} = \mathbf{X}^{n\times p}\boldsymbol{\beta}^{p\times 1} + \mathbf{e}^{n\times 1}, \qquad \mathbf{e} \sim \text{MVN}_n(\mathbf{0}, \sigma^2\mathbf{I}_n)$$

$$\text{OLS/MLE:} \quad \widehat{\boldsymbol{\beta}}_{OLS} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$$

$$\text{Ridge:} \quad \widehat{\boldsymbol{\beta}}_{RR} = (\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$$

$$\text{LASSO:} \quad \widehat{\boldsymbol{\beta}}_{LASSO} = \arg\min\{(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \lambda \left\|\boldsymbol{\beta}\right\|_1\}$$

- How to get the best of both worlds?
- Minimize L_1 and L_2 penalized squared error loss:

$$\widehat{\boldsymbol{\beta}}_{ENET} = \arg\min[(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \lambda\{(1-\alpha)\left\|\boldsymbol{\beta}\right\|_2^2 + 2\alpha\left\|\boldsymbol{\beta}\right\|_1\}]$$

- Typically cross-validation is used to determine λ for fixed values of α.

Elastic Net: $L_1 + L_2$ Penalization

$$\mathbf{y}^{n\times 1} = \mathbf{X}^{n\times p}\boldsymbol{\beta}^{p\times 1} + \mathbf{e}^{n\times 1}, \qquad \mathbf{e} \sim \mathrm{MVN}_n(\mathbf{0}, \sigma^2\mathbf{I}_n)$$

$$\text{OLS/MLE:} \quad \widehat{\boldsymbol{\beta}}_{OLS} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$$

$$\text{Ridge:} \quad \widehat{\boldsymbol{\beta}}_{RR} = (\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$$

$$\text{LASSO:} \quad \widehat{\boldsymbol{\beta}}_{LASSO} = \arg\min\{(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \lambda \|\boldsymbol{\beta}\|_1\}$$

- How to get the best of both worlds?
- Minimize L_1 and L_2 penalized squared error loss:

$$\widehat{\boldsymbol{\beta}}_{ENET} = \arg\min[(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \lambda\{(1-\alpha)\|\boldsymbol{\beta}\|_2^2 + 2\alpha\|\boldsymbol{\beta}\|_1\}]$$

- Typically cross-validation is used to determine λ for fixed values of α.

Elastic Net: $L_1 + L_2$ Penalization

$$\mathbf{y}^{n\times 1} = \mathbf{X}^{n\times p}\boldsymbol{\beta}^{p\times 1} + \mathbf{e}^{n\times 1}, \qquad \mathbf{e} \sim \mathrm{MVN}_n(\mathbf{0}, \sigma^2\mathbf{I}_n)$$

$$\text{OLS/MLE:} \quad \widehat{\boldsymbol{\beta}}_{OLS} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$$

$$\text{Ridge:} \quad \widehat{\boldsymbol{\beta}}_{RR} = (\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$$

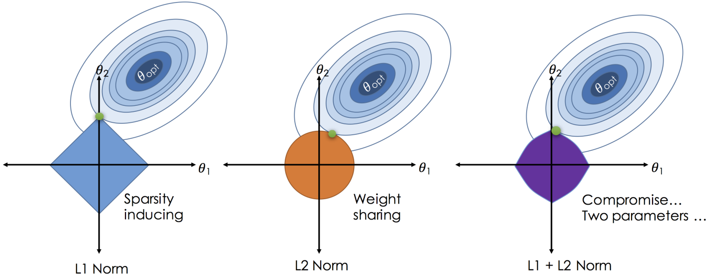
$$\text{LASSO:} \quad \widehat{\boldsymbol{\beta}}_{LASSO} = \arg\min\{(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \lambda\left\|\boldsymbol{\beta}\right\|_1\}$$

- How to get the best of both worlds?
- Minimize L_1 and L_2 penalized squared error loss:

$$\widehat{\boldsymbol{\beta}}_{ENET} = \arg\min[(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \lambda\{(1-\alpha)\left\|\boldsymbol{\beta}\right\|_2^2 + 2\alpha\left\|\boldsymbol{\beta}\right\|_1\}]$$

- Typically cross-validation is used to determine λ for fixed values of α.

Ridge vs LASSO vs Elastic Net



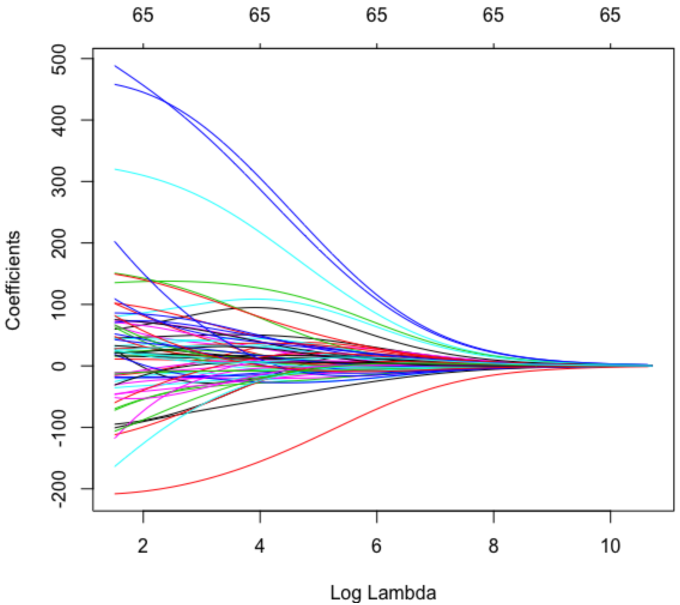
Contours of the error and constraint functions: the ellipses show the contours of the residual sum of squares (RSS), the regions around the origin show constraints corresponding to some given budget. The green dot is the smallest RSS that meets the budget.

Diabetes Data

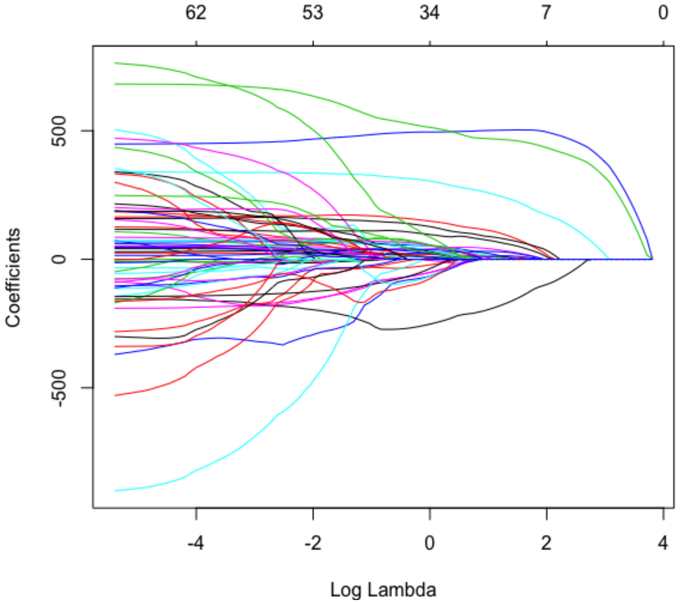
- Data set with $n = 442$ individuals and $p = 64$ covariates.
- The response and the associated covariate values for the first six individuals and the first seven covariates are listed below.

	y	age	sex	bmi	map	tc	ldl	hdl
1	151	0.038075906	0.05068012	0.06169621	0.021872355	-0.044223498	-0.03482076	-0.043400846
2	75	-0.001882017	-0.04464164	-0.05147406	-0.026327835	-0.008448724	-0.01916334	0.074411564
3	141	0.085298906	0.05068012	0.04445121	-0.005670611	-0.045599451	-0.03419447	-0.032355932
4	206	-0.089062939	-0.04464164	-0.01159501	-0.036656447	0.012190569	0.02499059	-0.036037570
5	135	0.005383060	-0.04464164	-0.03638469	0.021872355	0.003934852	0.01559614	0.008142084
6	97	-0.092695478	-0.04464164	-0.04069594	-0.019442093	-0.068990650	-0.07928784	0.041276824

Diabetes Data - Ridge



Diabetes Data- LASSO



$$\mathbf{y}^{n\times 1} = \mathbf{X}^{n\times p}\boldsymbol{\beta}^{p\times 1} + \mathbf{e}^{n\times 1}, \qquad \mathbf{e} \sim \mathrm{MVN}_n(\mathbf{0}, \sigma^2\mathbf{I}_n)$$

- Likelihood: $L(\boldsymbol{\beta}, \sigma^2) = \dfrac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{-\dfrac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right\}$

$$\propto \frac{1}{(\sigma^2)^{n/2}} \exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{\beta} - 2\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{y} + \mathbf{y}^{\mathrm{T}}\mathbf{y})\right\}$$

- Semi-conjugate priors: $p(\boldsymbol{\beta}, \sigma^2) = \mathrm{MVN}(\boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta) \cdot \text{Inv-Ga}(a_\sigma, b_\sigma)$

$$\propto \frac{1}{(\sigma^2)^{a_\sigma+1}} \exp\left(-\frac{b_\sigma}{\sigma^2}\right) \exp\left\{-\frac{1}{2}(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\Sigma}_\beta^{-1}\boldsymbol{\beta} - 2\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\Sigma}_\beta^{-1}\boldsymbol{\mu}_\beta + \boldsymbol{\mu}_\beta^{\mathrm{T}}\boldsymbol{\Sigma}_\beta^{-1}\boldsymbol{\mu}_\beta)\right\}$$

- Posterior: $p(\boldsymbol{\beta}, \sigma^2 \mid \mathbf{y}_{1:n}) \propto L(\boldsymbol{\beta}, \sigma^2) \times p(\boldsymbol{\beta}, \sigma^2)$
- Posterior full conditionals for block-Gibbs sampler:

Bayesian Linear Models

$$\mathbf{y}^{n\times 1} = \mathbf{X}^{n\times p}\boldsymbol{\beta}^{p\times 1} + \mathbf{e}^{n\times 1}, \qquad \mathbf{e} \sim \mathrm{MVN}_n(\mathbf{0}, \sigma^2\mathbf{I}_n)$$

- Likelihood: $L(\boldsymbol{\beta},\sigma^2) = \dfrac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{-\dfrac{1}{2\sigma^2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\right\}$

$$\propto \frac{1}{(\sigma^2)^{n/2}} \exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{\beta} - 2\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{y} + \mathbf{y}^{\mathrm{T}}\mathbf{y})\right\}$$

- Semi-conjugate priors: $p(\boldsymbol{\beta},\sigma^2) = \mathrm{MVN}(\boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta)\cdot \text{Inv-Ga}(a_\sigma, b_\sigma)$

$$\propto \exp\left[-\frac{1}{2}\{(\boldsymbol{\beta}-\boldsymbol{\mu}_\beta)^{\mathrm{T}}\boldsymbol{\Sigma}_\beta^{-1}(\boldsymbol{\beta}-\boldsymbol{\mu}_\beta)\}\right] \frac{1}{(\sigma^2)^{a_\sigma+1}} \exp\left(-\frac{b_\sigma}{\sigma^2}\right)$$

$$\propto \frac{1}{(\sigma^2)^{a_\sigma+1}} \exp\left(-\frac{b_\sigma}{\sigma^2}\right) \exp\left\{-\frac{1}{2}(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\Sigma}_\beta^{-1}\boldsymbol{\beta} - 2\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\Sigma}_\beta^{-1}\boldsymbol{\mu}_\beta + \boldsymbol{\mu}_\beta^{\mathrm{T}}\boldsymbol{\Sigma}_\beta^{-1}\boldsymbol{\mu}_\beta)\right\}$$

- Posterior: $p(\boldsymbol{\beta},\sigma^2 \mid \mathbf{y}_{1:n}) \propto L(\boldsymbol{\beta},\sigma^2)\times p(\boldsymbol{\beta},\sigma^2)$
- Posterior full conditionals for block-Gibbs sampler:

Bayesian Linear Models

$$\mathbf{y}^{n\times 1} = \mathbf{X}^{n\times p}\boldsymbol{\beta}^{p\times 1} + \mathbf{e}^{n\times 1}, \qquad \mathbf{e} \sim \mathrm{MVN}_n(\mathbf{0}, \sigma^2\mathbf{I}_n)$$

- Likelihood: $L(\boldsymbol{\beta}, \sigma^2) = \dfrac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{-\dfrac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right\}$

$$\propto \frac{1}{(\sigma^2)^{n/2}} \exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{\beta} - 2\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{y} + \mathbf{y}^{\mathrm{T}}\mathbf{y})\right\}$$

- Semi-conjugate priors: $p(\boldsymbol{\beta}, \sigma^2) = \mathrm{MVN}(\boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta) \cdot \text{Inv-Ga}(a_\sigma, b_\sigma)$

$$\propto \frac{1}{(\sigma^2)^{a_\sigma+1}} \exp\left(-\frac{b_\sigma}{\sigma^2}\right) \exp\left\{-\frac{1}{2}(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\Sigma}_\beta^{-1}\boldsymbol{\beta} - 2\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\Sigma}_\beta^{-1}\boldsymbol{\mu}_\beta + \boldsymbol{\mu}_\beta^{\mathrm{T}}\boldsymbol{\Sigma}_\beta^{-1}\boldsymbol{\mu}_\beta)\right\}$$

- Posterior: $p(\boldsymbol{\beta}, \sigma^2 \mid \mathbf{y}_{1:n}) \propto L(\boldsymbol{\beta}, \sigma^2) \times p(\boldsymbol{\beta}, \sigma^2)$
- Posterior full conditionals for block-Gibbs sampler:

Bayesian Linear Models

$$\mathbf{y}^{n\times 1} = \mathbf{X}^{n\times p}\boldsymbol{\beta}^{p\times 1} + \mathbf{e}^{n\times 1}, \qquad \mathbf{e} \sim \text{MVN}_n(\mathbf{0}, \sigma^2\mathbf{I}_n)$$

- Likelihood: $L(\boldsymbol{\beta}, \sigma^2) = \dfrac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{-\dfrac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right\}$

$$\propto \frac{1}{(\sigma^2)^{n/2}} \exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{\beta} - 2\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{y} + \mathbf{y}^{\mathrm{T}}\mathbf{y})\right\}$$

- Semi-conjugate priors: $p(\boldsymbol{\beta}, \sigma^2) = \text{MVN}(\boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta) \cdot \text{Inv-Ga}(a_\sigma, b_\sigma)$

$$\propto \frac{1}{(\sigma^2)^{a_\sigma+1}} \exp\left(-\frac{b_\sigma}{\sigma^2}\right) \exp\left\{-\frac{1}{2}(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\Sigma}_\beta^{-1}\boldsymbol{\beta} - 2\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\Sigma}_\beta^{-1}\boldsymbol{\mu}_\beta + \boldsymbol{\mu}_\beta^{\mathrm{T}}\boldsymbol{\Sigma}_\beta^{-1}\boldsymbol{\mu}_\beta)\right\}$$

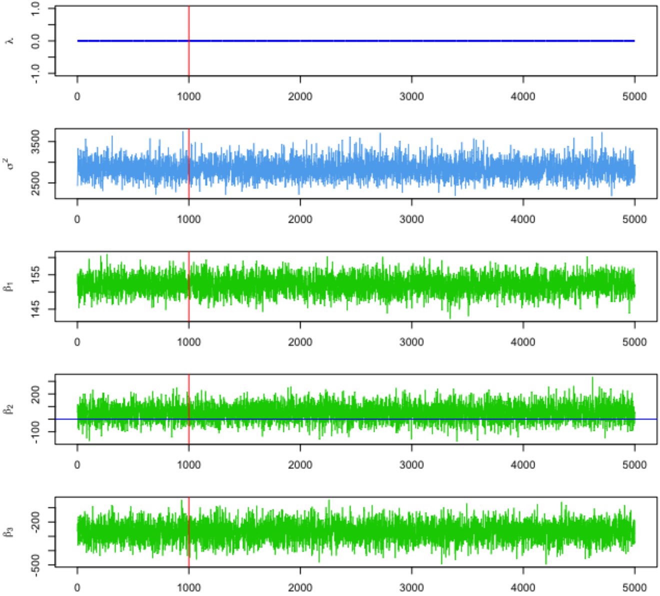
- Posterior: $p(\boldsymbol{\beta}, \sigma^2 \mid \mathbf{y}_{1:n}) \propto L(\boldsymbol{\beta}, \sigma^2) \times p(\boldsymbol{\beta}, \sigma^2)$
- Posterior full conditionals for block-Gibbs sampler:
  - $p(\boldsymbol{\beta} \mid -) \propto \exp\left[-\dfrac{1}{2}\left\{\boldsymbol{\beta}^{\mathrm{T}}(\boldsymbol{\Sigma}_\beta^{-1} + \sigma^{-2}\mathbf{X}^{\mathrm{T}}\mathbf{X})\boldsymbol{\beta} - 2\boldsymbol{\beta}^{\mathrm{T}}\left(\boldsymbol{\Sigma}_\beta^{-1}\boldsymbol{\mu}_\beta + \sigma^{-2}\mathbf{X}^{\mathrm{T}}\mathbf{y}\right)\right\}\right]$

    $\equiv \text{MVN}(\boldsymbol{\mu}_{\beta,n}, \boldsymbol{\Sigma}_{\beta,n})$,

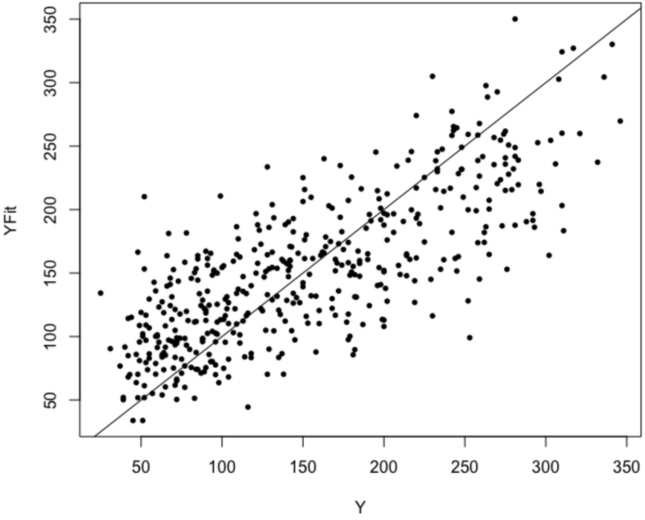
    $\boldsymbol{\Sigma}_{\beta,n} = (\boldsymbol{\Sigma}_\beta^{-1} + \sigma^{-2}\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}, \quad \boldsymbol{\mu}_{\beta,n} = \boldsymbol{\Sigma}_{\beta,n}\left(\boldsymbol{\Sigma}_\beta^{-1}\boldsymbol{\mu}_\beta + \sigma^{-2}\mathbf{X}^{\mathrm{T}}\mathbf{y}\right)$
  - $p(\sigma^2 \mid -) \propto \dfrac{1}{(\sigma^2)^{a_\sigma+\frac{n}{2}+1}} \exp\left[-\dfrac{1}{\sigma^2}\left\{b_\sigma + \dfrac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right\}\right]$

    $\equiv \text{Inv-Ga}\left\{a_\sigma + \dfrac{n}{2}, b_\sigma + \dfrac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right\}$

Diabetes Data - Bayesian Linear Models



Diabetes Data - Bayesian Linear Models



Bayesian Ridge

$$\mathbf{y}^{n\times 1} = \mathbf{X}^{n\times p}\boldsymbol{\beta}^{p\times 1} + \mathbf{e}^{n\times 1}, \qquad \mathbf{e} \sim \mathrm{MVN}_n(\mathbf{0}, \sigma^2\mathbf{I}_n)$$

- Ridge Regression:
  - $\widehat{\boldsymbol{\beta}}_{RR} = \arg\min\{(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) + \lambda\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\beta}\} = (\mathbf{X}^{\mathrm{T}}\mathbf{X}+\lambda\mathbf{I}_p)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$
- An alternative view:

Bayesian Ridge

$$\mathbf{y}^{n\times 1} = \mathbf{X}^{n\times p}\boldsymbol{\beta}^{p\times 1} + \mathbf{e}^{n\times 1}, \qquad \mathbf{e} \sim \mathrm{MVN}_n(\mathbf{0}, \sigma^2\mathbf{I}_n)$$

- Ridge Regression:
  - $\widehat{\boldsymbol{\beta}}_{RR} = \arg\min\{(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) + \lambda\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\beta}\} = (\mathbf{X}^{\mathrm{T}}\mathbf{X}+\lambda\mathbf{I}_p)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$
- An alternative view:
  - $\widehat{\boldsymbol{\beta}}_{RR} = \arg\min\left\{\frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) + \frac{\lambda}{2\sigma^2}\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\beta}\right\}$

$$= \arg\max\left\{-\frac{1}{2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\sigma^2\mathbf{I}_n)^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) - \frac{1}{2}\boldsymbol{\beta}^{\mathrm{T}}\left(\frac{\sigma^2}{\lambda}\mathbf{I}_p\right)^{-1}\boldsymbol{\beta}\right\}$$

$$= \arg\max\left\{\mathcal{L}(\boldsymbol{\beta}\mid\sigma^2) + \log p(\boldsymbol{\beta}\mid\lambda,\sigma^2)\right\}$$

where $p(\boldsymbol{\beta}\mid\lambda,\sigma^2) = \mathrm{MVN}\left(\mathbf{0}, \frac{\sigma^2}{\lambda}\mathbf{I}_p\right)$

  - $\widehat{\boldsymbol{\beta}}_{RR} = \widehat{\boldsymbol{\beta}}_{MAP} = (\mathbf{X}^{\mathrm{T}}\mathbf{X}+\lambda\mathbf{I}_p)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y} \quad$ with $\quad p(\boldsymbol{\beta}\mid\lambda,\sigma^2) = \mathrm{MVN}\left(\mathbf{0}, \frac{\sigma^2}{\lambda}\mathbf{I}_p\right)$.
  - Full conditionals:
  - λ can also be assigned $\mathrm{Ga}(a_\lambda, b_\lambda)$ hyper-prior and sampled from its full conditional $p(\lambda\mid -) = \mathrm{Ga}\left\{a_\lambda + p/2, b_\lambda + \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\beta}/(2\sigma^2)\right\}$.

Bayesian Ridge

$$\mathbf{y}^{n\times 1} = \mathbf{X}^{n\times p}\boldsymbol{\beta}^{p\times 1} + \mathbf{e}^{n\times 1}, \qquad \mathbf{e} \sim \mathrm{MVN}_n(\mathbf{0}, \sigma^2\mathbf{I}_n)$$

- Ridge Regression:
  - $\widehat{\boldsymbol{\beta}}_{RR} = \arg\min\{(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) + \lambda\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\beta}\} = (\mathbf{X}^{\mathrm{T}}\mathbf{X}+\lambda\mathbf{I}_p)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$
- An alternative view:
  - $\widehat{\boldsymbol{\beta}}_{RR} = \arg\min\left\{\frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) + \frac{\lambda}{2\sigma^2}\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\beta}\right\}$
  - $\widehat{\boldsymbol{\beta}}_{RR} = \widehat{\boldsymbol{\beta}}_{MAP} = (\mathbf{X}^{\mathrm{T}}\mathbf{X}+\lambda\mathbf{I}_p)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$ with $p(\boldsymbol{\beta}\mid\lambda,\sigma^2) = \mathrm{MVN}\left(\mathbf{0}, \frac{\sigma^2}{\lambda}\mathbf{I}_p\right)$.
  - Full conditionals:
- λ can also be assigned $\mathrm{Ga}(a_\lambda, b_\lambda)$ hyper-prior and sampled from its full conditional $p(\lambda\mid -) = \mathrm{Ga}\left\{a_\lambda + p/2, b_\lambda + \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\beta}/(2\sigma^2)\right\}$.

Bayesian Ridge

$$\mathbf{y}^{n\times 1} = \mathbf{X}^{n\times p}\boldsymbol{\beta}^{p\times 1} + \mathbf{e}^{n\times 1}, \qquad \mathbf{e} \sim \text{MVN}_n(\mathbf{0}, \sigma^2\mathbf{I}_n)$$

- Ridge Regression:
  - $\widehat{\boldsymbol{\beta}}_{RR} = \arg\min\{(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) + \lambda\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\beta}\} = (\mathbf{X}^{\mathrm{T}}\mathbf{X}+\lambda\mathbf{I}_p)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$
- An alternative view:
  - $\widehat{\boldsymbol{\beta}}_{RR} = \arg\min\left\{\frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) + \frac{\lambda}{2\sigma^2}\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\beta}\right\}$
  - $\widehat{\boldsymbol{\beta}}_{RR} = \widehat{\boldsymbol{\beta}}_{MAP} = (\mathbf{X}^{\mathrm{T}}\mathbf{X}+\lambda\mathbf{I}_p)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$ with $p(\boldsymbol{\beta} \mid \lambda, \sigma^2) = \text{MVN}\left(\mathbf{0}, \frac{\sigma^2}{\lambda}\mathbf{I}_p\right)$.
  - Full conditionals:
    - $p(\boldsymbol{\beta} \mid -) \equiv \text{MVN}(\boldsymbol{\mu}_{\beta,n}, \boldsymbol{\Sigma}_{\beta,n})$,

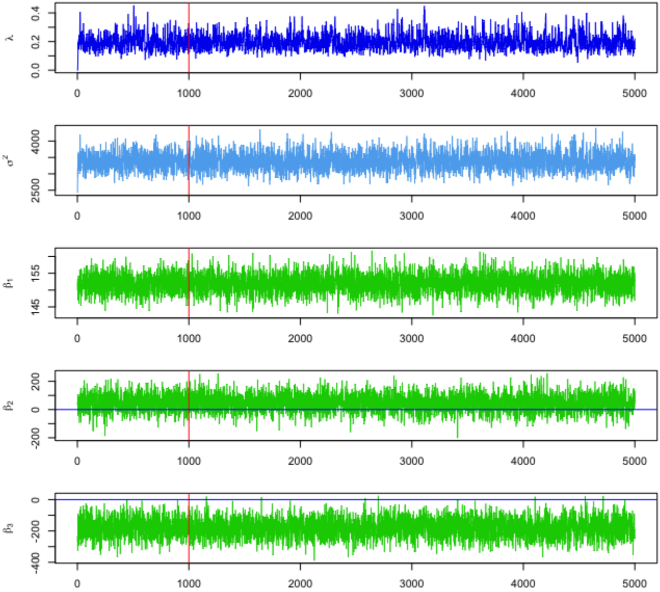
      $\boldsymbol{\Sigma}_{\beta,n} = \sigma^2(\mathbf{X}^{\mathrm{T}}\mathbf{X}+\lambda\mathbf{I}_p)^{-1}, \quad \boldsymbol{\mu}_{\beta,n} = \boldsymbol{\Sigma}_{\beta,n}(\sigma^{-2}\mathbf{X}^{\mathrm{T}}\mathbf{y}) = (\mathbf{X}^{\mathrm{T}}\mathbf{X}+\lambda\mathbf{I}_p)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$
    - $p(\sigma^2 \mid -) \equiv \text{Inv-Ga}\left\{a_\sigma + \frac{n}{2}, b_\sigma + \frac{1}{2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) + \frac{\lambda}{2}\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\beta}\right\}$
  - λ can also be assigned $\text{Ga}(a_\lambda, b_\lambda)$ hyper-prior and sampled from its full conditional $p(\lambda \mid -) = \text{Ga}\left\{a_\lambda + p/2, b_\lambda + \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\beta}/(2\sigma^2)\right\}$.

$$\mathbf{y}^{n\times 1} = \mathbf{X}^{n\times p}\boldsymbol{\beta}^{p\times 1} + \mathbf{e}^{n\times 1}, \qquad \mathbf{e} \sim \mathrm{MVN}_n(\mathbf{0}, \sigma^2\mathbf{I}_n)$$

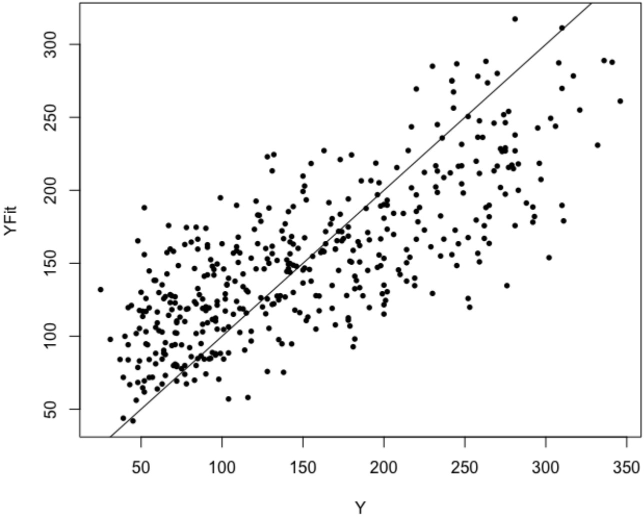
- Ridge Regression:
  - $\widehat{\boldsymbol{\beta}}_{RR} = \arg\min\{(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) + \lambda\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\beta}\} = (\mathbf{X}^{\mathrm{T}}\mathbf{X}+\lambda\mathbf{I}_p)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$
- An alternative view:
  - $\widehat{\boldsymbol{\beta}}_{RR} = \arg\min\left\{\frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) + \frac{\lambda}{2\sigma^2}\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\beta}\right\}$
  - $\widehat{\boldsymbol{\beta}}_{RR} = \widehat{\boldsymbol{\beta}}_{MAP} = (\mathbf{X}^{\mathrm{T}}\mathbf{X}+\lambda\mathbf{I}_p)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$ with $p(\boldsymbol{\beta}\mid\lambda,\sigma^2) = \mathrm{MVN}\left(\mathbf{0}, \frac{\sigma^2}{\lambda}\mathbf{I}_p\right)$.
  - Full conditionals:
    - $p(\boldsymbol{\beta}\mid -) \equiv \mathrm{MVN}(\boldsymbol{\mu}_{\beta,n}, \boldsymbol{\Sigma}_{\beta,n})$,

      $\boldsymbol{\Sigma}_{\beta,n} = \sigma^2(\mathbf{X}^{\mathrm{T}}\mathbf{X}+\lambda\mathbf{I}_p)^{-1}, \quad \boldsymbol{\mu}_{\beta,n} = \boldsymbol{\Sigma}_{\beta,n}(\sigma^{-2}\mathbf{X}^{\mathrm{T}}\mathbf{y}) = (\mathbf{X}^{\mathrm{T}}\mathbf{X}+\lambda\mathbf{I}_p)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$
    - $p(\sigma^2\mid -) \equiv \text{Inv-Ga}\left\{a_\sigma + \frac{n}{2}, b_\sigma + \frac{1}{2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) + \frac{\lambda}{2}\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\beta}\right\}$
  - λ can also be assigned $\mathrm{Ga}(a_\lambda, b_\lambda)$ hyper-prior and sampled from its full conditional $p(\lambda\mid -) = \mathrm{Ga}\left\{a_\lambda + p/2, b_\lambda + \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\beta}/(2\sigma^2)\right\}$.

Diabetes Data - Bayesian Ridge



Diabetes Data - Bayesian Ridge



Bayesian LASSO

$$\mathbf{y}^{n\times 1} = \mathbf{X}^{n\times p}\boldsymbol{\beta}^{p\times 1} + \mathbf{e}^{n\times 1}, \qquad \mathbf{e} \sim \mathrm{MVN}_n(\mathbf{0}, \sigma^2\mathbf{I}_n)$$

- LASSO Regression:
  - $\widehat{\boldsymbol{\beta}}_{LASSO} = \arg\min\{(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) + \lambda\,||\boldsymbol{\beta}||_1\}$
- An alternative view:

Bayesian LASSO

$$\mathbf{y}^{n\times 1} = \mathbf{X}^{n\times p}\boldsymbol{\beta}^{p\times 1} + \mathbf{e}^{n\times 1}, \qquad \mathbf{e} \sim \mathrm{MVN}_n(\mathbf{0}, \sigma^2\mathbf{I}_n)$$

- LASSO Regression:
  - $\widehat{\boldsymbol{\beta}}_{LASSO} = \arg\min\{(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) + \lambda\,||\boldsymbol{\beta}||_1\}$
- An alternative view:
  - $\widehat{\boldsymbol{\beta}}_{LASSO} = \arg\min\left\{\frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) + \frac{\lambda}{2\sigma^2}\sum_{j=1}^{p}|\boldsymbol{\beta}_j|\right\}$

    $= \arg\max\left\{-\frac{1}{2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\sigma^2\mathbf{I}_n)^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) - \frac{\lambda}{2\sigma^2}\sum_{j=1}^{p}|\boldsymbol{\beta}_j|\right\}$

    $= \arg\max\left\{\mathcal{L}(\boldsymbol{\beta} \mid \sigma^2) + \log p(\boldsymbol{\beta} \mid \lambda, \sigma^2)\right\}$

    where $p(\boldsymbol{\beta} \mid \lambda, \sigma^2) = \prod_{j=1}^{p} \text{Laplace}\left(\beta_j \mid 0, \frac{2\sigma^2}{\lambda}\right)$
  - $\widehat{\boldsymbol{\beta}}_{LASSO} = \widehat{\boldsymbol{\beta}}_{MAP}$ with $p(\boldsymbol{\beta} \mid \lambda, \sigma^2) = \prod_{j=1}^{p} \text{Laplace}\left(\beta_j \mid 0, \frac{2\sigma^2}{\lambda}\right)$.

Bayesian LASSO

$$\mathbf{y}^{n\times 1} = \mathbf{X}^{n\times p}\boldsymbol{\beta}^{p\times 1} + \mathbf{e}^{n\times 1}, \qquad \mathbf{e} \sim \mathrm{MVN}_n(\mathbf{0}, \sigma^2\mathbf{I}_n)$$

- LASSO Regression:
  - $\widehat{\boldsymbol{\beta}}_{LASSO} = \arg\min\{(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) + \lambda\,||\boldsymbol{\beta}||_1\}$
- An alternative view:
  - $\widehat{\boldsymbol{\beta}}_{LASSO} = \arg\min\left\{\frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) + \frac{\lambda}{2\sigma^2}\sum_{j=1}^{p}|\boldsymbol{\beta}_j|\right\}$
  - $\widehat{\boldsymbol{\beta}}_{LASSO} = \widehat{\boldsymbol{\beta}}_{MAP}$ with $p(\boldsymbol{\beta} \mid \lambda, \sigma^2) = \prod_{j=1}^{p} \text{Laplace}\left(\beta_j \mid 0, \frac{2\sigma^2}{\lambda}\right)$.

$$\mathbf{y}^{n\times 1} = \mathbf{X}^{n\times p}\boldsymbol{\beta}^{p\times 1} + \mathbf{e}^{n\times 1}, \qquad \mathbf{e} \sim \text{MVN}_n(\mathbf{0}, \sigma^2\mathbf{I}_n)$$

- LASSO Regression:
    - $\widehat{\boldsymbol{\beta}}_{LASSO} = \arg\min\{(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) + \lambda\,||\boldsymbol{\beta}||_1\}$
- Bayesian LASSO is motivated by the conditional Laplace prior

$$\begin{aligned}
p(\boldsymbol{\beta} \mid \lambda, \sigma^2) &= \prod_{j=1}^{p} \text{Laplace}\left(\beta_j \mid 0, \tfrac{\sigma}{\lambda}\right) = \prod_{j=1}^{p} \tfrac{\lambda}{2\sqrt{\sigma^2}} \exp\left(-\lambda\,|\beta_j|\,/\sqrt{\sigma^2}\right) \\
&= \prod_{j=1}^{p} \int_0^\infty \frac{1}{\sqrt{2\pi\tau_j^2}} \exp\left(-\frac{\beta_j^2}{2\tau_j^2}\right) \frac{\lambda^2}{2\sigma^2} \exp\left(-\frac{\lambda^2\tau_j^2}{2\sigma^2}\right) d\tau_j^2 \\
&= \prod_{j=1}^{p} \int_0^\infty \text{Normal}(\beta_j \mid 0, \tau_j^2)\ \text{Exp}\left(\tau_j^2 \mid \tfrac{\lambda^2}{2\sigma^2}\right) d\tau_j^2
\end{aligned}$$

Bayesian LASSO

$$\mathbf{y}^{n\times 1} = \mathbf{X}^{n\times p}\boldsymbol{\beta}^{p\times 1} + \mathbf{e}^{n\times 1}, \qquad \mathbf{e} \sim \text{MVN}_n(\mathbf{0}, \sigma^2\mathbf{I}_n)$$

- LASSO Regression:
  - $\widehat{\boldsymbol{\beta}}_{LASSO} = \arg\min\{(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) + \lambda\,||\boldsymbol{\beta}||_1\}$
- Bayesian LASSO is motivated by the conditional Laplace prior

$$p(\boldsymbol{\beta} \mid \lambda, \sigma^2) = \prod_{j=1}^{p} \frac{\lambda}{2\sqrt{\sigma^2}} \exp\left(-\frac{\lambda|\beta_j|}{\sqrt{\sigma^2}}\right) = \prod_{j=1}^{p} \int_0^\infty \text{Normal}(\beta_j \mid 0, \tau_j^2)\ \text{Exp}\left(\tau_j^2 \mid \frac{\lambda^2}{2\sigma^2}\right) d\tau_j^2$$

- Bayesian LASSO as a global-local prior

$$p(\boldsymbol{\beta} \mid \boldsymbol{\tau}^2) = \underbrace{\prod_{j=1}^{p} \text{Normal}(\beta_j \mid 0, \tau_j^2),}_{\text{Local components}} \qquad p(\boldsymbol{\tau}^2 \mid \lambda, \sigma^2) = \underbrace{\prod_{j=1}^{p} \text{Exp}\left(\tau_j^2 \mid \frac{\lambda^2}{2\sigma^2}\right)}_{\text{Global components}}$$

- Posterior full conditionals for block Gibbs sampler

Bayesian LASSO

$$\mathbf{y}^{n\times 1} = \mathbf{X}^{n\times p}\boldsymbol{\beta}^{p\times 1} + \mathbf{e}^{n\times 1}, \qquad \mathbf{e} \sim \text{MVN}_n(\mathbf{0}, \sigma^2\mathbf{I}_n)$$

- LASSO Regression:
  - $\widehat{\boldsymbol{\beta}}_{LASSO} = \arg\min\{(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) + \lambda\,||\boldsymbol{\beta}||_1\}$
- Bayesian LASSO is motivated by the conditional Laplace prior

$$p(\boldsymbol{\beta} \mid \lambda, \sigma^2) = \prod_{j=1}^{p} \frac{\lambda}{2\sqrt{\sigma^2}} \exp\left(-\frac{\lambda|\beta_j|}{\sqrt{\sigma^2}}\right) = \prod_{j=1}^{p} \int_0^\infty \text{Normal}(\beta_j \mid 0, \tau_j^2)\ \text{Exp}\left(\tau_j^2 \mid \frac{\lambda^2}{2\sigma^2}\right) d\tau_j^2$$

- Posterior full conditionals for block-Gibbs sampler:
  - $p(\boldsymbol{\beta} \mid -) = \text{MVN}(\boldsymbol{\mu}_{\beta,n}, \boldsymbol{\Sigma}_{\beta,n}), \qquad \mathbf{D}_\tau = \text{diag}(\tau_1^2, \ldots, \tau_p^2),$

    $\boldsymbol{\Sigma}_{\beta,n} = \sigma^2(\mathbf{X}^{\mathrm{T}}\mathbf{X} + \mathbf{D}_\tau^{-1})^{-1}, \quad \boldsymbol{\mu}_{\beta,n} = \boldsymbol{\Sigma}_{\beta,n}(\sigma^{-2}\mathbf{X}^{\mathrm{T}}\mathbf{y}) = (\mathbf{X}^{\mathrm{T}}\mathbf{X} + \mathbf{D}_\tau^{-1})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$
  - $p(\sigma^2 \mid -) = \text{Inv-Ga}\left\{a_\sigma + \frac{n+p}{2}, b_\sigma + \frac{(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})}{2} + \frac{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{D}_\tau^{-1}\boldsymbol{\beta}}{2}\right\}$
  - $p(\tau_j^2 \mid -) \propto \frac{1}{\tau_j} \exp\left(-\frac{\beta_j^2}{2\tau_j^2}\right) \exp\left(-\frac{\lambda^2\tau_j^2}{2\sigma^2}\right), \qquad \tau_j^2 \to w_j = \tau_j^{-2}$

$$p(w_j \mid -) \propto w_j^{-3/2} \exp\left(-\frac{\beta_j^2 w_j}{2}\right) \exp\left(-\frac{\lambda^2}{2\sigma^2 w_j}\right) \equiv \text{Inv-Gs}(\mu', \lambda'), \quad \mu' = \frac{\lambda}{\sigma|\beta_j|}, \ \lambda' = \frac{\lambda^2}{\sigma^2}.$$

  - λ^2 can also be assigned $\text{Ga}(a_\lambda, b_\lambda)$ hyper-prior and sampled from its full conditional

    $p(\lambda^2 \mid -) \propto (\lambda^2)^{a_\lambda - 1}\exp(-\lambda^2 b_\lambda) \prod_{j=1}^{p}\left\{\frac{\lambda^2}{2\sigma^2}\exp\left(-\frac{\lambda^2\tau_j^2}{2\sigma^2}\right)\right\}$

    $\equiv \text{Ga}\left\{a_\lambda + p, b_\lambda + \sum_{j=1}^{p}\tau_j^2/(2\sigma^2)\right\}.$

Bayesian LASSO

$$\mathbf{y}^{n\times 1} = \mathbf{X}^{n\times p}\boldsymbol{\beta}^{p\times 1} + \mathbf{e}^{n\times 1}, \qquad \mathbf{e} \sim \text{MVN}_n(\mathbf{0}, \sigma^2\mathbf{I}_n)$$

- LASSO Regression:
  - $\widehat{\boldsymbol{\beta}}_{LASSO} = \arg\min\{(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) + \lambda\,||\boldsymbol{\beta}||_1\}$
- Bayesian LASSO is motivated by the conditional Laplace prior

$$p(\boldsymbol{\beta} \mid \lambda, \sigma^2) = \prod_{j=1}^{p} \frac{\lambda}{2\sqrt{\sigma^2}} \exp\left(-\frac{\lambda|\beta_j|}{\sqrt{\sigma^2}}\right) = \prod_{j=1}^{p} \int_0^\infty \text{Normal}(\beta_j \mid 0, \tau_j^2)\ \text{Exp}\left(\tau_j^2 \mid \frac{\lambda^2}{2\sigma^2}\right) d\tau_j^2$$

- Posterior full conditionals for block-Gibbs sampler:
  - $p(\boldsymbol{\beta} \mid -) = \text{MVN}(\boldsymbol{\mu}_{\beta,n}, \boldsymbol{\Sigma}_{\beta,n}), \qquad \mathbf{D}_\tau = \text{diag}(\tau_1^2, \ldots, \tau_p^2),$

    $\boldsymbol{\Sigma}_{\beta,n} = \sigma^2(\mathbf{X}^{\mathrm{T}}\mathbf{X} + \mathbf{D}_\tau^{-1})^{-1}, \quad \boldsymbol{\mu}_{\beta,n} = \boldsymbol{\Sigma}_{\beta,n}(\sigma^{-2}\mathbf{X}^{\mathrm{T}}\mathbf{y}) = (\mathbf{X}^{\mathrm{T}}\mathbf{X} + \mathbf{D}_\tau^{-1})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$
  - $p(\sigma^2 \mid -) = \text{Inv-Ga}\left\{a_\sigma + \frac{n+p}{2}, b_\sigma + \frac{(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})}{2} + \frac{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{D}_\tau^{-1}\boldsymbol{\beta}}{2}\right\}$
  - $p(\tau_j^2 \mid -) \propto \frac{1}{\tau_j} \exp\left(-\frac{\beta_j^2}{2\tau_j^2}\right) \exp\left(-\frac{\lambda^2\tau_j^2}{2\sigma^2}\right), \quad \tau_j^2 \rightarrow w_j = \tau_j^{-2}$

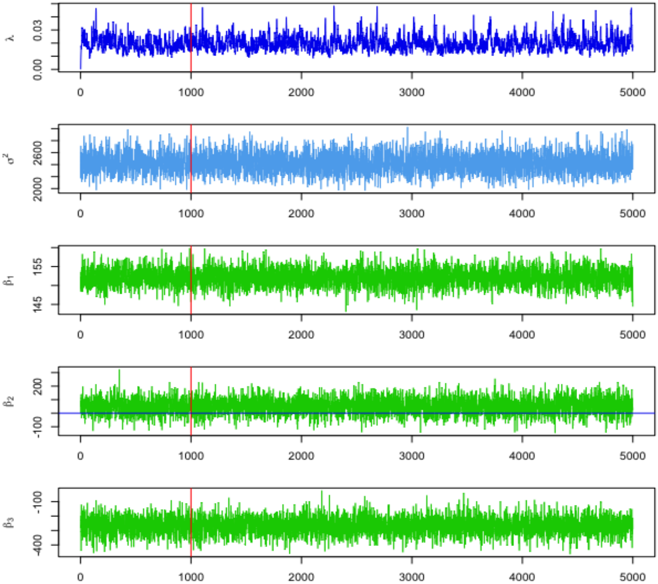
$$p(w_j \mid -) \propto w_j^{-3/2} \exp\left(-\frac{\beta_j^2 w_j}{2}\right) \exp\left(-\frac{\lambda^2}{2\sigma^2 w_j}\right) \equiv \text{Inv-Gs}(\mu', \lambda'), \ \ \mu' = \frac{\lambda}{\sigma|\beta_j|}, \ \lambda' = \frac{\lambda^2}{\sigma^2}.$$

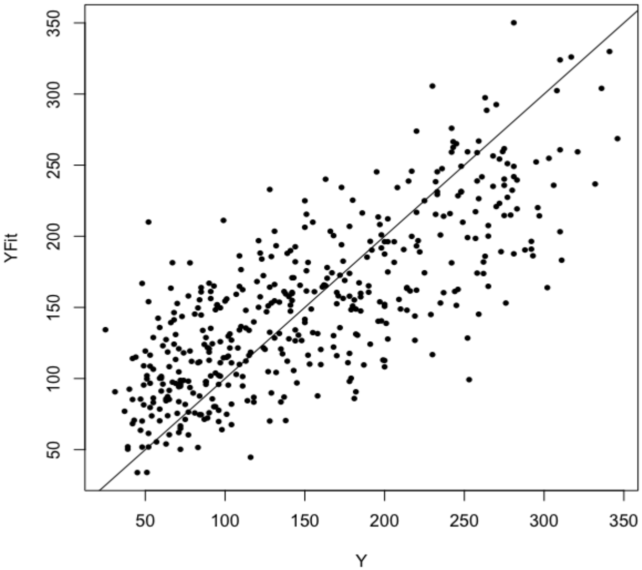
  - λ^2 can also be assigned $\text{Ga}(a_\lambda, b_\lambda)$ hyper-prior and sampled from its full conditional

    $p(\lambda^2 \mid -) \propto (\lambda^2)^{a_\lambda - 1}\exp(-\lambda^2 b_\lambda)\prod_{j=1}^{p}\left\{\frac{\lambda^2}{2\sigma^2}\exp\left(-\frac{\lambda^2\tau_j^2}{2\sigma^2}\right)\right\}$

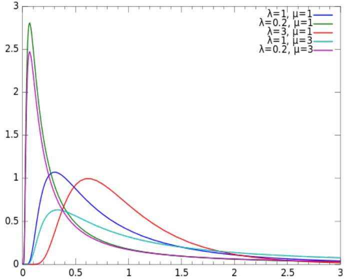
    $\equiv \text{Ga}\left\{a_\lambda + p, b_\lambda + \sum_{j=1}^{p}\tau_j^2/(2\sigma^2)\right\}.$

Diabetes Data - Bayesian LASSO




YFit
50
100
150
200
250
300
350
Y
50
100
150
200
250
300
350

- **Inverse Gaussian Distribution:**
  $$y \sim f(y \mid \mu, \lambda) = \sqrt{\frac{\lambda}{2\pi y^3}} \exp\left(-\frac{\lambda(y-\mu)^2}{2\mu^2 y}\right)$$
  $$y, \mu, \lambda \in \mathbb{R}^+$$
- $\mathbb{E}(y) = \mu$
- $\mathrm{var}(y) = \frac{\mu^3}{\lambda}$



Generalized / Weighted Least Squares

$$\mathbf{y}^{n\times 1} = \mathbf{X}^{n\times p}\boldsymbol{\beta}^{p\times 1} + \mathbf{e}^{n\times 1}, \quad \mathbb{E}(\mathbf{e}) = \mathbf{0},\ \text{cov}(\mathbf{e}) = \sigma^2\mathbf{V}, \quad \mathbf{V}\text{ known.}$$

- Minimize squared error loss:
  $$\widehat{\boldsymbol{\beta}} = \arg\min(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = (\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{y}$$
- $\mathbb{E}(\widehat{\boldsymbol{\beta}}) = (\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta}$
- $\text{var}(\widehat{\boldsymbol{\beta}}) = (\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\sigma^2\mathbf{V}\mathbf{V}^{-1}\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1} = \sigma^2(\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1}$
  - Hat matrix: $\mathbf{H}_V = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}$
  - Idempotent: $\mathbf{H}_V^2 = \mathbf{H}_V,\ (\mathbf{I}_n - \mathbf{H}_V)^2 = (\mathbf{I}_n - \mathbf{H}_V)$
  - $\text{trace}(\mathbf{H}_V) = \text{trace}(\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}) = \text{trace}(\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1}) = p$
  - $\text{trace}(\mathbf{I}_n - \mathbf{H}_V) = \text{trace}(\mathbf{I}_n) - \text{trace}(\mathbf{H}_V) = (n - p)$
  - $\mathbf{H}_V\mathbf{X} = \mathbf{X}$ and $(\mathbf{I}_n - \mathbf{H}_V)\mathbf{X} = (\mathbf{I}_n - \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}})\mathbf{X} = \mathbf{0}$
- Fitted values: $\widehat{\mathbf{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{y} = \mathbf{H}_V\mathbf{y}$
- Mean-Variance: $\mathbb{E}(\widehat{\mathbf{y}}) = \mathbf{X}\mathbb{E}(\widehat{\boldsymbol{\beta}}) = \mathbf{X}\boldsymbol{\beta}, \quad \text{var}(\widehat{\mathbf{y}}) = \sigma^2\mathbf{H}_V$
- Residuals: $\widehat{\mathbf{e}} = (\mathbf{y} - \widehat{\mathbf{y}}) = \mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}} = (\mathbf{I}_n - \mathbf{H}_V)\mathbf{y}$
- Mean-Variance: $\mathbb{E}(\widehat{\mathbf{e}}) = (\mathbf{I}_n - \mathbf{H}_V)\mathbf{X}\boldsymbol{\beta} = \mathbf{0}, \quad \text{var}(\widehat{\mathbf{e}}) = \mathbb{E}(\widehat{\mathbf{e}}\widehat{\mathbf{e}}^{\mathrm{T}}) = \sigma^2(\mathbf{I}_n - \mathbf{H}_V)$
- $\text{cov}(\widehat{\mathbf{y}}, \widehat{\mathbf{e}}) = \text{cov}\{\mathbf{H}_V\mathbf{y}, (\mathbf{I}_n - \mathbf{H}_V)\mathbf{y}) = \sigma^2\mathbf{H}_V(\mathbf{I}_n - \mathbf{H}_V) = \sigma^2(\mathbf{H}_V - \mathbf{H}_V^2) = \mathbf{0}$

Generalized / Weighted Least Squares

$$\mathbf{y}^{n\times 1} = \mathbf{X}^{n\times p}\boldsymbol{\beta}^{p\times 1} + \mathbf{e}^{n\times 1}, \quad \mathbb{E}(\mathbf{e}) = \mathbf{0},\ \mathrm{cov}(\mathbf{e}) = \sigma^2\mathbf{V}, \quad \mathbf{V}\ \text{known.}$$

- Minimize squared error loss:

$$\widehat{\boldsymbol{\beta}} = \arg\min(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = (\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{y}$$

- $\mathbb{E}(\widehat{\boldsymbol{\beta}}) = (\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta}$
- $\mathrm{var}(\widehat{\boldsymbol{\beta}}) = (\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\sigma^2\mathbf{V}\mathbf{V}^{-1}\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1} = \sigma^2(\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1}$
  - Hat matrix: $\mathbf{H}_V = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}$
  - Idempotent: $\mathbf{H}_V^2 = \mathbf{H}_V$, $(\mathbf{I}_n - \mathbf{H}_V)^2 = (\mathbf{I}_n - \mathbf{H}_V)$
  - $\mathrm{trace}(\mathbf{H}_V) = \mathrm{trace}(\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}) = \mathrm{trace}(\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1}) = p$
  - $\mathrm{trace}(\mathbf{I}_n - \mathbf{H}_V) = \mathrm{trace}(\mathbf{I}_n) - \mathrm{trace}(\mathbf{H}_V) = (n - p)$
  - $\mathbf{H}_V\mathbf{X} = \mathbf{X}$ and $(\mathbf{I}_n - \mathbf{H}_V)\mathbf{X} = (\mathbf{I}_n - \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}})\mathbf{X} = \mathbf{0}$
- Fitted values: $\widehat{\mathbf{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{y} = \mathbf{H}_V\mathbf{y}$
- Mean-Variance: $\mathbb{E}(\widehat{\mathbf{y}}) = \mathbf{X}\mathbb{E}(\widehat{\boldsymbol{\beta}}) = \mathbf{X}\boldsymbol{\beta}$, $\quad \mathrm{var}(\widehat{\mathbf{y}}) = \sigma^2\mathbf{H}_V$
- Residuals: $\widehat{\mathbf{e}} = (\mathbf{y} - \widehat{\mathbf{y}}) = \mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}} = (\mathbf{I}_n - \mathbf{H}_V)\mathbf{y}$
- Mean-Variance: $\mathbb{E}(\widehat{\mathbf{e}}) = (\mathbf{I}_n - \mathbf{H}_V)\mathbf{X}\boldsymbol{\beta} = \mathbf{0}$, $\quad \mathrm{var}(\widehat{\mathbf{e}}) = \mathbb{E}(\widehat{\mathbf{e}}\widehat{\mathbf{e}}^{\mathrm{T}}) = \sigma^2(\mathbf{I}_n - \mathbf{H}_V)$
- $\mathrm{cov}(\widehat{\mathbf{y}}, \widehat{\mathbf{e}}) = \mathrm{cov}\{\mathbf{H}_V\mathbf{y}, (\mathbf{I}_n - \mathbf{H}_V)\mathbf{y}) = \sigma^2\mathbf{H}_V(\mathbf{I}_n - \mathbf{H}_V) = \sigma^2(\mathbf{H}_V - \mathbf{H}_V^2) = \mathbf{0}$

Generalized / Weighted Least Squares

$$\mathbf{y}^{n\times 1} = \mathbf{X}^{n\times p}\boldsymbol{\beta}^{p\times 1} + \mathbf{e}^{n\times 1}, \quad \mathbb{E}(\mathbf{e}) = \mathbf{0},\ \mathrm{cov}(\mathbf{e}) = \sigma^2\mathbf{V}, \quad \mathbf{V}\ \text{known.}$$

- Minimize squared error loss:

$$\widehat{\boldsymbol{\beta}} = \arg\min(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = (\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{y}$$

- $\mathbb{E}(\widehat{\boldsymbol{\beta}}) = (\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta}$
- $\mathrm{var}(\widehat{\boldsymbol{\beta}}) = (\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\sigma^2\mathbf{V}\mathbf{V}^{-1}\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1} = \sigma^2(\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1}$
  - Hat matrix: $\mathbf{H}_V = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}$
  - Idempotent: $\mathbf{H}_V^2 = \mathbf{H}_V, \quad (\mathbf{I}_n - \mathbf{H}_V)^2 = (\mathbf{I}_n - \mathbf{H}_V)$
  - $\mathrm{trace}(\mathbf{H}_V) = \mathrm{trace}\{\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\} = \mathrm{trace}\{\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1}\} = p$
  - $\mathrm{trace}(\mathbf{I}_n - \mathbf{H}_V) = \mathrm{trace}(\mathbf{I}_n) - \mathrm{trace}(\mathbf{H}_V) = (n-p)$
  - $\mathbf{H}_V\mathbf{X} = \mathbf{X}$ and $(\mathbf{I}_n - \mathbf{H}_V)\mathbf{X} = \{\mathbf{I}_n - \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\}\mathbf{X} = \mathbf{0}$
- Fitted values: $\widehat{\mathbf{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{y} = \mathbf{H}_V\mathbf{y}$
- Mean-Variance: $\mathbb{E}(\widehat{\mathbf{y}}) = \mathbf{X}\mathbb{E}(\widehat{\boldsymbol{\beta}}) = \mathbf{X}\boldsymbol{\beta}, \quad \mathrm{var}(\widehat{\mathbf{y}}) = \sigma^2\mathbf{H}_V$
- Residuals: $\widehat{\mathbf{e}} = (\mathbf{y} - \widehat{\mathbf{y}}) = \mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}} = (\mathbf{I}_n - \mathbf{H}_V)\mathbf{y}$
- Mean-Variance: $\mathbb{E}(\widehat{\mathbf{e}}) = (\mathbf{I}_n - \mathbf{H}_V)\mathbf{X}\boldsymbol{\beta} = \mathbf{0}, \quad \mathrm{var}(\widehat{\mathbf{e}}) = \mathbb{E}(\widehat{\mathbf{e}}\widehat{\mathbf{e}}^{\mathrm{T}}) = \sigma^2(\mathbf{I}_n - \mathbf{H}_V)$
- $\mathrm{cov}(\widehat{\mathbf{y}}, \widehat{\mathbf{e}}) = \mathrm{cov}\{\mathbf{H}_V\mathbf{y}, (\mathbf{I}_n - \mathbf{H}_V)\mathbf{y}) = \sigma^2\mathbf{H}_V(\mathbf{I}_n - \mathbf{H}_V) = \sigma^2(\mathbf{H}_V - \mathbf{H}_V^2) = \mathbf{0}$

$$\mathbf{y}^{n\times 1} = \mathbf{X}^{n\times p}\boldsymbol{\beta}^{p\times 1} + \mathbf{e}^{n\times 1}, \qquad \mathbb{E}(\mathbf{e}) = \mathbf{0},\ \mathrm{cov}(\mathbf{e}) = \sigma^2\mathbf{V}, \quad \mathbf{V} \text{ known.}$$

- Minimize squared error loss:

$$\widehat{\boldsymbol{\beta}} = \arg\min(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = (\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{y}$$

- $\mathbb{E}(\widehat{\boldsymbol{\beta}}) = (\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta}$
- $\mathrm{var}(\widehat{\boldsymbol{\beta}}) = (\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\sigma^2\mathbf{V}\mathbf{V}^{-1}\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1} = \sigma^2(\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1}$
  - Hat matrix: $\mathbf{H}_V = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}$
  - Idempotent: $\mathbf{H}_V^2 = \mathbf{H}_V, \quad (\mathbf{I}_n - \mathbf{H}_V)^2 = (\mathbf{I}_n - \mathbf{H}_V)$
  - $\mathrm{trace}(\mathbf{H}_V) = \mathrm{trace}\{\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\} = \mathrm{trace}\{\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1}\} = p$
  - $\mathrm{trace}(\mathbf{I}_n - \mathbf{H}_V) = \mathrm{trace}(\mathbf{I}_n) - \mathrm{trace}(\mathbf{H}_V) = (n - p)$
  - $\mathbf{H}_V\mathbf{X} = \mathbf{X}$ and $(\mathbf{I}_n - \mathbf{H}_V)\mathbf{X} = \{\mathbf{I}_n - \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\}\mathbf{X} = \mathbf{0}$
- Fitted values: $\widehat{\mathbf{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{y} = \mathbf{H}_V\mathbf{y}$
- Mean-Variance: $\mathbb{E}(\widehat{\mathbf{y}}) = \mathbf{X}\mathbb{E}(\widehat{\boldsymbol{\beta}}) = \mathbf{X}\boldsymbol{\beta}, \quad \mathrm{var}(\widehat{\mathbf{y}}) = \sigma^2\mathbf{H}_V$
- Residuals: $\widehat{\mathbf{e}} = (\mathbf{y} - \widehat{\mathbf{y}}) = \mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}} = (\mathbf{I}_n - \mathbf{H}_V)\mathbf{y}$
- Mean-Variance: $\mathbb{E}(\widehat{\mathbf{e}}) = (\mathbf{I}_n - \mathbf{H}_V)\mathbf{X}\boldsymbol{\beta} = \mathbf{0}, \quad \mathrm{var}(\widehat{\mathbf{e}}) = \mathbb{E}(\widehat{\mathbf{e}}\widehat{\mathbf{e}}^{\mathrm{T}}) = \sigma^2(\mathbf{I}_n - \mathbf{H}_V)$
- $\mathrm{cov}(\widehat{\mathbf{y}}, \widehat{\mathbf{e}}) = \mathrm{cov}\{\mathbf{H}_V\mathbf{y}, (\mathbf{I}_n - \mathbf{H}_V)\mathbf{y}) = \sigma^2\mathbf{H}_V(\mathbf{I}_n - \mathbf{H}_V) = \sigma^2(\mathbf{H}_V - \mathbf{H}_V^2) = \mathbf{0}$

Generalized / Weighted Least Squares

$$\mathbf{y}^{n\times 1} = \mathbf{X}^{n\times p}\boldsymbol{\beta}^{p\times 1} + \mathbf{e}^{n\times 1}, \qquad \mathbb{E}(\mathbf{e}) = \mathbf{0},\ \operatorname{cov}(\mathbf{e}) = \sigma^2\mathbf{V}, \quad \mathbf{V}\text{ known.}$$

- Minimize squared error loss:
  $$\widehat{\boldsymbol{\beta}} = \arg\min(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) = (\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{y}$$
- $\mathbb{E}(\widehat{\boldsymbol{\beta}}) = (\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta}$
- $\operatorname{var}(\widehat{\boldsymbol{\beta}}) = (\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\sigma^2\mathbf{V}\mathbf{V}^{-1}\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1} = \sigma^2(\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1}$
  - Hat matrix: $\mathbf{H}_V = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}$
  - Idempotent: $\mathbf{H}_V^2 = \mathbf{H}_V, \quad (\mathbf{I}_n - \mathbf{H}_V)^2 = (\mathbf{I}_n - \mathbf{H}_V)$
  - $\operatorname{trace}(\mathbf{H}_V) = \operatorname{trace}\{\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\} = \operatorname{trace}\{\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1}\} = p$
  - $\operatorname{trace}(\mathbf{I}_n - \mathbf{H}_V) = \operatorname{trace}(\mathbf{I}_n) - \operatorname{trace}(\mathbf{H}_V) = (n-p)$
  - $\mathbf{H}_V\mathbf{X} = \mathbf{X}$ and $(\mathbf{I}_n - \mathbf{H}_V)\mathbf{X} = \{\mathbf{I}_n - \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\}\mathbf{X} = \mathbf{0}$
- Fitted values: $\widehat{\mathbf{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{y} = \mathbf{H}_V\mathbf{y}$
- Mean-Variance: $\mathbb{E}(\widehat{\mathbf{y}}) = \mathbf{X}\mathbb{E}(\widehat{\boldsymbol{\beta}}) = \mathbf{X}\boldsymbol{\beta}, \quad \operatorname{var}(\widehat{\mathbf{y}}) = \sigma^2\mathbf{H}_V$
- Residuals: $\widehat{\mathbf{e}} = (\mathbf{y}-\widehat{\mathbf{y}}) = \mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}} = (\mathbf{I}_n - \mathbf{H}_V)\mathbf{y}$
- Mean-Variance: $\mathbb{E}(\widehat{\mathbf{e}}) = (\mathbf{I}_n - \mathbf{H}_V)\mathbf{X}\boldsymbol{\beta} = \mathbf{0}, \quad \operatorname{var}(\widehat{\mathbf{e}}) = \mathbb{E}(\widehat{\mathbf{e}}\widehat{\mathbf{e}}^{\mathrm{T}}) = \sigma^2(\mathbf{I}_n - \mathbf{H}_V)$
- $\operatorname{cov}(\widehat{\mathbf{y}}, \widehat{\mathbf{e}}) = \operatorname{cov}\{\mathbf{H}_V\mathbf{y}, (\mathbf{I}_n - \mathbf{H}_V)\mathbf{y}) = \sigma^2\mathbf{H}_V(\mathbf{I}_n - \mathbf{H}_V) = \sigma^2(\mathbf{H}_V - \mathbf{H}_V^2) = \mathbf{0}$

Generalized / Weighted Least Squares

$$\mathbf{y}^{n\times 1} = \mathbf{X}^{n\times p}\boldsymbol{\beta}^{p\times 1} + \mathbf{e}^{n\times 1}, \quad \mathbb{E}(\mathbf{e}) = \mathbf{0},\ \mathrm{cov}(\mathbf{e}) = \sigma^2\mathbf{V}, \quad \mathbf{V}\ \text{known.}$$

- Minimize squared error loss:

$$\widehat{\boldsymbol{\beta}} = \arg\min(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = (\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{y}$$

- $\mathbb{E}(\widehat{\boldsymbol{\beta}}) = (\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta}$
- $\mathrm{var}(\widehat{\boldsymbol{\beta}}) = (\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\sigma^2\mathbf{V}\mathbf{V}^{-1}\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1} = \sigma^2(\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1}$
  - Hat matrix: $\mathbf{H}_V = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}$
  - Idempotent: $\mathbf{H}_V^2 = \mathbf{H}_V, \quad (\mathbf{I}_n - \mathbf{H}_V)^2 = (\mathbf{I}_n - \mathbf{H}_V)$
  - $\mathrm{trace}(\mathbf{H}_V) = \mathrm{trace}\{\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\} = \mathrm{trace}\{\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1}\} = p$
  - $\mathrm{trace}(\mathbf{I}_n - \mathbf{H}_V) = \mathrm{trace}(\mathbf{I}_n) - \mathrm{trace}(\mathbf{H}_V) = (n-p)$
  - $\mathbf{H}_V\mathbf{X} = \mathbf{X}$ and $(\mathbf{I}_n - \mathbf{H}_V)\mathbf{X} = \{\mathbf{I}_n - \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\}\mathbf{X} = \mathbf{0}$
- Fitted values: $\widehat{\mathbf{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{y} = \mathbf{H}_V\mathbf{y}$
- Mean-Variance: $\mathbb{E}(\widehat{\mathbf{y}}) = \mathbf{X}\mathbb{E}(\widehat{\boldsymbol{\beta}}) = \mathbf{X}\boldsymbol{\beta}, \quad \mathrm{var}(\widehat{\mathbf{y}}) = \sigma^2\mathbf{H}_V$
- Residuals: $\widehat{\mathbf{e}} = (\mathbf{y} - \widehat{\mathbf{y}}) = \mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}} = (\mathbf{I}_n - \mathbf{H}_V)\mathbf{y}$
- Mean-Variance: $\mathbb{E}(\widehat{\mathbf{e}}) = (\mathbf{I}_n - \mathbf{H}_V)\mathbf{X}\boldsymbol{\beta} = \mathbf{0}, \quad \mathrm{var}(\widehat{\mathbf{e}}) = \mathbb{E}(\widehat{\mathbf{e}}\widehat{\mathbf{e}}^{\mathrm{T}}) = \sigma^2(\mathbf{I}_n - \mathbf{H}_V)$
- $\mathrm{cov}(\widehat{\mathbf{y}}, \widehat{\mathbf{e}}) = \mathrm{cov}\{\mathbf{H}_V\mathbf{y}, (\mathbf{I}_n - \mathbf{H}_V)\mathbf{y}) = \sigma^2\mathbf{H}_V(\mathbf{I}_n - \mathbf{H}_V) = \sigma^2(\mathbf{H}_V - \mathbf{H}_V^2) = \mathbf{0}$

Generalized / Weighted Least Squares

$$\mathbf{y}^{n\times 1} = \mathbf{X}^{n\times p}\boldsymbol{\beta}^{p\times 1} + \mathbf{e}^{n\times 1}, \qquad \mathbb{E}(\mathbf{e}) = \mathbf{0},\ \mathrm{cov}(\mathbf{e}) = \sigma^2\mathbf{V}, \quad \mathbf{V} \text{ known.}$$

- Minimize squared error loss:
$$\widehat{\boldsymbol{\beta}} = \arg\min(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) = (\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{y}$$
- $\mathbb{E}(\widehat{\boldsymbol{\beta}}) = (\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta}$
- $\mathrm{var}(\widehat{\boldsymbol{\beta}}) = (\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\sigma^2\mathbf{V}\mathbf{V}^{-1}\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1} = \sigma^2(\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1}$
  - Hat matrix: $\mathbf{H}_V = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}$
  - Idempotent: $\mathbf{H}_V^2 = \mathbf{H}_V, \quad (\mathbf{I}_n - \mathbf{H}_V)^2 = (\mathbf{I}_n - \mathbf{H}_V)$
  - $\mathrm{trace}(\mathbf{H}_V) = \mathrm{trace}\{\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\} = \mathrm{trace}\{\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1}\} = p$
  - $\mathrm{trace}(\mathbf{I}_n - \mathbf{H}_V) = \mathrm{trace}(\mathbf{I}_n) - \mathrm{trace}(\mathbf{H}_V) = (n-p)$
  - $\mathbf{H}_V\mathbf{X} = \mathbf{X}$ and $(\mathbf{I}_n - \mathbf{H}_V)\mathbf{X} = \{\mathbf{I}_n - \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\}\mathbf{X} = \mathbf{0}$
- Fitted values: $\widehat{\mathbf{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{y} = \mathbf{H}_V\mathbf{y}$
- Mean-Variance: $\mathbb{E}(\widehat{\mathbf{y}}) = \mathbf{X}\mathbb{E}(\widehat{\boldsymbol{\beta}}) = \mathbf{X}\boldsymbol{\beta}, \quad \mathrm{var}(\widehat{\mathbf{y}}) = \sigma^2\mathbf{H}_V$
- Residuals: $\widehat{\mathbf{e}} = (\mathbf{y} - \widehat{\mathbf{y}}) = \mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}} = (\mathbf{I}_n - \mathbf{H}_V)\mathbf{y}$
- Mean-Variance: $\mathbb{E}(\widehat{\mathbf{e}}) = (\mathbf{I}_n - \mathbf{H}_V)\mathbf{X}\boldsymbol{\beta} = \mathbf{0}, \quad \mathrm{var}(\widehat{\mathbf{e}}) = \mathbb{E}(\widehat{\mathbf{e}}\widehat{\mathbf{e}}^{\mathrm{T}}) = \sigma^2(\mathbf{I}_n - \mathbf{H}_V)$
- $\mathrm{cov}(\widehat{\mathbf{y}}, \widehat{\mathbf{e}}) = \mathrm{cov}\{\mathbf{H}_V\mathbf{y}, (\mathbf{I}_n - \mathbf{H}_V)\mathbf{y}) = \sigma^2\mathbf{H}_V(\mathbf{I}_n - \mathbf{H}_V) = \sigma^2(\mathbf{H}_V - \mathbf{H}_V^2) = \mathbf{0}$

Generalized / Weighted Least Squares

$$\mathbf{y}^{n\times 1} = \mathbf{X}^{n\times p}\boldsymbol{\beta}^{p\times 1} + \mathbf{e}^{n\times 1}, \qquad \mathbb{E}(\mathbf{e}) = \mathbf{0},\ \operatorname{cov}(\mathbf{e}) = \sigma^2\mathbf{V}, \quad \mathbf{V} \text{ known.}$$

- Minimize squared error loss:

$$\widehat{\boldsymbol{\beta}} = \arg\min(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = (\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{y}$$

- $\mathbb{E}(\widehat{\boldsymbol{\beta}}) = (\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta}$
- $\operatorname{var}(\widehat{\boldsymbol{\beta}}) = (\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\sigma^2\mathbf{V}\mathbf{V}^{-1}\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1} = \sigma^2(\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1}$
  - Hat matrix: $\mathbf{H}_V = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}$
  - Idempotent: $\mathbf{H}_V^2 = \mathbf{H}_V, \quad (\mathbf{I}_n - \mathbf{H}_V)^2 = (\mathbf{I}_n - \mathbf{H}_V)$
  - $\operatorname{trace}(\mathbf{H}_V) = \operatorname{trace}\{\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\} = \operatorname{trace}\{\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1}\} = p$
  - $\operatorname{trace}(\mathbf{I}_n - \mathbf{H}_V) = \operatorname{trace}(\mathbf{I}_n) - \operatorname{trace}(\mathbf{H}_V) = (n-p)$
  - $\mathbf{H}_V\mathbf{X} = \mathbf{X}$ and $(\mathbf{I}_n - \mathbf{H}_V)\mathbf{X} = \{\mathbf{I}_n - \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\}\mathbf{X} = \mathbf{0}$
- Fitted values: $\widehat{\mathbf{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{y} = \mathbf{H}_V\mathbf{y}$
- Mean-Variance: $\mathbb{E}(\widehat{\mathbf{y}}) = \mathbf{X}\mathbb{E}(\widehat{\boldsymbol{\beta}}) = \mathbf{X}\boldsymbol{\beta}, \quad \operatorname{var}(\widehat{\mathbf{y}}) = \sigma^2\mathbf{H}_V$
- Residuals: $\widehat{\mathbf{e}} = (\mathbf{y} - \widehat{\mathbf{y}}) = \mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}} = (\mathbf{I}_n - \mathbf{H}_V)\mathbf{y}$
- Mean-Variance: $\mathbb{E}(\widehat{\mathbf{e}}) = (\mathbf{I}_n - \mathbf{H}_V)\mathbf{X}\boldsymbol{\beta} = \mathbf{0}, \quad \operatorname{var}(\widehat{\mathbf{e}}) = \mathbb{E}(\widehat{\mathbf{e}}\widehat{\mathbf{e}}^{\mathrm{T}}) = \sigma^2(\mathbf{I}_n - \mathbf{H}_V)$
- $\operatorname{cov}(\widehat{\mathbf{y}}, \widehat{\mathbf{e}}) = \operatorname{cov}\{\mathbf{H}_V\mathbf{y}, (\mathbf{I}_n - \mathbf{H}_V)\mathbf{y}) = \sigma^2\mathbf{H}_V(\mathbf{I}_n - \mathbf{H}_V) = \sigma^2(\mathbf{H}_V - \mathbf{H}_V^2) = \mathbf{0}$

Generalized / Weighted Least Squares

$$\mathbf{y}^{n\times 1} = \mathbf{X}^{n\times p}\boldsymbol{\beta}^{p\times 1} + \mathbf{e}^{n\times 1}, \qquad \mathbb{E}(\mathbf{e}) = \mathbf{0},\ \mathrm{cov}(\mathbf{e}) = \sigma^2\mathbf{V}, \quad \mathbf{V} \text{ known.}$$

- Minimize squared error loss:

$$\widehat{\boldsymbol{\beta}} = \arg\min(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = (\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{y}$$

- $\mathbb{E}(\widehat{\boldsymbol{\beta}}) = (\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta}$
- $\mathrm{var}(\widehat{\boldsymbol{\beta}}) = (\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\sigma^2\mathbf{V}\mathbf{V}^{-1}\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1} = \sigma^2(\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1}$
    - Hat matrix: $\mathbf{H}_V = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}$
    - Idempotent: $\mathbf{H}_V^2 = \mathbf{H}_V, \quad (\mathbf{I}_n - \mathbf{H}_V)^2 = (\mathbf{I}_n - \mathbf{H}_V)$
    - $\mathrm{trace}(\mathbf{H}_V) = \mathrm{trace}\{\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\} = \mathrm{trace}\{\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1}\} = p$
    - $\mathrm{trace}(\mathbf{I}_n - \mathbf{H}_V) = \mathrm{trace}(\mathbf{I}_n) - \mathrm{trace}(\mathbf{H}_V) = (n-p)$
    - $\mathbf{H}_V\mathbf{X} = \mathbf{X}$ and $(\mathbf{I}_n - \mathbf{H}_V)\mathbf{X} = \{\mathbf{I}_n - \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\}\mathbf{X} = \mathbf{0}$
- Fitted values: $\widehat{\mathbf{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{y} = \mathbf{H}_V\mathbf{y}$
- Mean-Variance: $\mathbb{E}(\widehat{\mathbf{y}}) = \mathbf{X}\mathbb{E}(\widehat{\boldsymbol{\beta}}) = \mathbf{X}\boldsymbol{\beta}, \quad \mathrm{var}(\widehat{\mathbf{y}}) = \sigma^2\mathbf{H}_V$
- Residuals: $\widehat{\mathbf{e}} = (\mathbf{y} - \widehat{\mathbf{y}}) = \mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}} = (\mathbf{I}_n - \mathbf{H}_V)\mathbf{y}$
- Mean-Variance: $\mathbb{E}(\widehat{\mathbf{e}}) = (\mathbf{I}_n - \mathbf{H}_V)\mathbf{X}\boldsymbol{\beta} = \mathbf{0}, \quad \mathrm{var}(\widehat{\mathbf{e}}) = \mathbb{E}(\widehat{\mathbf{e}}\widehat{\mathbf{e}}^{\mathrm{T}}) = \sigma^2(\mathbf{I}_n - \mathbf{H}_V)$
- $\mathrm{cov}(\widehat{\mathbf{y}}, \widehat{\mathbf{e}}) = \mathrm{cov}\{\mathbf{H}_V\mathbf{y}, (\mathbf{I}_n - \mathbf{H}_V)\mathbf{y}) = \sigma^2\mathbf{H}_V(\mathbf{I}_n - \mathbf{H}_V) = \sigma^2(\mathbf{H}_V - \mathbf{H}_V^2) = \mathbf{0}$

Generalized / Weighted Least Squares

$$\mathbf{y}^{n\times 1} = \mathbf{X}^{n\times p}\boldsymbol{\beta}^{p\times 1} + \mathbf{e}^{n\times 1}, \qquad \mathbb{E}(\mathbf{e}) = \mathbf{0},\ \mathrm{cov}(\mathbf{e}) = \sigma^2\mathbf{V}, \quad \mathbf{V} \text{ known.}$$

- Minimize squared error loss:

$$\widehat{\boldsymbol{\beta}} = \arg\min(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) = (\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{y}$$

- $\mathbb{E}(\widehat{\boldsymbol{\beta}}) = (\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta}$
- $\mathrm{var}(\widehat{\boldsymbol{\beta}}) = (\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\sigma^2\mathbf{V}\mathbf{V}^{-1}\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1} = \sigma^2(\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1}$
  - Hat matrix: $\mathbf{H}_V = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}$
  - Idempotent: $\mathbf{H}_V^2 = \mathbf{H}_V, \quad (\mathbf{I}_n - \mathbf{H}_V)^2 = (\mathbf{I}_n - \mathbf{H}_V)$
  - $\mathrm{trace}(\mathbf{H}_V) = \mathrm{trace}\{\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\} = \mathrm{trace}\{\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1}\} = p$
  - $\mathrm{trace}(\mathbf{I}_n - \mathbf{H}_V) = \mathrm{trace}(\mathbf{I}_n) - \mathrm{trace}(\mathbf{H}_V) = (n-p)$
  - $\mathbf{H}_V\mathbf{X} = \mathbf{X}$ and $(\mathbf{I}_n - \mathbf{H}_V)\mathbf{X} = \{\mathbf{I}_n - \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\}\mathbf{X} = \mathbf{0}$
- Fitted values: $\widehat{\mathbf{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{y} = \mathbf{H}_V\mathbf{y}$
- Mean-Variance: $\mathbb{E}(\widehat{\mathbf{y}}) = \mathbf{X}\mathbb{E}(\widehat{\boldsymbol{\beta}}) = \mathbf{X}\boldsymbol{\beta}, \quad \mathrm{var}(\widehat{\mathbf{y}}) = \sigma^2\mathbf{H}_V$
- Residuals: $\widehat{\mathbf{e}} = (\mathbf{y} - \widehat{\mathbf{y}}) = \mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}} = (\mathbf{I}_n - \mathbf{H}_V)\mathbf{y}$
- Mean-Variance: $\mathbb{E}(\widehat{\mathbf{e}}) = (\mathbf{I}_n - \mathbf{H}_V)\mathbf{X}\boldsymbol{\beta} = \mathbf{0}, \quad \mathrm{var}(\widehat{\mathbf{e}}) = \mathbb{E}(\widehat{\mathbf{e}}\widehat{\mathbf{e}}^{\mathrm{T}}) = \sigma^2(\mathbf{I}_n - \mathbf{H}_V)$
- $\mathrm{cov}(\widehat{\mathbf{y}}, \widehat{\mathbf{e}}) = \mathrm{cov}\{\mathbf{H}_V\mathbf{y}, (\mathbf{I}_n - \mathbf{H}_V)\mathbf{y}) = \sigma^2\mathbf{H}_V(\mathbf{I}_n - \mathbf{H}_V) = \sigma^2(\mathbf{H}_V - \mathbf{H}_V^2) = \mathbf{0}$

Generalized / Weighted Least Squares

$$\mathbf{y}^{n\times 1} = \mathbf{X}^{n\times p}\boldsymbol{\beta}^{p\times 1} + \mathbf{e}^{n\times 1}, \quad \mathbb{E}(\mathbf{e}) = \mathbf{0},\ \text{cov}(\mathbf{e}) = \sigma^2\mathbf{V}, \quad \mathbf{V}\ \text{known.}$$

- Minimize squared error loss:

$$\widehat{\boldsymbol{\beta}} = \arg\min(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = (\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{y}$$

- $\mathbb{E}(\widehat{\boldsymbol{\beta}}) = (\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta}$
- $\text{var}(\widehat{\boldsymbol{\beta}}) = (\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\sigma^2\mathbf{V}\mathbf{V}^{-1}\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1} = \sigma^2(\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1}$
    - Hat matrix: $\mathbf{H}_V = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}$
    - Idempotent: $\mathbf{H}_V^2 = \mathbf{H}_V, \quad (\mathbf{I}_n - \mathbf{H}_V)^2 = (\mathbf{I}_n - \mathbf{H}_V)$
    - $\text{trace}(\mathbf{H}_V) = \text{trace}\{\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\} = \text{trace}\{\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1}\} = p$
    - $\text{trace}(\mathbf{I}_n - \mathbf{H}_V) = \text{trace}(\mathbf{I}_n) - \text{trace}(\mathbf{H}_V) = (n-p)$
    - $\mathbf{H}_V\mathbf{X} = \mathbf{X}$ and $(\mathbf{I}_n - \mathbf{H}_V)\mathbf{X} = \{\mathbf{I}_n - \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\}\mathbf{X} = \mathbf{0}$
- Fitted values: $\widehat{\mathbf{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{y} = \mathbf{H}_V\mathbf{y}$
- Mean-Variance: $\mathbb{E}(\widehat{\mathbf{y}}) = \mathbf{X}\mathbb{E}(\widehat{\boldsymbol{\beta}}) = \mathbf{X}\boldsymbol{\beta}, \quad \text{var}(\widehat{\mathbf{y}}) = \sigma^2\mathbf{H}_V$
- Residuals: $\widehat{\mathbf{e}} = (\mathbf{y} - \widehat{\mathbf{y}}) = \mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}} = (\mathbf{I}_n - \mathbf{H}_V)\mathbf{y}$
- Mean-Variance: $\mathbb{E}(\widehat{\mathbf{e}}) = (\mathbf{I}_n - \mathbf{H}_V)\mathbf{X}\boldsymbol{\beta} = \mathbf{0}, \quad \text{var}(\widehat{\mathbf{e}}) = \mathbb{E}(\widehat{\mathbf{e}}\widehat{\mathbf{e}}^{\mathrm{T}}) = \sigma^2(\mathbf{I}_n - \mathbf{H}_V)$
- $\text{cov}(\widehat{\mathbf{y}}, \widehat{\mathbf{e}}) = \text{cov}\{\mathbf{H}_V\mathbf{y}, (\mathbf{I}_n - \mathbf{H}_V)\mathbf{y}) = \sigma^2\mathbf{H}_V(\mathbf{I}_n - \mathbf{H}_V) = \sigma^2(\mathbf{H}_V - \mathbf{H}_V^2) = \mathbf{0}$

Generalized / Weighted Least Squares

$$\mathbf{y}^{n\times 1} = \mathbf{X}^{n\times p}\boldsymbol{\beta}^{p\times 1} + \mathbf{e}^{n\times 1}, \quad \mathbb{E}(\mathbf{e}) = \mathbf{0},\ \mathrm{cov}(\mathbf{e}) = \sigma^2\mathbf{V}, \quad \mathbf{V} \text{ known.}$$

- Minimize squared error loss:

$$\widehat{\boldsymbol{\beta}} = \arg\min(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = (\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{y}$$

- $\mathbb{E}(\widehat{\boldsymbol{\beta}}) = (\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta}$
- $\mathrm{var}(\widehat{\boldsymbol{\beta}}) = (\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\sigma^2\mathbf{V}\mathbf{V}^{-1}\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1} = \sigma^2(\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1}$
  - Hat matrix: $\mathbf{H}_V = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}$
  - Idempotent: $\mathbf{H}_V^2 = \mathbf{H}_V, \quad (\mathbf{I}_n - \mathbf{H}_V)^2 = (\mathbf{I}_n - \mathbf{H}_V)$
  - $\mathrm{trace}(\mathbf{H}_V) = \mathrm{trace}\{\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\} = \mathrm{trace}\{\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1}\} = p$
  - $\mathrm{trace}(\mathbf{I}_n - \mathbf{H}_V) = \mathrm{trace}(\mathbf{I}_n) - \mathrm{trace}(\mathbf{H}_V) = (n-p)$
  - $\mathbf{H}_V\mathbf{X} = \mathbf{X}$ and $(\mathbf{I}_n - \mathbf{H}_V)\mathbf{X} = \{\mathbf{I}_n - \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\}\mathbf{X} = \mathbf{0}$
- Fitted values: $\widehat{\mathbf{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{y} = \mathbf{H}_V\mathbf{y}$
- Mean-Variance: $\mathbb{E}(\widehat{\mathbf{y}}) = \mathbf{X}\mathbb{E}(\widehat{\boldsymbol{\beta}}) = \mathbf{X}\boldsymbol{\beta}, \quad \mathrm{var}(\widehat{\mathbf{y}}) = \sigma^2\mathbf{H}_V$
- Residuals: $\widehat{\mathbf{e}} = (\mathbf{y} - \widehat{\mathbf{y}}) = \mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}} = (\mathbf{I}_n - \mathbf{H}_V)\mathbf{y}$
- Mean-Variance: $\mathbb{E}(\widehat{\mathbf{e}}) = (\mathbf{I}_n - \mathbf{H}_V)\mathbf{X}\boldsymbol{\beta} = \mathbf{0}, \quad \mathrm{var}(\widehat{\mathbf{e}}) = \mathbb{E}(\widehat{\mathbf{e}}\widehat{\mathbf{e}}^{\mathrm{T}}) = \sigma^2(\mathbf{I}_n - \mathbf{H}_V)$
- $\mathrm{cov}(\widehat{\mathbf{y}}, \widehat{\mathbf{e}}) = \mathrm{cov}\{\mathbf{H}_V\mathbf{y}, (\mathbf{I}_n - \mathbf{H}_V)\mathbf{y}\} = \sigma^2\mathbf{H}_V(\mathbf{I}_n - \mathbf{H}_V) = \sigma^2(\mathbf{H}_V - \mathbf{H}_V^2) = \mathbf{0}$

Generalized / Weighted Least Squares

$$\mathbf{y}^{n\times 1} = \mathbf{X}^{n\times p}\boldsymbol{\beta}^{p\times 1} + \mathbf{e}^{n\times 1}, \quad \mathbb{E}(\mathbf{e}) = \mathbf{0},\ \text{cov}(\mathbf{e}) = \sigma^2\mathbf{V}, \quad \mathbf{V}\text{ known.}$$

- Minimize squared error loss:
$$\widehat{\boldsymbol{\beta}} = \arg\min(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) = (\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{y}$$
- $\mathbb{E}(\widehat{\boldsymbol{\beta}}) = (\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta}$
- $\text{var}(\widehat{\boldsymbol{\beta}}) = (\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\sigma^2\mathbf{V}\mathbf{V}^{-1}\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1} = \sigma^2(\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1}$
  - Hat matrix: $\mathbf{H}_V = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}$
  - Idempotent: $\mathbf{H}_V^2 = \mathbf{H}_V, \quad (\mathbf{I}_n - \mathbf{H}_V)^2 = (\mathbf{I}_n - \mathbf{H}_V)$
  - $\text{trace}(\mathbf{H}_V) = \text{trace}\{\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\} = \text{trace}\{\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1}\} = p$
  - $\text{trace}(\mathbf{I}_n - \mathbf{H}_V) = \text{trace}(\mathbf{I}_n) - \text{trace}(\mathbf{H}_V) = (n-p)$
  - $\mathbf{H}_V\mathbf{X} = \mathbf{X}$ and $(\mathbf{I}_n - \mathbf{H}_V)\mathbf{X} = \{\mathbf{I}_n - \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\}\mathbf{X} = \mathbf{0}$
- Fitted values: $\widehat{\mathbf{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{y} = \mathbf{H}_V\mathbf{y}$
- Mean-Variance: $\mathbb{E}(\widehat{\mathbf{y}}) = \mathbf{X}\mathbb{E}(\widehat{\boldsymbol{\beta}}) = \mathbf{X}\boldsymbol{\beta}, \quad \text{var}(\widehat{\mathbf{y}}) = \sigma^2\mathbf{H}_V$
- Residuals: $\widehat{\mathbf{e}} = (\mathbf{y} - \widehat{\mathbf{y}}) = \mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}} = (\mathbf{I}_n - \mathbf{H}_V)\mathbf{y}$
- Mean-Variance: $\mathbb{E}(\widehat{\mathbf{e}}) = (\mathbf{I}_n - \mathbf{H}_V)\mathbf{X}\boldsymbol{\beta} = \mathbf{0}, \quad \text{var}(\widehat{\mathbf{e}}) = \mathbb{E}(\widehat{\mathbf{e}}\widehat{\mathbf{e}}^{\mathrm{T}}) = \sigma^2(\mathbf{I}_n - \mathbf{H}_V)$
- $\text{cov}(\widehat{\mathbf{y}}, \widehat{\mathbf{e}}) = \text{cov}\{\mathbf{H}_V\mathbf{y}, (\mathbf{I}_n - \mathbf{H}_V)\mathbf{y}) = \sigma^2\mathbf{H}_V(\mathbf{I}_n - \mathbf{H}_V) = \sigma^2(\mathbf{H}_V - \mathbf{H}_V^2) = \mathbf{0}$

$$\mathbf{y}^{n\times 1} = \mathbf{X}^{n\times p}\boldsymbol{\beta}^{p\times 1} + \mathbf{e}^{n\times 1}, \quad \mathbb{E}(\mathbf{e}) = \mathbf{0},\ \mathrm{cov}(\mathbf{e}) = \sigma^2\mathbf{V}, \quad \mathbf{V} \text{ known.}$$

- Likelihood: $L(\boldsymbol{\beta}, \sigma^2) = \dfrac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{-\dfrac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right\}$

$$\propto \frac{1}{(\sigma^2)^{n/2}} \exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X}\boldsymbol{\beta} - 2\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{y} + \mathbf{y}^{\mathrm{T}}\mathbf{y})\right\}$$

- Semi-conjugate priors: $p(\boldsymbol{\beta}, \sigma^2) = \mathrm{MVN}(\boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta) \cdot \text{Inv-Ga}(a_\sigma, b_\sigma)$

$$\propto \frac{1}{(\sigma^2)^{a_\sigma+1}} \exp\left(-\frac{b_\sigma}{\sigma^2}\right) \exp\left\{-\frac{1}{2}(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\Sigma}_\beta^{-1}\boldsymbol{\beta} - 2\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\Sigma}_\beta^{-1}\boldsymbol{\mu}_\beta + \boldsymbol{\mu}_\beta^{\mathrm{T}}\boldsymbol{\Sigma}_\beta^{-1}\boldsymbol{\mu}_\beta)\right\}$$

- Posterior: $p(\boldsymbol{\beta}, \sigma^2 \mid \mathbf{y}_{1:n}) \propto L(\boldsymbol{\beta}, \sigma^2) \times p(\boldsymbol{\beta}, \sigma^2)$
- Posterior full conditionals for block-Gibbs sampler:

$$\mathbf{y}^{n\times 1} = \mathbf{X}^{n\times p}\boldsymbol{\beta}^{p\times 1} + \mathbf{e}^{n\times 1}, \qquad \mathbb{E}(\mathbf{e}) = \mathbf{0},\ \mathrm{cov}(\mathbf{e}) = \sigma^2\mathbf{V}, \quad \mathbf{V}\ \text{known.}$$

- Likelihood: $L(\boldsymbol{\beta}, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{-\frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\right\}$

$$\propto \frac{1}{(\sigma^2)^{n/2}} \exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X}\boldsymbol{\beta} - 2\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{y} + \mathbf{y}^{\mathrm{T}}\mathbf{y})\right\}$$

- Semi-conjugate priors: $p(\boldsymbol{\beta}, \sigma^2) = \mathrm{MVN}(\boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta) \cdot \text{Inv-Ga}(a_\sigma, b_\sigma)$

$$\propto \frac{1}{(\sigma^2)^{a_\sigma+1}} \exp\left(-\frac{b_\sigma}{\sigma^2}\right) \exp\left\{-\frac{1}{2}(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\Sigma}_\beta^{-1}\boldsymbol{\beta} - 2\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\Sigma}_\beta^{-1}\boldsymbol{\mu}_\beta + \boldsymbol{\mu}_\beta^{\mathrm{T}}\boldsymbol{\Sigma}_\beta^{-1}\boldsymbol{\mu}_\beta)\right\}$$

- Posterior: $p(\boldsymbol{\beta}, \sigma^2 \mid \mathbf{y}_{1:n}) \propto L(\boldsymbol{\beta}, \sigma^2) \times p(\boldsymbol{\beta}, \sigma^2)$
- Posterior full conditionals for block-Gibbs sampler:

Bayesian Generalized / Weighted Linear Models

$$\mathbf{y}^{n\times 1} = \mathbf{X}^{n\times p}\boldsymbol{\beta}^{p\times 1} + \mathbf{e}^{n\times 1}, \qquad \mathbb{E}(\mathbf{e}) = \mathbf{0},\ \mathrm{cov}(\mathbf{e}) = \sigma^2\mathbf{V}, \quad \mathbf{V} \text{ known.}$$

- Likelihood: $L(\boldsymbol{\beta}, \sigma^2) = \dfrac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{-\dfrac{1}{2\sigma^2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\right\}$

$$\propto \frac{1}{(\sigma^2)^{n/2}} \exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X}\boldsymbol{\beta} - 2\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{y} + \mathbf{y}^{\mathrm{T}}\mathbf{y})\right\}$$

- Semi-conjugate priors: $p(\boldsymbol{\beta}, \sigma^2) = \mathrm{MVN}(\boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta) \cdot \text{Inv-Ga}(a_\sigma, b_\sigma)$

$$\propto \frac{1}{(\sigma^2)^{a_\sigma+1}} \exp\left(-\frac{b_\sigma}{\sigma^2}\right) \exp\left\{-\frac{1}{2}(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\Sigma}_\beta^{-1}\boldsymbol{\beta} - 2\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\Sigma}_\beta^{-1}\boldsymbol{\mu}_\beta + \boldsymbol{\mu}_\beta^{\mathrm{T}}\boldsymbol{\Sigma}_\beta^{-1}\boldsymbol{\mu}_\beta)\right\}$$

- Posterior: $p(\boldsymbol{\beta}, \sigma^2 \mid \mathbf{y}_{1:n}) \propto L(\boldsymbol{\beta}, \sigma^2) \times p(\boldsymbol{\beta}, \sigma^2)$
- Posterior full conditionals for block-Gibbs sampler:

Bayesian Generalized / Weighted Linear Models

$$\mathbf{y}^{n\times 1} = \mathbf{X}^{n\times p}\boldsymbol{\beta}^{p\times 1} + \mathbf{e}^{n\times 1}, \qquad \mathbb{E}(\mathbf{e}) = \mathbf{0},\ \mathrm{cov}(\mathbf{e}) = \sigma^2\mathbf{V}, \quad \mathbf{V}\ \text{known}.$$

- Likelihood: $L(\boldsymbol{\beta},\sigma^2) = \dfrac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{-\dfrac{1}{2\sigma^2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\right\}$

$$\propto \frac{1}{(\sigma^2)^{n/2}} \exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X}\boldsymbol{\beta} - 2\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{y} + \mathbf{y}^{\mathrm{T}}\mathbf{y})\right\}$$

- Semi-conjugate priors: $p(\boldsymbol{\beta},\sigma^2) = \mathrm{MVN}(\boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta)\cdot \text{Inv-Ga}(a_\sigma, b_\sigma)$

$$\propto \frac{1}{(\sigma^2)^{a_\sigma+1}} \exp\left(-\frac{b_\sigma}{\sigma^2}\right) \exp\left\{-\frac{1}{2}(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\Sigma}_\beta^{-1}\boldsymbol{\beta} - 2\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\Sigma}_\beta^{-1}\boldsymbol{\mu}_\beta + \boldsymbol{\mu}_\beta^{\mathrm{T}}\boldsymbol{\Sigma}_\beta^{-1}\boldsymbol{\mu}_\beta)\right\}$$

- Posterior: $p(\boldsymbol{\beta},\sigma^2 \mid \mathbf{y}_{1:n}) \propto L(\boldsymbol{\beta},\sigma^2)\times p(\boldsymbol{\beta},\sigma^2)$
- Posterior full conditionals for block-Gibbs sampler:
  - $p(\boldsymbol{\beta} \mid -) \propto \exp\left[-\dfrac{1}{2}\left\{\boldsymbol{\beta}^{\mathrm{T}}(\boldsymbol{\Sigma}_\beta^{-1} + \sigma^{-2}\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})\boldsymbol{\beta} - 2\boldsymbol{\beta}^{\mathrm{T}}\left(\boldsymbol{\Sigma}_\beta^{-1}\boldsymbol{\mu}_\beta + \sigma^{-2}\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{y}\right)\right\}\right]$

    $\equiv \mathrm{MVN}(\boldsymbol{\mu}_{\beta,n}, \boldsymbol{\Sigma}_{\beta,n})$,

    $\boldsymbol{\Sigma}_{\beta,n} = (\boldsymbol{\Sigma}_\beta^{-1} + \sigma^{-2}\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1}, \quad \boldsymbol{\mu}_{\beta,n} = \boldsymbol{\Sigma}_{\beta,n}\left(\boldsymbol{\Sigma}_\beta^{-1}\boldsymbol{\mu}_\beta + \sigma^{-2}\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{y}\right)$
  - $p(\sigma^2 \mid -) \propto \dfrac{1}{(\sigma^2)^{a_\sigma + \frac{n}{2} + 1}} \exp\left[-\dfrac{1}{\sigma^2}\left\{b_\sigma + \dfrac{1}{2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\right\}\right]$

    $\equiv \text{Inv-Ga}\left\{a_\sigma + \dfrac{n}{2}, b_\sigma + \dfrac{1}{2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\right\}$

$$\mathbf{y}^{n\times 1} = \mathbf{X}^{n\times p}\boldsymbol{\beta}^{p\times 1} + \mathbf{e}^{n\times 1}, \quad \mathbb{E}(\mathbf{e}) = \mathbf{0},\ \mathrm{cov}(\mathbf{e}) = \sigma^2\mathbf{V}, \quad \mathbf{V} \text{ known}.$$

- Likelihood: $L(\boldsymbol{\beta}, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right\}$
$$\propto \frac{1}{(\sigma^2)^{n/2}} \exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X}\boldsymbol{\beta} - 2\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{y} + \mathbf{y}^{\mathrm{T}}\mathbf{y})\right\}$$
- Semi-conjugate priors: $p(\boldsymbol{\beta}, \sigma^2) = \mathrm{MVN}(\boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta) \cdot \text{Inv-Ga}(a_\sigma, b_\sigma)$
- Posterior: $p(\boldsymbol{\beta}, \sigma^2 \mid \mathbf{y}_{1:n}) \propto L(\boldsymbol{\beta}, \sigma^2) \times p(\boldsymbol{\beta}, \sigma^2)$
- Posterior full conditionals for block-Gibbs sampler:
  - $p(\boldsymbol{\beta} \mid -) = \mathrm{MVN}(\boldsymbol{\mu}_{\beta,n}, \boldsymbol{\Sigma}_{\beta,n})$,
  $$\boldsymbol{\Sigma}_{\beta,n} = (\boldsymbol{\Sigma}_\beta^{-1} + \sigma^{-2}\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1}, \quad \boldsymbol{\mu}_{\beta,n} = \boldsymbol{\Sigma}_{\beta,n}\left(\boldsymbol{\Sigma}_\beta^{-1}\boldsymbol{\mu}_\beta + \sigma^{-2}\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{y}\right)$$
  - $p(\sigma^2 \mid -) = \text{Inv-Ga}\left\{a_\sigma + \frac{n}{2}, b_\sigma + \frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right\}$
- Bayesian weighted Ridge and LASSO linear models can be similarly developed.

Linear Mixed Models

$$y_{i,j} = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \mathbf{z}_i^{\mathrm{T}}\mathbf{u}_i + \epsilon_{i,j}, \quad \mathbf{u}_i \overset{iid}{\sim} f_{\mathbf{u}}, \quad \epsilon_{i,j} \overset{iid}{\sim} f_{\epsilon},$$
$$i = 1, \ldots, n, \quad j = 1, \ldots, m_i$$
$$\text{with} \quad \mathbb{E}_{f_{\mathbf{u}}}(\mathbf{u}) = \mathbf{0} \quad \text{and} \quad \mathbb{E}_{f_{\epsilon}}(\epsilon) = 0.$$

- $y_{i,j} = \beta_0 + x_i\beta_1 + u_i + \epsilon_{i,j}, \quad u_i \overset{iid}{\sim} \text{Normal}(0, \sigma_u^2), \quad \epsilon_{i,j} \overset{iid}{\sim} \text{Normal}(0, \sigma_\epsilon^2)$
- $y_{i,j} = \beta_0 + x_i\beta_1 + x_i^2\beta_2 + u_i + \epsilon_{i,j}, \quad u_i \overset{iid}{\sim} \text{Normal}(0, \sigma_u^2), \quad \epsilon_{i,j} \overset{iid}{\sim} \text{Normal}(0, \sigma_\epsilon^2)$
- $y_{i,j} = \beta_0 + x_{i,1}\beta_1 + x_{i,2}\beta_2 + \cdots + x_{i,p}\beta_p + u_i + \epsilon_{i,j},$
  $u_i \overset{iid}{\sim} \text{Normal}(0, \sigma_u^2), \quad \epsilon_{i,j} \overset{iid}{\sim} \text{Normal}(0, \sigma_\epsilon^2)$
- $y_{i,j} = \beta_0 + x_i\beta_1 + u_{0,i} + x_i u_{1,i} + \epsilon_{i,j}, \quad \mathbf{u}_i \overset{iid}{\sim} \text{MVN}(\mathbf{0}, \boldsymbol{\Sigma}_u), \quad \epsilon_{i,j} \overset{iid}{\sim} \text{Normal}(0, \sigma_\epsilon^2)$
- $y_{i,j} = \beta_0 + x_i\beta_1 + x_i^2\beta_2 + u_{0,i} + x_i u_{1,i} + \epsilon_{i,j},$
  $\mathbf{u}_i \overset{iid}{\sim} \text{MVN}(\mathbf{0}, \boldsymbol{\Sigma}_u), \quad \epsilon_{i,j} \overset{iid}{\sim} \text{Normal}(0, \sigma_\epsilon^2)$
- $y_{i,j} = \beta_0 + x_{i,1}\beta_1 + x_{i,2}\beta_2 + u_{0,i} + x_{1,i}u_{1,i} + \epsilon_{i,j},$
  $\mathbf{u}_i \overset{iid}{\sim} \text{MVN}(\mathbf{0}, \boldsymbol{\Sigma}_u), \quad \epsilon_{i,j} \overset{iid}{\sim} \text{Normal}(0, \sigma_\epsilon^2)$
- $y_{i,j} = \beta_0 + x_i\beta_1 + u_{0,i} + z_i u_{1,i} + \epsilon_{i,j},$
  $\mathbf{u}_i \overset{iid}{\sim} \text{MVN}(\mathbf{0}, \boldsymbol{\Sigma}_u), \quad \epsilon_{i,j} \overset{iid}{\sim} \text{Normal}(0, \sigma_\epsilon^2)$

$$
\begin{gathered}
y_{i,j} = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \mathbf{z}_i^{\mathrm{T}}\mathbf{u}_i + \epsilon_{i,j}, \quad \mathbf{u}_i \overset{iid}{\sim} f_{\mathbf{u}}, \quad \epsilon_{i,j} \overset{iid}{\sim} f_{\epsilon}, \\
i = 1, \dots, n, \quad j = 1, \dots, m_i \\
\text{with} \quad \mathbb{E}_{f_{\mathbf{u}}}(\mathbf{u}) = \mathbf{0} \quad \text{and} \quad \mathbb{E}_{f_{\epsilon}}(\epsilon) = 0.
\end{gathered}
$$

- $y_{i,j} = \beta_0 + x_i\beta_1 + u_i + \epsilon_{i,j}, \quad u_i \overset{iid}{\sim} \text{Normal}(0, \sigma_u^2), \quad \epsilon_{i,j} \overset{iid}{\sim} \text{Normal}(0, \sigma_\epsilon^2)$
- $y_{i,j} = \beta_0 + x_i\beta_1 + x_i^2\beta_2 + u_i + \epsilon_{i,j}, \quad u_i \overset{iid}{\sim} \text{Normal}(0, \sigma_u^2), \quad \epsilon_{i,j} \overset{iid}{\sim} \text{Normal}(0, \sigma_\epsilon^2)$
- $y_{i,j} = \beta_0 + x_{i,1}\beta_1 + x_{i,2}\beta_2 + \cdots + x_{i,p}\beta_p + u_i + \epsilon_{i,j}, \quad u_i \overset{iid}{\sim} \text{Normal}(0, \sigma_u^2), \quad \epsilon_{i,j} \overset{iid}{\sim} \text{Normal}(0, \sigma_\epsilon^2)$
- $y_{i,j} = \beta_0 + x_i\beta_1 + u_{0,i} + x_iu_{1,i} + \epsilon_{i,j}, \quad \mathbf{u}_i \overset{iid}{\sim} \text{MVN}(\mathbf{0}, \boldsymbol{\Sigma}_u), \quad \epsilon_{i,j} \overset{iid}{\sim} \text{Normal}(0, \sigma_\epsilon^2)$
- $y_{i,j} = \beta_0 + x_i\beta_1 + x_i^2\beta_2 + u_{0,i} + x_iu_{1,i} + \epsilon_{i,j}, \quad \mathbf{u}_i \overset{iid}{\sim} \text{MVN}(\mathbf{0}, \boldsymbol{\Sigma}_u), \quad \epsilon_{i,j} \overset{iid}{\sim} \text{Normal}(0, \sigma_\epsilon^2)$
- $y_{i,j} = \beta_0 + x_{i,1}\beta_1 + x_{i,2}\beta_2 + u_{0,i} + x_{1,i}u_{1,i} + \epsilon_{i,j}, \quad \mathbf{u}_i \overset{iid}{\sim} \text{MVN}(\mathbf{0}, \boldsymbol{\Sigma}_u), \quad \epsilon_{i,j} \overset{iid}{\sim} \text{Normal}(0, \sigma_\epsilon^2)$
- $y_{i,j} = \beta_0 + x_i\beta_1 + u_{0,i} + z_iu_{1,i} + \epsilon_{i,j}, \quad \mathbf{u}_i \overset{iid}{\sim} \text{MVN}(\mathbf{0}, \boldsymbol{\Sigma}_u), \quad \epsilon_{i,j} \overset{iid}{\sim} \text{Normal}(0, \sigma_\epsilon^2)$

$$
\begin{gathered}
y_{i,j} = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \mathbf{z}_i^{\mathrm{T}}\mathbf{u}_i + \epsilon_{i,j}, \quad \mathbf{u}_i \overset{iid}{\sim} f_{\mathbf{u}}, \quad \epsilon_{i,j} \overset{iid}{\sim} f_{\epsilon}, \\
i = 1, \ldots, n, \quad j = 1, \ldots, m_i \\
\text{with} \quad \mathbb{E}_{f_{\mathbf{u}}}(\mathbf{u}) = \mathbf{0} \quad \text{and} \quad \mathbb{E}_{f_{\epsilon}}(\epsilon) = 0.
\end{gathered}
$$

- $y_{i,j} = \beta_0 + x_i\beta_1 + u_i + \epsilon_{i,j}, \qquad u_i \overset{iid}{\sim} \text{Normal}(0, \sigma_u^2), \quad \epsilon_{i,j} \overset{iid}{\sim} \text{Normal}(0, \sigma_\epsilon^2)$
- $y_{i,j} = \beta_0 + x_i\beta_1 + x_i^2\beta_2 + u_i + \epsilon_{i,j}, \qquad u_i \overset{iid}{\sim} \text{Normal}(0, \sigma_u^2), \quad \epsilon_{i,j} \overset{iid}{\sim} \text{Normal}(0, \sigma_\epsilon^2)$
- $y_{i,j} = \beta_0 + x_{i,1}\beta_1 + x_{i,2}\beta_2 + \cdots + x_{i,p}\beta_p + u_i + \epsilon_{i,j},$
  $u_i \overset{iid}{\sim} \text{Normal}(0, \sigma_u^2), \quad \epsilon_{i,j} \overset{iid}{\sim} \text{Normal}(0, \sigma_\epsilon^2)$
- $y_{i,j} = \beta_0 + x_i\beta_1 + u_{0,i} + x_iu_{1,i} + \epsilon_{i,j}, \qquad \mathbf{u}_i \overset{iid}{\sim} \text{MVN}(\mathbf{0}, \boldsymbol{\Sigma}_u), \quad \epsilon_{i,j} \overset{iid}{\sim} \text{Normal}(0, \sigma_\epsilon^2)$
- $y_{i,j} = \beta_0 + x_i\beta_1 + x_i^2\beta_2 + u_{0,i} + x_iu_{1,i} + \epsilon_{i,j},$
  $\mathbf{u}_i \overset{iid}{\sim} \text{MVN}(\mathbf{0}, \boldsymbol{\Sigma}_u), \quad \epsilon_{i,j} \overset{iid}{\sim} \text{Normal}(0, \sigma_\epsilon^2)$
- $y_{i,j} = \beta_0 + x_{i,1}\beta_1 + x_{i,2}\beta_2 + u_{0,i} + x_{1,i}u_{1,i} + \epsilon_{i,j},$
  $\mathbf{u}_i \overset{iid}{\sim} \text{MVN}(\mathbf{0}, \boldsymbol{\Sigma}_u), \quad \epsilon_{i,j} \overset{iid}{\sim} \text{Normal}(0, \sigma_\epsilon^2)$
- $y_{i,j} = \beta_0 + x_i\beta_1 + u_{0,i} + z_iu_{1,i} + \epsilon_{i,j},$
  $\mathbf{u}_i \overset{iid}{\sim} \text{MVN}(\mathbf{0}, \boldsymbol{\Sigma}_u), \quad \epsilon_{i,j} \overset{iid}{\sim} \text{Normal}(0, \sigma_\epsilon^2)$

Linear Mixed Models

$$y_{i,j} = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \mathbf{z}_i^{\mathrm{T}}\mathbf{u}_i + \epsilon_{i,j}, \quad \mathbf{u}_i \overset{iid}{\sim} f_{\mathbf{u}}, \quad \epsilon_{i,j} \overset{iid}{\sim} f_\epsilon,$$
$$i = 1, \ldots, n, \quad j = 1, \ldots, m_i$$
$$\text{with} \quad \mathbb{E}_{f_{\mathbf{u}}}(\mathbf{u}) = \mathbf{0} \quad \text{and} \quad \mathbb{E}_{f_\epsilon}(\epsilon) = 0.$$

- $y_{i,j} = \beta_0 + x_i\beta_1 + u_i + \epsilon_{i,j}, \qquad u_i \overset{iid}{\sim} \text{Normal}(0, \sigma_u^2), \quad \epsilon_{i,j} \overset{iid}{\sim} \text{Normal}(0, \sigma_\epsilon^2)$
- $y_{i,j} = \beta_0 + x_i\beta_1 + x_i^2\beta_2 + u_i + \epsilon_{i,j}, \qquad u_i \overset{iid}{\sim} \text{Normal}(0, \sigma_u^2), \quad \epsilon_{i,j} \overset{iid}{\sim} \text{Normal}(0, \sigma_\epsilon^2)$
- $y_{i,j} = \beta_0 + x_{i,1}\beta_1 + x_{i,2}\beta_2 + \cdots + x_{i,p}\beta_p + u_i + \epsilon_{i,j},$
  $u_i \overset{iid}{\sim} \text{Normal}(0, \sigma_u^2), \quad \epsilon_{i,j} \overset{iid}{\sim} \text{Normal}(0, \sigma_\epsilon^2)$
- $y_{i,j} = \beta_0 + x_i\beta_1 + u_{0,i} + x_iu_{1,i} + \epsilon_{i,j}, \qquad \mathbf{u}_i \overset{iid}{\sim} \text{MVN}(\mathbf{0}, \boldsymbol{\Sigma}_u), \quad \epsilon_{i,j} \overset{iid}{\sim} \text{Normal}(0, \sigma_\epsilon^2)$
- $y_{i,j} = \beta_0 + x_i\beta_1 + x_i^2\beta_2 + u_{0,i} + x_iu_{1,i} + \epsilon_{i,j},$
  $\mathbf{u}_i \overset{iid}{\sim} \text{MVN}(\mathbf{0}, \boldsymbol{\Sigma}_u), \quad \epsilon_{i,j} \overset{iid}{\sim} \text{Normal}(0, \sigma_\epsilon^2)$
- $y_{i,j} = \beta_0 + x_{i,1}\beta_1 + x_{i,2}\beta_2 + u_{0,i} + x_{1,i}u_{1,i} + \epsilon_{i,j},$
  $\mathbf{u}_i \overset{iid}{\sim} \text{MVN}(\mathbf{0}, \boldsymbol{\Sigma}_u), \quad \epsilon_{i,j} \overset{iid}{\sim} \text{Normal}(0, \sigma_\epsilon^2)$
- $y_{i,j} = \beta_0 + x_i\beta_1 + u_{0,i} + z_iu_{1,i} + \epsilon_{i,j},$
  $\mathbf{u}_i \overset{iid}{\sim} \text{MVN}(\mathbf{0}, \boldsymbol{\Sigma}_u), \quad \epsilon_{i,j} \overset{iid}{\sim} \text{Normal}(0, \sigma_\epsilon^2)$

Linear Mixed Models

$$y_{i,j} = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \mathbf{z}_i^{\mathrm{T}}\mathbf{u}_i + \epsilon_{i,j}, \quad \mathbf{u}_i \overset{iid}{\sim} f_{\mathbf{u}}, \quad \epsilon_{i,j} \overset{iid}{\sim} f_{\epsilon},$$
$$i = 1, \ldots, n, \quad j = 1, \ldots, m_i$$
$$\text{with} \quad \mathbb{E}_{f_{\mathbf{u}}}(\mathbf{u}) = \mathbf{0} \quad \text{and} \quad \mathbb{E}_{f_{\epsilon}}(\epsilon) = 0.$$

- $y_{i,j} = \beta_0 + x_i\beta_1 + u_i + \epsilon_{i,j}, \qquad u_i \overset{iid}{\sim} \text{Normal}(0, \sigma_u^2), \quad \epsilon_{i,j} \overset{iid}{\sim} \text{Normal}(0, \sigma_\epsilon^2)$
- $y_{i,j} = \beta_0 + x_i\beta_1 + x_i^2\beta_2 + u_i + \epsilon_{i,j}, \qquad u_i \overset{iid}{\sim} \text{Normal}(0, \sigma_u^2), \quad \epsilon_{i,j} \overset{iid}{\sim} \text{Normal}(0, \sigma_\epsilon^2)$
- $y_{i,j} = \beta_0 + x_{i,1}\beta_1 + x_{i,2}\beta_2 + \cdots + x_{i,p}\beta_p + u_i + \epsilon_{i,j},$
  $u_i \overset{iid}{\sim} \text{Normal}(0, \sigma_u^2), \quad \epsilon_{i,j} \overset{iid}{\sim} \text{Normal}(0, \sigma_\epsilon^2)$
- $y_{i,j} = \beta_0 + x_i\beta_1 + u_{0,i} + x_i u_{1,i} + \epsilon_{i,j}, \qquad \mathbf{u}_i \overset{iid}{\sim} \text{MVN}(\mathbf{0}, \boldsymbol{\Sigma}_u), \quad \epsilon_{i,j} \overset{iid}{\sim} \text{Normal}(0, \sigma_\epsilon^2)$
- $y_{i,j} = \beta_0 + x_i\beta_1 + x_i^2\beta_2 + u_{0,i} + x_i u_{1,i} + \epsilon_{i,j},$
  $\mathbf{u}_i \overset{iid}{\sim} \text{MVN}(\mathbf{0}, \boldsymbol{\Sigma}_u), \quad \epsilon_{i,j} \overset{iid}{\sim} \text{Normal}(0, \sigma_\epsilon^2)$
- $y_{i,j} = \beta_0 + x_{i,1}\beta_1 + x_{i,2}\beta_2 + u_{0,i} + x_{1,i}u_{1,i} + \epsilon_{i,j},$
  $\mathbf{u}_i \overset{iid}{\sim} \text{MVN}(\mathbf{0}, \boldsymbol{\Sigma}_u), \quad \epsilon_{i,j} \overset{iid}{\sim} \text{Normal}(0, \sigma_\epsilon^2)$
- $y_{i,j} = \beta_0 + x_i\beta_1 + u_{0,i} + z_i u_{1,i} + \epsilon_{i,j},$
  $\mathbf{u}_i \overset{iid}{\sim} \text{MVN}(\mathbf{0}, \boldsymbol{\Sigma}_u), \quad \epsilon_{i,j} \overset{iid}{\sim} \text{Normal}(0, \sigma_\epsilon^2)$

Linear Mixed Models

$$y_{i,j} = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \mathbf{z}_i^{\mathrm{T}}\mathbf{u}_i + \epsilon_{i,j}, \quad \mathbf{u}_i \overset{iid}{\sim} f_{\mathbf{u}}, \quad \epsilon_{i,j} \overset{iid}{\sim} f_{\epsilon},$$
$$i = 1, \ldots, n, \quad j = 1, \ldots, m_i$$
$$\text{with} \quad \mathbb{E}_{f_{\mathbf{u}}}(\mathbf{u}) = \mathbf{0} \quad \text{and} \quad \mathbb{E}_{f_{\epsilon}}(\epsilon) = 0.$$

- $y_{i,j} = \beta_0 + x_i\beta_1 + u_i + \epsilon_{i,j}, \quad u_i \overset{iid}{\sim} \text{Normal}(0, \sigma_u^2), \quad \epsilon_{i,j} \overset{iid}{\sim} \text{Normal}(0, \sigma_\epsilon^2)$
- $y_{i,j} = \beta_0 + x_i\beta_1 + x_i^2\beta_2 + u_i + \epsilon_{i,j}, \quad u_i \overset{iid}{\sim} \text{Normal}(0, \sigma_u^2), \quad \epsilon_{i,j} \overset{iid}{\sim} \text{Normal}(0, \sigma_\epsilon^2)$
- $y_{i,j} = \beta_0 + x_{i,1}\beta_1 + x_{i,2}\beta_2 + \cdots + x_{i,p}\beta_p + u_i + \epsilon_{i,j},$
  $u_i \overset{iid}{\sim} \text{Normal}(0, \sigma_u^2), \quad \epsilon_{i,j} \overset{iid}{\sim} \text{Normal}(0, \sigma_\epsilon^2)$
- $y_{i,j} = \beta_0 + x_i\beta_1 + u_{0,i} + x_iu_{1,i} + \epsilon_{i,j}, \quad \mathbf{u}_i \overset{iid}{\sim} \text{MVN}(\mathbf{0}, \boldsymbol{\Sigma}_u), \quad \epsilon_{i,j} \overset{iid}{\sim} \text{Normal}(0, \sigma_\epsilon^2)$
- $y_{i,j} = \beta_0 + x_i\beta_1 + x_i^2\beta_2 + u_{0,i} + x_iu_{1,i} + \epsilon_{i,j},$
  $\mathbf{u}_i \overset{iid}{\sim} \text{MVN}(\mathbf{0}, \boldsymbol{\Sigma}_u), \quad \epsilon_{i,j} \overset{iid}{\sim} \text{Normal}(0, \sigma_\epsilon^2)$
- $y_{i,j} = \beta_0 + x_{i,1}\beta_1 + x_{i,2}\beta_2 + u_{0,i} + x_{1,i}u_{1,i} + \epsilon_{i,j},$
  $\mathbf{u}_i \overset{iid}{\sim} \text{MVN}(\mathbf{0}, \boldsymbol{\Sigma}_u), \quad \epsilon_{i,j} \overset{iid}{\sim} \text{Normal}(0, \sigma_\epsilon^2)$
- $y_{i,j} = \beta_0 + x_i\beta_1 + u_{0,i} + z_iu_{1,i} + \epsilon_{i,j},$
  $\mathbf{u}_i \overset{iid}{\sim} \text{MVN}(\mathbf{0}, \boldsymbol{\Sigma}_u), \quad \epsilon_{i,j} \overset{iid}{\sim} \text{Normal}(0, \sigma_\epsilon^2)$

$$y_{i,j} = \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta} + \mathbf{z}_i^{\mathrm{T}} \mathbf{u}_i + \epsilon_{i,j}, \quad \mathbf{u}_i \overset{iid}{\sim} f_{\mathbf{u}}, \quad \epsilon_{i,j} \overset{iid}{\sim} f_{\epsilon},$$
$$i = 1, \ldots, n, \quad j = 1, \ldots, m_i$$
$$\text{with} \quad \mathbb{E}_{f_{\mathbf{u}}}(\mathbf{u}) = \mathbf{0} \quad \text{and} \quad \mathbb{E}_{f_{\epsilon}}(\epsilon) = 0.$$

- $y_{i,j} = \beta_0 + x_i\beta_1 + u_i + \epsilon_{i,j}, \quad u_i \overset{iid}{\sim} \text{Normal}(0, \sigma_u^2), \quad \epsilon_{i,j} \overset{iid}{\sim} \text{Normal}(0, \sigma_\epsilon^2)$
- $y_{i,j} = \beta_0 + x_i\beta_1 + x_i^2\beta_2 + u_i + \epsilon_{i,j}, \quad u_i \overset{iid}{\sim} \text{Normal}(0, \sigma_u^2), \quad \epsilon_{i,j} \overset{iid}{\sim} \text{Normal}(0, \sigma_\epsilon^2)$
- $y_{i,j} = \beta_0 + x_{i,1}\beta_1 + x_{i,2}\beta_2 + \cdots + x_{i,p}\beta_p + u_i + \epsilon_{i,j},$
  $u_i \overset{iid}{\sim} \text{Normal}(0, \sigma_u^2), \quad \epsilon_{i,j} \overset{iid}{\sim} \text{Normal}(0, \sigma_\epsilon^2)$
- $y_{i,j} = \beta_0 + x_i\beta_1 + u_{0,i} + x_i u_{1,i} + \epsilon_{i,j}, \quad \mathbf{u}_i \overset{iid}{\sim} \text{MVN}(\mathbf{0}, \boldsymbol{\Sigma}_u), \quad \epsilon_{i,j} \overset{iid}{\sim} \text{Normal}(0, \sigma_\epsilon^2)$
- $y_{i,j} = \beta_0 + x_i\beta_1 + x_i^2\beta_2 + u_{0,i} + x_i u_{1,i} + \epsilon_{i,j},$
  $\mathbf{u}_i \overset{iid}{\sim} \text{MVN}(\mathbf{0}, \boldsymbol{\Sigma}_u), \quad \epsilon_{i,j} \overset{iid}{\sim} \text{Normal}(0, \sigma_\epsilon^2)$
- $y_{i,j} = \beta_0 + x_{i,1}\beta_1 + x_{i,2}\beta_2 + u_{0,i} + x_{1,i} u_{1,i} + \epsilon_{i,j},$
  $\mathbf{u}_i \overset{iid}{\sim} \text{MVN}(\mathbf{0}, \boldsymbol{\Sigma}_u), \quad \epsilon_{i,j} \overset{iid}{\sim} \text{Normal}(0, \sigma_\epsilon^2)$
- $y_{i,j} = \beta_0 + x_i\beta_1 + u_{0,i} + z_i u_{1,i} + \epsilon_{i,j},$
  $\mathbf{u}_i \overset{iid}{\sim} \text{MVN}(\mathbf{0}, \boldsymbol{\Sigma}_u), \quad \epsilon_{i,j} \overset{iid}{\sim} \text{Normal}(0, \sigma_\epsilon^2)$

$$\text{Individual level model:} \quad \mathbb{E}_{\epsilon}(y_{i,j}) = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \mathbf{z}_i^{\mathrm{T}}\mathbf{u}_i,$$
$$\text{Population level model:} \quad \mathbb{E}_{\epsilon,\mathbf{u}}(y_{i,j}) = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}.$$

- $y_{i,j} = \beta_0 + x_i\beta_1 + u_i + \epsilon_{i,j},$ $\quad \mathbb{E}_{\epsilon,\mathbf{u}}(y_{i,j}) = \beta_0 + x_i\beta_1$
- $y_{i,j} = \beta_0 + x_i\beta_1 + x_i^2\beta_2 + u_i + \epsilon_{i,j},$ $\quad \mathbb{E}_{\epsilon,\mathbf{u}}(y_{i,j}) = \beta_0 + x_i\beta_1 + x_i^2\beta_2$
- $y_{i,j} = \beta_0 + x_{i,1}\beta_1 + \cdots + x_{i,p}\beta_p + u_i + \epsilon_{i,j},$ $\quad \mathbb{E}_{\epsilon,\mathbf{u}}(y_{i,j}) = \beta_0 + x_{i,1}\beta_1 + \cdots + x_{i,p}\beta_p$
- $y_{i,j} = \beta_0 + x_i\beta_1 + u_{0,i} + x_iu_{1,i} + \epsilon_{i,j},$ $\quad \mathbb{E}_{\epsilon,\mathbf{u}}(y_{i,j}) = \beta_0 + x_i\beta_1$
- $y_{i,j} = \beta_0 + x_i\beta_1 + x_i^2\beta_2 + u_{0,i} + x_iu_{1,i} + \epsilon_{i,j},$ $\quad \mathbb{E}_{\epsilon,\mathbf{u}}(y_{i,j}) = \beta_0 + x_i\beta_1 + x_i^2\beta_2$
- $y_{i,j} = \beta_0 + x_{i,1}\beta_1 + x_{i,2}\beta_2 + u_{0,i} + x_{1,i}u_{1,i} + \epsilon_{i,j},$ $\quad \mathbb{E}_{\epsilon,\mathbf{u}}(y_{i,j}) = \beta_0 + x_{i,1}\beta_1 + x_{i,2}\beta_2$
- $y_{i,j} = \beta_0 + x_{i,1}\beta_1 + u_{0,i} + z_iu_{1,i} + \epsilon_{i,j},$ $\quad \mathbb{E}_{\epsilon,\mathbf{u}}(y_{i,j}) = \beta_0 + x_{i,1}\beta_1$

REML - Restricted Maximum Likelihood

$$\text{Likelihood:} \quad \mathcal{L}(\boldsymbol{\theta} \mid \mathbf{y}) = \log \prod_{i=1}^{n} \prod_{j=1}^{m_i} f_y(y_{i,j} \mid \boldsymbol{\theta})$$

$$= \log \prod_{i=1}^{n} \prod_{j=1}^{m_i} \int_{\mathbb{R}^q} f_y(y_{i,j}, \mathbf{u}_i \mid \boldsymbol{\theta}) f_{\mathbf{u}}(\mathbf{u}_i \mid \boldsymbol{\theta}) d\mathbf{u}_i$$

$$\text{MLE:} \quad \widehat{\boldsymbol{\theta}} = \arg\max \mathcal{L}(\boldsymbol{\theta} \mid \mathbf{y})$$

- We have $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e} \sim \text{MVN}(\mathbf{X}\boldsymbol{\beta}, \mathbf{Z}\boldsymbol{\Sigma}_u\mathbf{Z}^{\mathrm{T}} + \sigma^2\mathbf{I}) = \text{MVN}(\mathbf{X}\boldsymbol{\beta}, \mathbf{V})$.
- If the variance components $\mathbf{V}$ were known, we could estimate β by

$$\widehat{\beta} = (\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{y}.$$

- If there exists a $\mathbf{B}$, such that $\mathbf{B}\mathbf{X} = \mathbf{0}$, we have

$$\mathbf{B}\mathbf{y} = \mathbf{B}\mathbf{X}\beta + \mathbf{B}\mathbf{Z}\mathbf{u} + \mathbf{B}\mathbf{e} = \mathbf{B}\mathbf{Z}\mathbf{u} + \mathbf{B}\mathbf{e} \sim \text{MVN}(\mathbf{0}, \mathbf{B}\mathbf{V}\mathbf{B}^{\mathrm{T}}).$$

- $(\mathbf{I} - \mathbf{H})$, with $\mathbf{H} = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}$, is a candidate for $\mathbf{B}$ since $(\mathbf{I} - \mathbf{H})\mathbf{X} = \mathbf{0}$.
- Use $\mathbf{B}\mathbf{y}$ to estimate $\mathbf{V}$ by some $\widehat{\mathbf{V}}_{REML}$.
- Estimate β then by $\widehat{\beta}_{REML} = (\mathbf{X}^{\mathrm{T}}\widehat{\mathbf{V}}_{REML}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\widehat{\mathbf{V}}_{REML}^{-1}\mathbf{y}$.

REML - Restricted Maximum Likelihood

$$\text{Likelihood:} \quad \mathcal{L}(\boldsymbol{\theta} \mid \mathbf{y}) = \log \prod_{i=1}^{n} \prod_{j=1}^{m_i} f_y(y_{i,j} \mid \boldsymbol{\theta})$$
$$= \log \prod_{i=1}^{n} \prod_{j=1}^{m_i} \int_{\mathbb{R}^q} f_y(y_{i,j}, \mathbf{u}_i \mid \boldsymbol{\theta}) f_{\mathbf{u}}(\mathbf{u}_i \mid \boldsymbol{\theta}) d\mathbf{u}_i$$
$$\text{MLE:} \quad \widehat{\boldsymbol{\theta}} = \arg\max \mathcal{L}(\boldsymbol{\theta} \mid \mathbf{y})$$

- We have $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e} \sim \text{MVN}(\mathbf{X}\boldsymbol{\beta}, \mathbf{Z}\boldsymbol{\Sigma}_u\mathbf{Z}^{\mathrm{T}} + \sigma^2\mathbf{I}) = \text{MVN}(\mathbf{X}\boldsymbol{\beta}, \mathbf{V})$.
- If the variance components $\mathbf{V}$ were known, we could estimate $\boldsymbol{\beta}$ by

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{y}.$$

- If there exists a $\mathbf{B}$, such that $\mathbf{B}\mathbf{X} = \mathbf{0}$, we have

$$\mathbf{B}\mathbf{y} = \mathbf{B}\mathbf{X}\boldsymbol{\beta} + \mathbf{B}\mathbf{Z}\mathbf{u} + \mathbf{B}\mathbf{e} = \mathbf{B}\mathbf{Z}\mathbf{u} + \mathbf{B}\mathbf{e} \sim \text{MVN}(\mathbf{0}, \mathbf{B}\mathbf{V}\mathbf{B}^{\mathrm{T}}).$$

- $(\mathbf{I} - \mathbf{H})$, with $\mathbf{H} = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}$, is a candidate for $\mathbf{B}$ since $(\mathbf{I} - \mathbf{H})\mathbf{X} = \mathbf{0}$.
- Use $\mathbf{B}\mathbf{y}$ to estimate $\mathbf{V}$ by some $\widehat{\mathbf{V}}_{REML}$.
- Estimate $\boldsymbol{\beta}$ then by $\widehat{\boldsymbol{\beta}}_{REML} = (\mathbf{X}^{\mathrm{T}}\widehat{\mathbf{V}}_{REML}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\widehat{\mathbf{V}}_{REML}^{-1}\mathbf{y}$.

$$\text{Likelihood:}\quad \mathcal{L}(\boldsymbol{\theta} \mid \mathbf{y}) = \log \prod_{i=1}^{n} \prod_{j=1}^{m_i} f_y(y_{i,j} \mid \boldsymbol{\theta})$$
$$= \log \prod_{i=1}^{n} \prod_{j=1}^{m_i} \int_{\mathbb{R}^q} f_y(y_{i,j}, \mathbf{u}_i \mid \boldsymbol{\theta}) f_{\mathbf{u}}(\mathbf{u}_i \mid \boldsymbol{\theta}) d\mathbf{u}_i$$
$$\text{MLE:}\quad \widehat{\boldsymbol{\theta}} = \arg\max \mathcal{L}(\boldsymbol{\theta} \mid \mathbf{y})$$

- We have $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e} \sim \text{MVN}(\mathbf{X}\boldsymbol{\beta}, \mathbf{Z}\boldsymbol{\Sigma}_u\mathbf{Z}^{\mathrm{T}} + \sigma^2\mathbf{I}) = \text{MVN}(\mathbf{X}\boldsymbol{\beta}, \mathbf{V})$.
- If the variance components $\mathbf{V}$ were known, we could estimate $\boldsymbol{\beta}$ by

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{y}.$$

- If there exists a $\mathbf{B}$, such that $\mathbf{B}\mathbf{X} = \mathbf{0}$, we have

$$\mathbf{B}\mathbf{y} = \mathbf{B}\mathbf{X}\boldsymbol{\beta} + \mathbf{B}\mathbf{Z}\mathbf{u} + \mathbf{B}\mathbf{e} = \mathbf{B}\mathbf{Z}\mathbf{u} + \mathbf{B}\mathbf{e} \sim \text{MVN}(\mathbf{0}, \mathbf{B}\mathbf{V}\mathbf{B}^{\mathrm{T}}).$$

- $(\mathbf{I} - \mathbf{H})$, with $\mathbf{H} = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}$, is a candidate for $\mathbf{B}$ since $(\mathbf{I} - \mathbf{H})\mathbf{X} = \mathbf{0}$.
- Use $\mathbf{B}\mathbf{y}$ to estimate $\mathbf{V}$ by some $\widehat{\mathbf{V}}_{REML}$.
- Estimate $\boldsymbol{\beta}$ then by $\widehat{\boldsymbol{\beta}}_{REML} = (\mathbf{X}^{\mathrm{T}}\widehat{\mathbf{V}}_{REML}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\widehat{\mathbf{V}}_{REML}^{-1}\mathbf{y}$.

REML - Restricted Maximum Likelihood

$$\text{Likelihood:} \quad \mathcal{L}(\boldsymbol{\theta} \mid \mathbf{y}) = \log \prod_{i=1}^{n} \prod_{j=1}^{m_i} f_y(y_{i,j} \mid \boldsymbol{\theta})$$

$$= \log \prod_{i=1}^{n} \prod_{j=1}^{m_i} \int_{\mathbb{R}^q} f_y(y_{i,j}, \mathbf{u}_i \mid \boldsymbol{\theta}) f_{\mathbf{u}}(\mathbf{u}_i \mid \boldsymbol{\theta}) d\mathbf{u}_i$$

$$\text{MLE:} \quad \widehat{\boldsymbol{\theta}} = \arg\max \mathcal{L}(\boldsymbol{\theta} \mid \mathbf{y})$$

- We have $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e} \sim \text{MVN}(\mathbf{X}\boldsymbol{\beta}, \mathbf{Z}\boldsymbol{\Sigma}_u\mathbf{Z}^{\mathrm{T}} + \sigma^2\mathbf{I}) = \text{MVN}(\mathbf{X}\boldsymbol{\beta}, \mathbf{V})$.
- If the variance components $\mathbf{V}$ were known, we could estimate $\boldsymbol{\beta}$ by

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{y}.$$

- If there exists a $\mathbf{B}$, such that $\mathbf{B}\mathbf{X} = \mathbf{0}$, we have

$$\mathbf{B}\mathbf{y} = \mathbf{B}\mathbf{X}\boldsymbol{\beta} + \mathbf{B}\mathbf{Z}\mathbf{u} + \mathbf{B}\mathbf{e} = \mathbf{B}\mathbf{Z}\mathbf{u} + \mathbf{B}\mathbf{e} \sim \text{MVN}(\mathbf{0}, \mathbf{B}\mathbf{V}\mathbf{B}^{\mathrm{T}}).$$

- $(\mathbf{I} - \mathbf{H})$, with $\mathbf{H} = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}$, is a candidate for $\mathbf{B}$ since $(\mathbf{I} - \mathbf{H})\mathbf{X} = \mathbf{0}$.
- Use $\mathbf{B}\mathbf{y}$ to estimate $\mathbf{V}$ by some $\widehat{\mathbf{V}}_{REML}$.
- Estimate $\boldsymbol{\beta}$ then by $\widehat{\boldsymbol{\beta}}_{REML} = (\mathbf{X}^{\mathrm{T}}\widehat{\mathbf{V}}_{REML}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\widehat{\mathbf{V}}_{REML}^{-1}\mathbf{y}$.

REML - Restricted Maximum Likelihood

$$\text{Likelihood:} \quad \mathcal{L}(\boldsymbol{\theta} \mid \mathbf{y}) = \log \prod_{i=1}^{n} \prod_{j=1}^{m_i} f_y(y_{i,j} \mid \boldsymbol{\theta})$$
$$= \log \prod_{i=1}^{n} \prod_{j=1}^{m_i} \int_{\mathbb{R}^q} f_y(y_{i,j}, \mathbf{u}_i \mid \boldsymbol{\theta}) f_{\mathbf{u}}(\mathbf{u}_i \mid \boldsymbol{\theta}) d\mathbf{u}_i$$
$$\text{MLE:} \quad \widehat{\boldsymbol{\theta}} = \arg\max \mathcal{L}(\boldsymbol{\theta} \mid \mathbf{y})$$

- We have $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e} \sim \text{MVN}(\mathbf{X}\boldsymbol{\beta}, \mathbf{Z}\boldsymbol{\Sigma}_u\mathbf{Z}^{\mathrm{T}} + \sigma^2\mathbf{I}) = \text{MVN}(\mathbf{X}\boldsymbol{\beta}, \mathbf{V})$.
- If the variance components $\mathbf{V}$ were known, we could estimate $\boldsymbol{\beta}$ by

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{y}.$$

- If there exists a $\mathbf{B}$, such that $\mathbf{B}\mathbf{X} = \mathbf{0}$, we have

$$\mathbf{B}\mathbf{y} = \mathbf{B}\mathbf{X}\boldsymbol{\beta} + \mathbf{B}\mathbf{Z}\mathbf{u} + \mathbf{B}\mathbf{e} = \mathbf{B}\mathbf{Z}\mathbf{u} + \mathbf{B}\mathbf{e} \sim \text{MVN}(\mathbf{0}, \mathbf{B}\mathbf{V}\mathbf{B}^{\mathrm{T}}).$$

- $(\mathbf{I} - \mathbf{H})$, with $\mathbf{H} = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}$, is a candidate for $\mathbf{B}$ since $(\mathbf{I} - \mathbf{H})\mathbf{X} = \mathbf{0}$.
- Use $\mathbf{B}\mathbf{y}$ to estimate $\mathbf{V}$ by some $\widehat{\mathbf{V}}_{REML}$.
- Estimate $\boldsymbol{\beta}$ then by $\widehat{\boldsymbol{\beta}}_{REML} = (\mathbf{X}^{\mathrm{T}}\widehat{\mathbf{V}}_{REML}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\widehat{\mathbf{V}}_{REML}^{-1}\mathbf{y}$.

$$\text{Likelihood:}\quad \mathcal{L}(\boldsymbol{\theta} \mid \mathbf{y}) = \log \prod_{i=1}^{n} \prod_{j=1}^{m_i} f_y(y_{i,j} \mid \boldsymbol{\theta})$$
$$= \log \prod_{i=1}^{n} \prod_{j=1}^{m_i} \int_{\mathbb{R}^q} f_y(y_{i,j}, \mathbf{u}_i \mid \boldsymbol{\theta}) f_{\mathbf{u}}(\mathbf{u}_i \mid \boldsymbol{\theta}) d\mathbf{u}_i$$
$$\text{MLE:}\quad \widehat{\boldsymbol{\theta}} = \arg\max \mathcal{L}(\boldsymbol{\theta} \mid \mathbf{y})$$

- We have $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e} \sim \text{MVN}(\mathbf{X}\boldsymbol{\beta}, \mathbf{Z}\boldsymbol{\Sigma}_u\mathbf{Z}^{\mathrm{T}} + \sigma^2\mathbf{I}) = \text{MVN}(\mathbf{X}\boldsymbol{\beta}, \mathbf{V})$.
- If the variance components $\mathbf{V}$ were known, we could estimate $\boldsymbol{\beta}$ by

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{y}.$$

- If there exists a $\mathbf{B}$, such that $\mathbf{B}\mathbf{X} = \mathbf{0}$, we have

$$\mathbf{B}\mathbf{y} = \mathbf{B}\mathbf{X}\boldsymbol{\beta} + \mathbf{B}\mathbf{Z}\mathbf{u} + \mathbf{B}\mathbf{e} = \mathbf{B}\mathbf{Z}\mathbf{u} + \mathbf{B}\mathbf{e} \sim \text{MVN}(\mathbf{0}, \mathbf{B}\mathbf{V}\mathbf{B}^{\mathrm{T}}).$$

- $(\mathbf{I} - \mathbf{H})$, with $\mathbf{H} = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}$, is a candidate for $\mathbf{B}$ since $(\mathbf{I} - \mathbf{H})\mathbf{X} = \mathbf{0}$.
- Use $\mathbf{B}\mathbf{y}$ to estimate $\mathbf{V}$ by some $\widehat{\mathbf{V}}_{REML}$.
- Estimate $\boldsymbol{\beta}$ then by $\widehat{\boldsymbol{\beta}}_{REML} = (\mathbf{X}^{\mathrm{T}}\widehat{\mathbf{V}}_{REML}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\widehat{\mathbf{V}}_{REML}^{-1}\mathbf{y}$.

Sleep Study Data

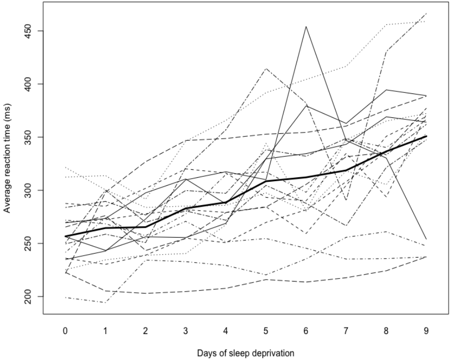
- $n = 18$ individuals, each measured $m_i = 10$ times.
- On day 0 the subjects had their normal amount of sleep. Starting that night, they were restricted to 3 hours of sleep per night.

Sleep Study Data

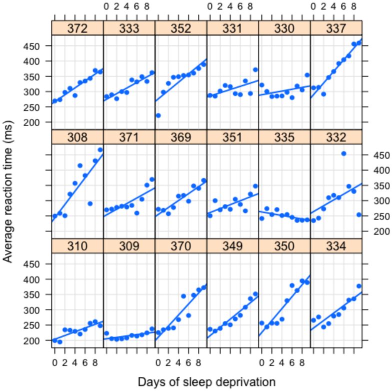
- $n = 18$ individuals, each measured $m_i = 10$ times.
- On day 0 the subjects had their normal amount of sleep. Starting that night, they were restricted to 3 hours of sleep per night.

Sleep Study Data

- $n = 18$ individuals, each measured $m_i = 10$ times.
- On day 0 the subjects had their normal amount of sleep. Starting that night, they were restricted to 3 hours of sleep per night.
- The figure below shows average reaction time versus days of sleep deprivation for different subjects super-imposed over their averages in solid bold.

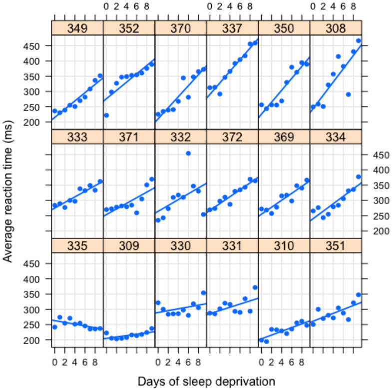


- $n = 18$ individuals, each measured $m_i = 10$ times.
- On day 0 the subjects had their normal amount of sleep. Starting that night, they were restricted to 3 hours of sleep per night.
- The figure below shows average reaction time versus days of sleep deprivation by subject. Subjects ordered (from bottom-left to top-right) by increasing intercept of subject specific linear regressions.



Sleep Study Data

- $n = 18$ individuals, each measured $m_i = 10$ times.
- On day 0 the subjects had their normal amount of sleep. Starting that night, they were restricted to 3 hours of sleep per night.



- The figure above shows average reaction time versus days of sleep deprivation by subject. Subjects ordered (from bottom-left to top-right) by increasing slope of subject specific linear regressions.

Linear Mixed Models with Penalized Fixed Effects

$$y_{i,j} = \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta} + \mathbf{z}_i^{\mathrm{T}} \mathbf{u}_i + \epsilon_{i,j}, \quad \mathbf{u}_i \overset{iid}{\sim} f_{\mathbf{u}}, \quad \epsilon_{i,j} \overset{iid}{\sim} f_{\epsilon},$$
$$i = 1, \ldots, n, \quad j = 1, \ldots, m_i$$
$$\text{with} \quad \mathbb{E}_{f_{\mathbf{u}}}(\mathbf{u}) = \mathbf{0} \quad \text{and} \quad \mathbb{E}_{f_{\epsilon}}(\epsilon) = 0.$$

$$\text{Likelihood:} \quad \mathcal{L}(\boldsymbol{\theta} \mid \mathbf{y}) = \log \prod_{i=1}^{n} \prod_{j=1}^{m_i} f_y(y_{i,j} \mid \boldsymbol{\theta})$$
$$= \log \prod_{i=1}^{n} \prod_{j=1}^{m_i} \int_{\mathbb{R}^q} f_y(y_{i,j}, \mathbf{u}_i \mid \boldsymbol{\theta}) f_{\mathbf{u}}(\mathbf{u}_i \mid \boldsymbol{\theta}) d\mathbf{u}_i$$
$$\text{Unrestricted MLE:} \quad \widehat{\boldsymbol{\theta}} = \arg\max \mathcal{L}(\boldsymbol{\theta} \mid \mathbf{y})$$

- Penalized likelihood estimates:

  $$\widehat{\boldsymbol{\beta}}_{ENET} = \arg\min[\mathcal{L}(\boldsymbol{\theta} \mid \mathbf{y}) + \lambda\{(1-\alpha) \|\boldsymbol{\beta}\|_2^2 + 2\alpha \|\boldsymbol{\beta}\|_1\}]$$

- Conceptually straightforward but computationally extremely challenging!

Linear Mixed Models with Penalized Fixed Effects

$$y_{i,j} = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \mathbf{z}_i^{\mathrm{T}}\mathbf{u}_i + \epsilon_{i,j}, \quad \mathbf{u}_i \overset{iid}{\sim} f_{\mathbf{u}}, \quad \epsilon_{i,j} \overset{iid}{\sim} f_{\epsilon},$$
$$i = 1, \ldots, n, \quad j = 1, \ldots, m_i$$
$$\text{with} \quad \mathbb{E}_{f_{\mathbf{u}}}(\mathbf{u}) = \mathbf{0} \quad \text{and} \quad \mathbb{E}_{f_{\epsilon}}(\epsilon) = 0.$$

$$\text{Likelihood:} \quad \mathcal{L}(\boldsymbol{\theta} \mid \mathbf{y}) = \log \prod_{i=1}^{n} \prod_{j=1}^{m_i} f_y(y_{i,j} \mid \boldsymbol{\theta})$$
$$= \log \prod_{i=1}^{n} \prod_{j=1}^{m_i} \int_{\mathbb{R}^q} f_y(y_{i,j}, \mathbf{u}_i \mid \boldsymbol{\theta}) f_{\mathbf{u}}(\mathbf{u}_i \mid \boldsymbol{\theta}) d\mathbf{u}_i$$
$$\text{Unrestricted MLE:} \quad \widehat{\boldsymbol{\theta}} = \arg\max \mathcal{L}(\boldsymbol{\theta} \mid \mathbf{y})$$

- Penalized likelihood estimates:

  $$\widehat{\boldsymbol{\beta}}_{ENET} = \arg\min[\mathcal{L}(\boldsymbol{\theta} \mid \mathbf{y}) + \lambda\{(1-\alpha)\left\|\boldsymbol{\beta}\right\|_2^2 + 2\alpha\left\|\boldsymbol{\beta}\right\|_1\}]$$
- Conceptually straightforward but computationally extremely challenging!

$$y_{i,j} = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \mathbf{z}_i^{\mathrm{T}}\mathbf{u}_i + \epsilon_{i,j}, \quad \mathbf{u}_i \overset{iid}{\sim} \mathrm{MVN}(\mathbf{0}, \boldsymbol{\Sigma}_u), \quad \epsilon_{i,j} \overset{iid}{\sim} \mathrm{Normal}(0, \sigma_\epsilon^2),$$
$$i = 1, \ldots, n, \quad j = 1, \ldots, m_i, \quad N = \textstyle\sum_{i=1}^n m_i.$$

- Conditional likelihood:

$$L(\boldsymbol{\beta}, \sigma_\epsilon^2 \mid \mathbf{u}_{1:n}) = \frac{1}{(2\pi\sigma_\epsilon^2)^{N/2}} \exp\left\{-\frac{1}{2\sigma_\epsilon^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u})^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u})\right\}$$
$$\propto \frac{1}{(\sigma_\epsilon^2)^{N/2}} \exp\left[-\frac{1}{2\sigma_\epsilon^2}\left\{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{\beta} - 2\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}(\mathbf{y} - \mathbf{Z}\mathbf{u}) + (\mathbf{y} - \mathbf{Z}\mathbf{u})^{\mathrm{T}}(\mathbf{y} - \mathbf{Z}\mathbf{u})\right\}\right]$$
$$\propto \frac{1}{(\sigma_\epsilon^2)^{N/2}} \exp\left\{-\frac{1}{2\sigma_\epsilon^2}\left(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{\beta} - 2\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\widetilde{\mathbf{y}} + \widetilde{\mathbf{y}}^{\mathrm{T}}\widetilde{\mathbf{y}}\right)\right\} \text{ with } \widetilde{\mathbf{y}} = \mathbf{y} - \mathbf{Z}\mathbf{u}$$

- Semi-conjugate priors on $(\boldsymbol{\beta}, \sigma_\epsilon^2)$: $p(\boldsymbol{\beta}, \sigma_\epsilon^2) = \mathrm{MVN}(\boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta) \cdot \text{Inv-Ga}(a_\sigma, b_\sigma)$

$$\propto \frac{1}{(\sigma_\epsilon^2)^{a_\sigma+1}} \exp\left(-\frac{b_\sigma}{\sigma_\epsilon^2}\right) \exp\left\{-\frac{1}{2}(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\Sigma}_\beta^{-1}\boldsymbol{\beta} - 2\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\Sigma}_\beta^{-1}\boldsymbol{\mu}_\beta + \boldsymbol{\mu}_\beta^{\mathrm{T}}\boldsymbol{\Sigma}_\beta^{-1}\boldsymbol{\mu}_\beta)\right\}$$

- Posterior full conditionals of $\boldsymbol{\beta}, \sigma_\epsilon^2$:

- Conjugate prior on $\boldsymbol{\Sigma}_u$:

$p(\boldsymbol{\Sigma}_u) = \mathrm{IW}(\boldsymbol{\Sigma}_u \mid \nu_0, \boldsymbol{\Sigma}_0) \propto |\boldsymbol{\Sigma}_u|^{-\left(\frac{\nu_0+d+1}{2}+\frac{1}{2}\right)} \exp\left[-\frac{1}{2}\left\{\mathrm{trace}(\boldsymbol{\Sigma}_u^{-1}\boldsymbol{\Sigma}_0)\right\}\right]$

- Posterior full conditional of $\boldsymbol{\Sigma}_u$:

$$y_{i,j} = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \mathbf{z}_i^{\mathrm{T}}\mathbf{u}_i + \epsilon_{i,j}, \quad \mathbf{u}_i \overset{iid}{\sim} \mathrm{MVN}(\mathbf{0}, \boldsymbol{\Sigma}_u), \quad \epsilon_{i,j} \overset{iid}{\sim} \mathrm{Normal}(0, \sigma_\epsilon^2),$$
$$i = 1, \ldots, n, \quad j = 1, \ldots, m_i, \quad N = \textstyle\sum_{i=1}^n m_i.$$

- Conditional likelihood:
  $$L(\boldsymbol{\beta}, \sigma_\epsilon^2 \mid \mathbf{u}_{1:n}) \quad \propto \frac{1}{(\sigma_\epsilon^2)^{N/2}} \exp\left\{-\frac{1}{2\sigma_\epsilon^2}\left(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{\beta} - 2\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\widetilde{\mathbf{y}} + \widetilde{\mathbf{y}}^{\mathrm{T}}\widetilde{\mathbf{y}}\right)\right\} \text{ with } \widetilde{\mathbf{y}} = \mathbf{y} - \mathbf{Zu}$$
- Semi-conjugate priors on $(\boldsymbol{\beta}, \sigma_\epsilon^2)$: $p(\boldsymbol{\beta}, \sigma_\epsilon^2) = \mathrm{MVN}(\boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta) \cdot \text{Inv-Ga}(a_\sigma, b_\sigma)$
  $$\propto \frac{1}{(\sigma_\epsilon^2)^{a_\sigma+1}} \exp\left(-\frac{b_\sigma}{\sigma_\epsilon^2}\right) \exp\left\{-\frac{1}{2}(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\Sigma}_\beta^{-1}\boldsymbol{\beta} - 2\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\Sigma}_\beta^{-1}\boldsymbol{\mu}_\beta + \boldsymbol{\mu}_\beta^{\mathrm{T}}\boldsymbol{\Sigma}_\beta^{-1}\boldsymbol{\mu}_\beta)\right\}$$
- Posterior full conditionals of β, σ_ϵ^2:
- Conjugate prior on $\boldsymbol{\Sigma}_u$:
  $$p(\boldsymbol{\Sigma}_u) = \mathrm{IW}\left(\boldsymbol{\Sigma}_u \mid \nu_0, \boldsymbol{\Sigma}_0\right) \propto |\boldsymbol{\Sigma}_u|^{-\left(\frac{\nu_0+d+1}{2}+\frac{1}{2}\right)} \exp\left[-\tfrac{1}{2}\left\{\mathrm{trace}(\boldsymbol{\Sigma}_u^{-1}\boldsymbol{\Sigma}_0)\right\}\right]$$
- Posterior full conditional of $\boldsymbol{\Sigma}_u$:
- Posterior full conditional of $\mathbf{u}_i$:

Bayesian Linear Mixed Models

$$y_{i,j} = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \mathbf{z}_i^{\mathrm{T}}\mathbf{u}_i + \epsilon_{i,j}, \quad \mathbf{u}_i \overset{iid}{\sim} \mathrm{MVN}(\mathbf{0}, \boldsymbol{\Sigma}_u), \quad \epsilon_{i,j} \overset{iid}{\sim} \mathrm{Normal}(0, \sigma_\epsilon^2),$$
$$i = 1, \ldots, n, \quad j = 1, \ldots, m_i, \quad N = \textstyle\sum_{i=1}^n m_i.$$

- Conditional likelihood:
  $$L(\boldsymbol{\beta}, \sigma_\epsilon^2 \mid \mathbf{u}_{1:n}) \quad \propto \frac{1}{(\sigma_\epsilon^2)^{N/2}} \exp\left\{-\frac{1}{2\sigma_\epsilon^2}\left(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{\beta} - 2\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\widetilde{\mathbf{y}} + \widetilde{\mathbf{y}}^{\mathrm{T}}\widetilde{\mathbf{y}}\right)\right\} \text{ with } \widetilde{\mathbf{y}} = \mathbf{y} - \mathbf{Z}\mathbf{u}$$
- Semi-conjugate priors on $(\boldsymbol{\beta}, \sigma_\epsilon^2)$: $p(\boldsymbol{\beta}, \sigma_\epsilon^2) = \mathrm{MVN}(\boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta) \cdot \text{Inv-Ga}(a_\sigma, b_\sigma)$
  $$\propto \frac{1}{(\sigma_\epsilon^2)^{a_\sigma+1}} \exp\left(-\frac{b_\sigma}{\sigma_\epsilon^2}\right) \exp\left\{-\frac{1}{2}(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\Sigma}_\beta^{-1}\boldsymbol{\beta} - 2\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\Sigma}_\beta^{-1}\boldsymbol{\mu}_\beta + \boldsymbol{\mu}_\beta^{\mathrm{T}}\boldsymbol{\Sigma}_\beta^{-1}\boldsymbol{\mu}_\beta)\right\}$$
- Posterior full conditionals of $\boldsymbol{\beta}, \sigma_\epsilon^2$:
  - $p(\boldsymbol{\beta} \mid -) \propto \exp\left[-\frac{1}{2}\left\{\boldsymbol{\beta}^{\mathrm{T}}(\boldsymbol{\Sigma}_\beta^{-1} + \sigma_\epsilon^{-2}\mathbf{X}^{\mathrm{T}}\mathbf{X})\boldsymbol{\beta} - 2\boldsymbol{\beta}^{\mathrm{T}}\left(\boldsymbol{\Sigma}_\beta^{-1}\boldsymbol{\mu}_\beta + \sigma_\epsilon^{-2}\mathbf{X}^{\mathrm{T}}\widetilde{\mathbf{y}}\right)\right\}\right]$
    $$\equiv \mathrm{MVN}(\boldsymbol{\mu}_{\beta,N}, \boldsymbol{\Sigma}_{\beta,N}),$$
    $$\boldsymbol{\Sigma}_{\beta,N} = (\boldsymbol{\Sigma}_\beta^{-1} + \sigma_\epsilon^{-2}\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}, \quad \boldsymbol{\mu}_{\beta,N} = \boldsymbol{\Sigma}_{\beta,N}\left(\boldsymbol{\Sigma}_\beta^{-1}\boldsymbol{\mu}_\beta + \sigma_\epsilon^{-2}\mathbf{X}^{\mathrm{T}}\widetilde{\mathbf{y}}\right)$$
  - $p(\sigma_\epsilon^2 \mid -) \propto \frac{1}{(\sigma_\epsilon^2)^{a_\sigma+\frac{N}{2}+1}} \exp\left[-\frac{1}{\sigma_\epsilon^2}\left\{b_\sigma + \frac{1}{2}(\widetilde{\mathbf{y}} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\widetilde{\mathbf{y}} - \mathbf{X}\boldsymbol{\beta})\right\}\right]$
    $$\equiv \text{Inv-Ga}\left\{a_\sigma + \frac{N}{2}, b_\sigma + \frac{1}{2}(\widetilde{\mathbf{y}} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\widetilde{\mathbf{y}} - \mathbf{X}\boldsymbol{\beta})\right\}$$
- Conjugate prior on $\boldsymbol{\Sigma}_u$:
  $p(\boldsymbol{\Sigma}_u) = \mathrm{IW}(\boldsymbol{\Sigma}_u \mid \nu_0, \boldsymbol{\Sigma}_0) \propto |\boldsymbol{\Sigma}_u|^{-\left(\frac{\nu_0+d+1}{2}+\frac{1}{2}\right)} \exp\left[-\frac{1}{2}\left\{\mathrm{trace}(\boldsymbol{\Sigma}_u^{-1}\boldsymbol{\Sigma}_0)\right\}\right]$
- Posterior full conditional of $\boldsymbol{\Sigma}_u$:

$$y_{i,j} = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \mathbf{z}_i^{\mathrm{T}}\mathbf{u}_i + \epsilon_{i,j}, \quad \mathbf{u}_i \overset{iid}{\sim} \mathrm{MVN}(\mathbf{0}, \boldsymbol{\Sigma}_u), \quad \epsilon_{i,j} \overset{iid}{\sim} \mathrm{Normal}(0, \sigma_\epsilon^2),$$
$$i = 1, \ldots, n, \quad j = 1, \ldots, m_i, \quad N = \textstyle\sum_{i=1}^n m_i.$$

- Conditional likelihood:
  $$L(\boldsymbol{\beta}, \sigma_\epsilon^2 \mid \mathbf{u}_{1:n}) \quad \propto \frac{1}{(\sigma_\epsilon^2)^{N/2}} \exp\left\{-\frac{1}{2\sigma_\epsilon^2}\left(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{\beta} - 2\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\widetilde{\mathbf{y}} + \widetilde{\mathbf{y}}^{\mathrm{T}}\widetilde{\mathbf{y}}\right)\right\} \text{ with } \widetilde{\mathbf{y}} = \mathbf{y} - \mathbf{Z}\mathbf{u}$$
- Semi-conjugate priors on $(\boldsymbol{\beta}, \sigma_\epsilon^2)$: $p(\boldsymbol{\beta}, \sigma_\epsilon^2) = \mathrm{MVN}(\boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta) \cdot \text{Inv-Ga}(a_\sigma, b_\sigma)$
- Posterior full conditionals of $\boldsymbol{\beta}, \sigma_\epsilon^2$:
  - $p(\boldsymbol{\beta} \mid -) = \mathrm{MVN}(\boldsymbol{\mu}_{\beta,N}, \boldsymbol{\Sigma}_{\beta,N})$,
    $$\boldsymbol{\Sigma}_{\beta,N} = (\boldsymbol{\Sigma}_\beta^{-1} + \sigma_\epsilon^{-2}\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}, \quad \boldsymbol{\mu}_{\beta,N} = \boldsymbol{\Sigma}_{\beta,N}\left(\boldsymbol{\Sigma}_\beta^{-1}\boldsymbol{\mu}_\beta + \sigma_\epsilon^{-2}\mathbf{X}^{\mathrm{T}}\widetilde{\mathbf{y}}\right)$$
  - $p(\sigma_\epsilon^2 \mid -) = \text{Inv-Ga}\left\{a_\sigma + \frac{N}{2}, b_\sigma + \frac{1}{2}(\widetilde{\mathbf{y}} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\widetilde{\mathbf{y}} - \mathbf{X}\boldsymbol{\beta})\right\}$
- Conjugate prior on $\boldsymbol{\Sigma}_u$:
  $p(\boldsymbol{\Sigma}_u) = \mathrm{IW}(\boldsymbol{\Sigma}_u \mid \nu_0, \boldsymbol{\Sigma}_0) \propto |\boldsymbol{\Sigma}_u|^{-\left(\frac{\nu_0+d+1}{2}+\frac{1}{2}\right)} \exp\left[-\frac{1}{2}\left\{\mathrm{trace}(\boldsymbol{\Sigma}_u^{-1}\boldsymbol{\Sigma}_0)\right\}\right]$
- Posterior full conditional of $\boldsymbol{\Sigma}_u$:
  - $p(\boldsymbol{\Sigma}_u \mid -) = \mathrm{IW}(n + \nu_0, \boldsymbol{\Sigma}_0 + \sum_{i=1}^n \mathbf{u}_i\mathbf{u}_i^{\mathrm{T}})$
- Posterior full conditional of $\mathbf{u}_i$:
  - $p(\mathbf{u}_i \mid -)$

Bayesian Linear Mixed Models

$$y_{i,j} = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \mathbf{z}_i^{\mathrm{T}}\mathbf{u}_i + \epsilon_{i,j}, \quad \mathbf{u}_i \overset{iid}{\sim} \mathrm{MVN}(\mathbf{0}, \boldsymbol{\Sigma}_u), \quad \epsilon_{i,j} \overset{iid}{\sim} \mathrm{Normal}(0, \sigma_\epsilon^2),$$
$$i = 1, \ldots, n, \quad j = 1, \ldots, m_i, \quad N = \textstyle\sum_{i=1}^n m_i.$$

- Conditional likelihood:
  $$L(\boldsymbol{\beta}, \sigma_\epsilon^2 \mid \mathbf{u}_{1:n}) \quad \propto \frac{1}{(\sigma_\epsilon^2)^{N/2}} \exp\left\{-\frac{1}{2\sigma_\epsilon^2}\left(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{\beta} - 2\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\widetilde{\mathbf{y}} + \widetilde{\mathbf{y}}^{\mathrm{T}}\widetilde{\mathbf{y}}\right)\right\} \text{ with } \widetilde{\mathbf{y}} = \mathbf{y} - \mathbf{Z}\mathbf{u}$$
- Semi-conjugate priors on $(\boldsymbol{\beta}, \sigma_\epsilon^2)$: $p(\boldsymbol{\beta}, \sigma_\epsilon^2) = \mathrm{MVN}(\boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta) \cdot \text{Inv-Ga}(a_\sigma, b_\sigma)$
- Posterior full conditionals of $\boldsymbol{\beta}, \sigma_\epsilon^2$:
  - $p(\boldsymbol{\beta} \mid -) = \mathrm{MVN}(\boldsymbol{\mu}_{\beta,N}, \boldsymbol{\Sigma}_{\beta,N})$,
    $$\boldsymbol{\Sigma}_{\beta,N} = (\boldsymbol{\Sigma}_\beta^{-1} + \sigma_\epsilon^{-2}\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}, \quad \boldsymbol{\mu}_{\beta,N} = \boldsymbol{\Sigma}_{\beta,N}\left(\boldsymbol{\Sigma}_\beta^{-1}\boldsymbol{\mu}_\beta + \sigma_\epsilon^{-2}\mathbf{X}^{\mathrm{T}}\widetilde{\mathbf{y}}\right)$$
  - $p(\sigma_\epsilon^2 \mid -) = \text{Inv-Ga}\left\{a_\sigma + \frac{N}{2}, b_\sigma + \frac{1}{2}(\widetilde{\mathbf{y}} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\widetilde{\mathbf{y}} - \mathbf{X}\boldsymbol{\beta})\right\}$
- Conjugate prior on $\boldsymbol{\Sigma}_u$:
  $$p(\boldsymbol{\Sigma}_u) = \mathrm{IW}(\boldsymbol{\Sigma}_u \mid \nu_0, \boldsymbol{\Sigma}_0) \propto |\boldsymbol{\Sigma}_u|^{-\left(\frac{\nu_0+d+1}{2}+\frac{1}{2}\right)} \exp\left[-\tfrac{1}{2}\left\{\mathrm{trace}(\boldsymbol{\Sigma}_u^{-1}\boldsymbol{\Sigma}_0)\right\}\right]$$
- Posterior full conditional of $\boldsymbol{\Sigma}_u$:
  - $p(\boldsymbol{\Sigma}_u \mid -) = \mathrm{IW}(n + \nu_0, \boldsymbol{\Sigma}_0 + \sum_{i=1}^n \mathbf{u}_i\mathbf{u}_i^{\mathrm{T}})$
- Posterior full conditional of $\mathbf{u}_i$:
  - $p(\mathbf{u}_i \mid -)$

$$y_{i,j} = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \mathbf{z}_i^{\mathrm{T}}\mathbf{u}_i + \epsilon_{i,j}, \quad \mathbf{u}_i \overset{iid}{\sim} \mathrm{MVN}(\mathbf{0}, \boldsymbol{\Sigma}_u), \quad \epsilon_{i,j} \overset{iid}{\sim} \mathrm{Normal}(0, \sigma_\epsilon^2),$$
$$i = 1, \ldots, n, \quad j = 1, \ldots, m_i, \quad N = \textstyle\sum_{i=1}^n m_i.$$

- Conditional likelihood:
$L(\boldsymbol{\beta}, \sigma_\epsilon^2 \mid \mathbf{u}_{1:n}) \quad \propto \frac{1}{(\sigma_\epsilon^2)^{N/2}} \exp\left\{-\frac{1}{2\sigma_\epsilon^2}\left(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{\beta} - 2\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\widetilde{\mathbf{y}} + \widetilde{\mathbf{y}}^{\mathrm{T}}\widetilde{\mathbf{y}}\right)\right\}$ with $\widetilde{\mathbf{y}} = \mathbf{y} - \mathbf{Z}\mathbf{u}$
- Semi-conjugate priors on $(\boldsymbol{\beta}, \sigma_\epsilon^2)$: $p(\boldsymbol{\beta}, \sigma_\epsilon^2) = \mathrm{MVN}(\boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta) \cdot \text{Inv-Ga}(a_\sigma, b_\sigma)$
- Posterior full conditionals of $\boldsymbol{\beta}, \sigma_\epsilon^2$:
  - $p(\boldsymbol{\beta} \mid -) = \mathrm{MVN}(\boldsymbol{\mu}_{\beta,N}, \boldsymbol{\Sigma}_{\beta,N})$,
  $$\boldsymbol{\Sigma}_{\beta,N} = (\boldsymbol{\Sigma}_\beta^{-1} + \sigma_\epsilon^{-2}\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}, \quad \boldsymbol{\mu}_{\beta,N} = \boldsymbol{\Sigma}_{\beta,N}\left(\boldsymbol{\Sigma}_\beta^{-1}\boldsymbol{\mu}_\beta + \sigma_\epsilon^{-2}\mathbf{X}^{\mathrm{T}}\widetilde{\mathbf{y}}\right)$$
  - $p(\sigma_\epsilon^2 \mid -) = \text{Inv-Ga}\left\{a_\sigma + \frac{N}{2}, b_\sigma + \frac{1}{2}(\widetilde{\mathbf{y}} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\widetilde{\mathbf{y}} - \mathbf{X}\boldsymbol{\beta})\right\}$
- Conjugate prior on $\boldsymbol{\Sigma}_u$:
$p(\boldsymbol{\Sigma}_u) = \mathrm{IW}(\boldsymbol{\Sigma}_u \mid \nu_0, \boldsymbol{\Sigma}_0) \propto |\boldsymbol{\Sigma}_u|^{-\left(\frac{\nu_0+d+1}{2}+\frac{1}{2}\right)} \exp\left[-\frac{1}{2}\left\{\mathrm{trace}(\boldsymbol{\Sigma}_u^{-1}\boldsymbol{\Sigma}_0)\right\}\right]$
- Posterior full conditional of $\boldsymbol{\Sigma}_u$:
  - $p(\boldsymbol{\Sigma}_u \mid -) = \mathrm{IW}(n + \nu_0, \boldsymbol{\Sigma}_0 + \sum_{i=1}^n \mathbf{u}_i\mathbf{u}_i^{\mathrm{T}})$
- Posterior full conditional of $\mathbf{u}_i$:
  - $p(\mathbf{u}_i \mid -) \propto \exp\left[-\frac{1}{2}\left\{\mathbf{u}_i^{\mathrm{T}}\boldsymbol{\Sigma}_u^{-1}\mathbf{u}_i + \sigma_\epsilon^{-2}\sum_{j=1}^{m_i}(y_{i,j} - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} - \mathbf{z}_i^{\mathrm{T}}\mathbf{u}_i)^2\right\}\right]$
  $\propto \exp\left[-\frac{1}{2}\left\{\mathbf{u}_i^{\mathrm{T}}(\boldsymbol{\Sigma}_u^{-1} + \sigma_\epsilon^{-2}m_i\mathbf{z}_i\mathbf{z}_i^{\mathrm{T}})\mathbf{u}_i - 2\sigma_\epsilon^{-2}\sum_{j=1}^{m_i}(y_{i,j} - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})\mathbf{z}_i^{\mathrm{T}}\mathbf{u}_i\right\}\right]$

Bayesian Linear Mixed Models

$$y_{i,j} = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \mathbf{z}_i^{\mathrm{T}}\mathbf{u}_i + \epsilon_{i,j}, \quad \mathbf{u}_i \overset{iid}{\sim} \mathrm{MVN}(\mathbf{0}, \boldsymbol{\Sigma}_u), \quad \epsilon_{i,j} \overset{iid}{\sim} \mathrm{Normal}(0, \sigma_\epsilon^2),$$
$$i = 1, \ldots, n, \quad j = 1, \ldots, m_i, \quad N = \textstyle\sum_{i=1}^n m_i.$$

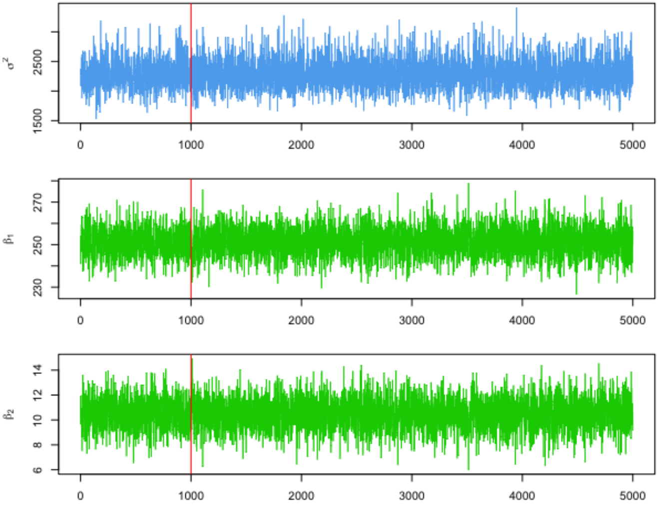
- Conditional likelihood:
  $L(\boldsymbol{\beta}, \sigma_\epsilon^2 \mid \mathbf{u}_{1:n}) \quad \propto \frac{1}{(\sigma_\epsilon^2)^{N/2}} \exp\left\{-\frac{1}{2\sigma_\epsilon^2}\left(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{\beta} - 2\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\widetilde{\mathbf{y}} + \widetilde{\mathbf{y}}^{\mathrm{T}}\widetilde{\mathbf{y}}\right)\right\}$ with $\widetilde{\mathbf{y}} = \mathbf{y} - \mathbf{Z}\mathbf{u}$
- Semi-conjugate priors on $(\boldsymbol{\beta}, \sigma_\epsilon^2)$: $p(\boldsymbol{\beta}, \sigma_\epsilon^2) = \mathrm{MVN}(\boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta) \cdot \text{Inv-Ga}(a_\sigma, b_\sigma)$
- Posterior full conditionals of $\boldsymbol{\beta}, \sigma_\epsilon^2$:
  - $p(\boldsymbol{\beta} \mid -) = \mathrm{MVN}(\boldsymbol{\mu}_{\beta,N}, \boldsymbol{\Sigma}_{\beta,N})$,

    $$\boldsymbol{\Sigma}_{\beta,N} = (\boldsymbol{\Sigma}_\beta^{-1} + \sigma_\epsilon^{-2}\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}, \quad \boldsymbol{\mu}_{\beta,N} = \boldsymbol{\Sigma}_{\beta,N}\left(\boldsymbol{\Sigma}_\beta^{-1}\boldsymbol{\mu}_\beta + \sigma_\epsilon^{-2}\mathbf{X}^{\mathrm{T}}\widetilde{\mathbf{y}}\right)$$
  - $p(\sigma_\epsilon^2 \mid -) = \text{Inv-Ga}\left\{a_\sigma + \frac{N}{2}, b_\sigma + \frac{1}{2}(\widetilde{\mathbf{y}} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\widetilde{\mathbf{y}} - \mathbf{X}\boldsymbol{\beta})\right\}$
- Conjugate prior on $\boldsymbol{\Sigma}_u$:
  $p(\boldsymbol{\Sigma}_u) = \mathrm{IW}\left(\boldsymbol{\Sigma}_u \mid \nu_0, \boldsymbol{\Sigma}_0\right) \propto |\boldsymbol{\Sigma}_u|^{-\left(\frac{\nu_0+d+1}{2}+\frac{1}{2}\right)} \exp\left[-\frac{1}{2}\left\{\mathrm{trace}(\boldsymbol{\Sigma}_u^{-1}\boldsymbol{\Sigma}_0)\right\}\right]$
- Posterior full conditional of $\boldsymbol{\Sigma}_u$:
  - $p(\boldsymbol{\Sigma}_u \mid -) = \mathrm{IW}(n + \nu_0, \boldsymbol{\Sigma}_0 + \sum_{i=1}^n \mathbf{u}_i\mathbf{u}_i^{\mathrm{T}})$
- Posterior full conditional of $\mathbf{u}_i$:
  - $p(\mathbf{u}_i \mid -) = \mathrm{MVN}(\boldsymbol{\mu}_{i,u,N}, \boldsymbol{\Sigma}_{i,u,N})$,

$$\boldsymbol{\Sigma}_{i,u,N} = (\boldsymbol{\Sigma}_u^{-1} + \sigma_\epsilon^{-2}m_i\mathbf{z}_i\mathbf{z}_i^{\mathrm{T}})^{-1}, \quad \boldsymbol{\mu}_{i,u,N} = \boldsymbol{\Sigma}_{i,u,N}\left\{\sigma_\epsilon^{-2}\textstyle\sum_{j=1}^{m_i}(y_{i,j} - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})\mathbf{z}_i\right\}$$

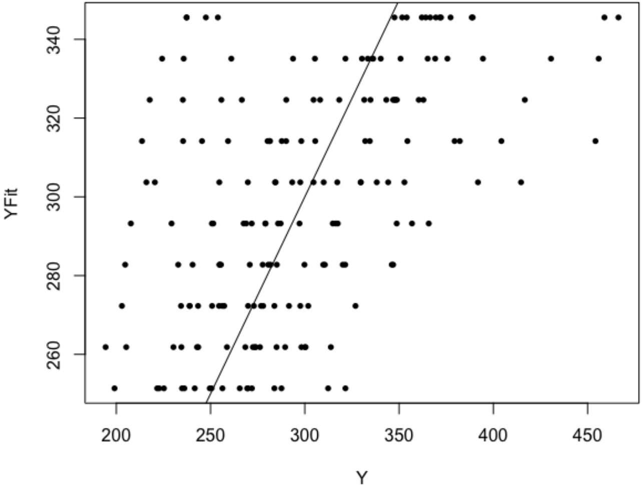
Sleep Study Data - Bayesian LM - Fixed Effects Only

$$\text{ReactionTime}_{i,j} = \beta_0 + \beta_1 \text{Day}_j + \epsilon_{i,j}, \quad \epsilon_{i,j} \overset{iid}{\sim} \text{Normal}(0, \sigma_\epsilon^2),$$
$$i = 1, \ldots, n = 18, \quad j = 1, \ldots, m_i = 10, \quad N = \textstyle\sum_{i=1}^n m_i = 180.$$



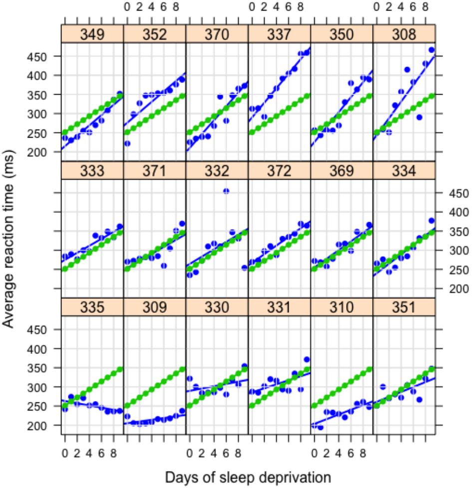
Sleep Study Data - Bayesian LM - Fixed Effects Only

$$\text{ReactionTime}_{i,j} = \beta_0 + \beta_1 \text{Day}_j + \epsilon_{i,j}, \quad \epsilon_{i,j} \overset{iid}{\sim} \text{Normal}(0, \sigma_\epsilon^2),$$
$$i = 1, \ldots, n = 18, \quad j = 1, \ldots, m_i = 10, \quad N = \sum_{i=1}^{n} m_i = 180.$$

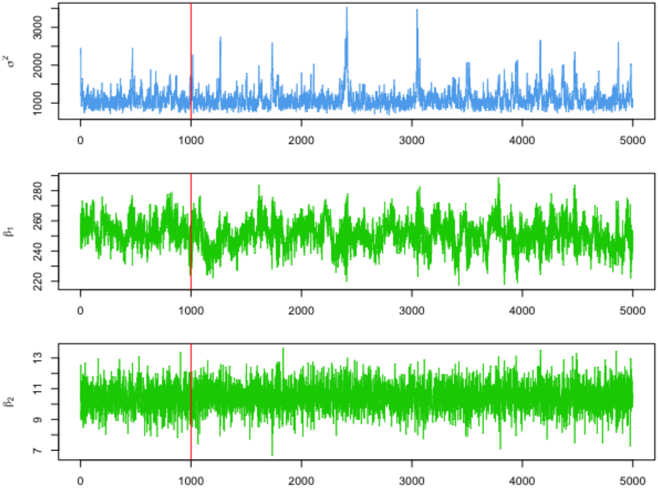


Sleep Study Data - Bayesian LM - Fixed Effects Only

$$\text{ReactionTime}_{i,j} = \beta_0 + \beta_1 \text{Day}_j + \epsilon_{i,j}, \quad \epsilon_{i,j} \overset{iid}{\sim} \text{Normal}(0, \sigma_\epsilon^2),$$
$$i = 1, \ldots, n = 18, \quad j = 1, \ldots, m_i = 10, \quad N = \textstyle\sum_{i=1}^{n} m_i = 180.$$

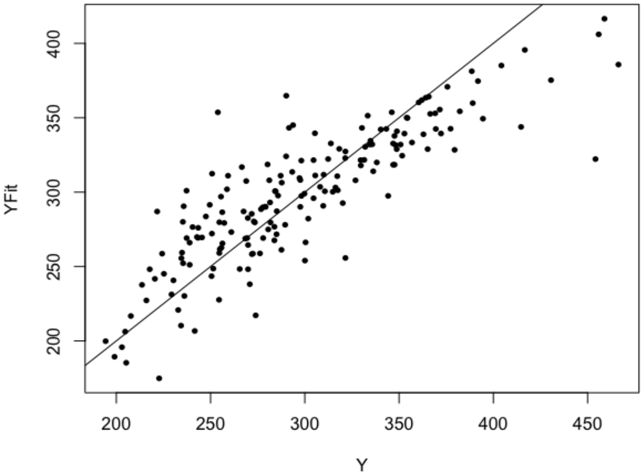


$$\text{ReactionTime}_{i,j} = \beta_0 + \beta_1 \text{Day}_j + u_i + \epsilon_{i,j}, \quad u_i \sim \text{Normal}(\sigma_u^2), \quad \epsilon_{i,j} \overset{iid}{\sim} \text{Normal}(0, \sigma_\epsilon^2),$$
$$i = 1, \ldots, n = 18, \quad j = 1, \ldots, m_i = 10, \quad N = \sum_{i=1}^{n} m_i = 180.$$



Sleep Study Data - Bayesian LMM - Random Intercept Only

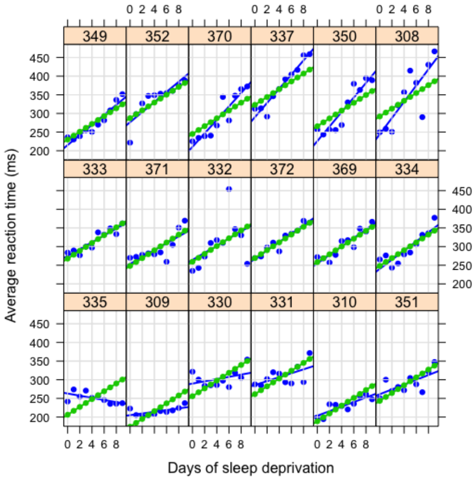
$$\text{ReactionTime}_{i,j} = \beta_0 + \beta_1 \text{Day}_j + u_i + \epsilon_{i,j}, \quad u_i \sim \text{Normal}(\sigma_u^2), \quad \epsilon_{i,j} \overset{iid}{\sim} \text{Normal}(0, \sigma_\epsilon^2),$$
$$i = 1, \ldots, n = 18, \quad j = 1, \ldots, m_i = 10, \quad N = \sum_{i=1}^{n} m_i = 180.$$



Sleep Study Data - Bayesian LMM - Random Intercept Only

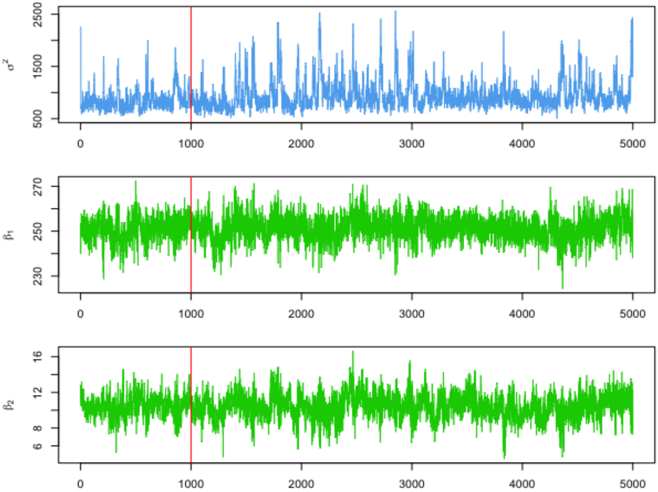
$$\text{ReactionTime}_{i,j} = \beta_0 + \beta_1 \text{Day}_j + u_i + \epsilon_{i,j}, \quad u_i \sim \text{Normal}(\sigma_u^2), \quad \epsilon_{i,j} \overset{iid}{\sim} \text{Normal}(0, \sigma_\epsilon^2),$$

$$i = 1, \ldots, n = 18, \quad j = 1, \ldots, m_i = 10, \quad N = \sum_{i=1}^{n} m_i = 180.$$



$$\text{ReactionTime}_{i,j} = \beta_0 + \beta_1 \text{Day}_j + u_{0,i} + u_{1,i}\text{Day}_j + \epsilon_{i,j}, \quad u_i \sim \text{Normal}(\sigma_u^2),$$

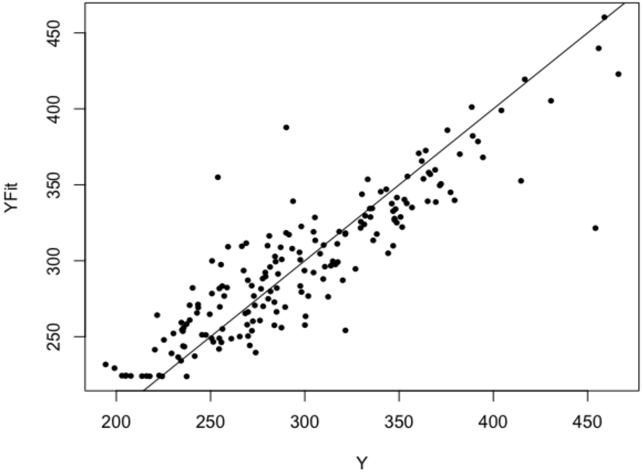
$$\epsilon_{i,j} \overset{iid}{\sim} \text{Normal}(0, \sigma_\epsilon^2), \quad i = 1, \ldots, n = 18, \quad j = 1, \ldots, m_i = 10, \quad N = \sum_{i=1}^n m_i = 180.$$



Sleep Study Data - Bayesian LMM - Random Intercept and Slope

$$\text{ReactionTime}_{i,j} = \beta_0 + \beta_1 \text{Day}_j + u_{0,i} + u_{1,i}\text{Day}_j + \epsilon_{i,j}, \quad u_i \sim \text{Normal}(\sigma_u^2),$$

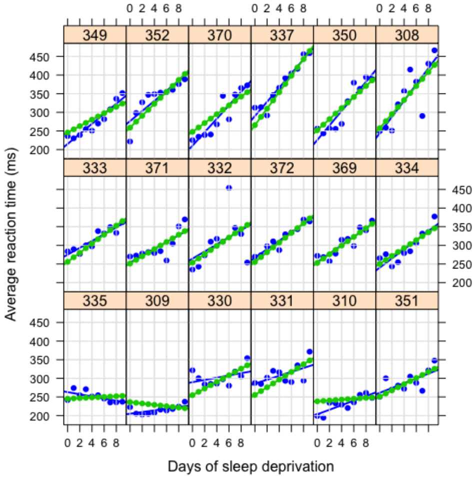
$$\epsilon_{i,j} \overset{iid}{\sim} \text{Normal}(0, \sigma_\epsilon^2), \quad i = 1, \ldots, n = 18, \quad j = 1, \ldots, m_i = 10, \quad N = \sum_{i=1}^{n} m_i = 180.$$



Sleep Study Data - Bayesian LMM - Random Intercept and Slope

$$\text{ReactionTime}_{i,j} = \beta_0 + \beta_1 \text{Day}_j + u_{0,i} + u_{1,i}\text{Day}_j + \epsilon_{i,j}, \quad u_i \sim \text{Normal}(\sigma_u^2),$$

$$\epsilon_{i,j} \overset{iid}{\sim} \text{Normal}(0, \sigma_\epsilon^2), \quad i = 1, \ldots, n = 18, \quad j = 1, \ldots, m_i = 10, \quad N = \sum_{i=1}^n m_i = 180.$$



Bayesian Ridge for Linear Mixed Models

$$y_{i,j} = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \mathbf{z}_i^{\mathrm{T}}\mathbf{u}_i + \epsilon_{i,j}, \quad \mathbf{u}_i \overset{iid}{\sim} \mathrm{MVN}(\mathbf{0}, \boldsymbol{\Sigma}_u), \quad \epsilon_{i,j} \overset{iid}{\sim} \mathrm{Normal}(0, \sigma_\epsilon^2),$$
$$i = 1, \ldots, n, \quad j = 1, \ldots, m_i, \quad N = \sum_{i=1}^{n} m_i.$$

- Semi-conjugate priors on $(\boldsymbol{\beta}, \sigma_\epsilon^2)$:
  $p(\boldsymbol{\beta}, \sigma_\epsilon^2) = \mathrm{MVN}(\boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta) \cdot \text{Inv-Ga}(a_\sigma, b_\sigma)$ with $\boldsymbol{\Sigma}_\beta = \lambda^{-1}\sigma_\epsilon^2\mathbf{I}_p$
- Posterior full conditionals of β, σ_ϵ^2:
- Conjugate prior on λ: $p(\lambda) = \mathrm{Ga}(a_\lambda, b_\lambda)$
- Posterior full conditional of λ:
- Conjugate prior on $\boldsymbol{\Sigma}_u$: $p(\boldsymbol{\Sigma}_u) = \mathrm{IW}(\boldsymbol{\Sigma}_u \mid \nu_0, \boldsymbol{\Sigma}_0)$
- Posterior full conditional of $\boldsymbol{\Sigma}_u$:
- Posterior full conditional of $\mathbf{u}_i$:

$$y_{i,j} = \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta} + \mathbf{z}_i^{\mathrm{T}} \mathbf{u}_i + \epsilon_{i,j}, \quad \mathbf{u}_i \overset{iid}{\sim} \mathrm{MVN}(\mathbf{0}, \boldsymbol{\Sigma}_u), \quad \epsilon_{i,j} \overset{iid}{\sim} \mathrm{Normal}(0, \sigma_\epsilon^2),$$
$$i = 1, \ldots, n, \quad j = 1, \ldots, m_i, \quad N = \textstyle\sum_{i=1}^n m_i.$$

- Semi-conjugate priors on $(\boldsymbol{\beta}, \sigma_\epsilon^2)$:
  $p(\boldsymbol{\beta}, \sigma_\epsilon^2) = \mathrm{MVN}(\boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta) \cdot \text{Inv-Ga}(a_\sigma, b_\sigma)$ with $\boldsymbol{\Sigma}_\beta = \lambda^{-1} \sigma_\epsilon^2 \mathbf{I}_p$
- Posterior full conditionals of $\boldsymbol{\beta}, \sigma_\epsilon^2$:
  - $p(\boldsymbol{\beta} \mid -) = \mathrm{MVN}(\boldsymbol{\mu}_{\beta,N}, \boldsymbol{\Sigma}_{\beta,N})$,
    $$\boldsymbol{\Sigma}_{\beta,N} = (\boldsymbol{\Sigma}_\beta^{-1} + \sigma_\epsilon^{-2} \mathbf{X}^{\mathrm{T}} \mathbf{X})^{-1}, \quad \boldsymbol{\mu}_{\beta,N} = \boldsymbol{\Sigma}_{\beta,N} \left(\boldsymbol{\Sigma}_\beta^{-1} \boldsymbol{\mu}_\beta + \sigma_\epsilon^{-2} \mathbf{X}^{\mathrm{T}} \widetilde{\mathbf{y}} \right)$$
  - $p(\sigma_\epsilon^2 \mid -) = \text{Inv-Ga} \left\{ a_\sigma + \frac{N}{2}, b_\sigma + \frac{1}{2} (\widetilde{\mathbf{y}} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}} (\widetilde{\mathbf{y}} - \mathbf{X}\boldsymbol{\beta}) + \frac{\lambda}{2} \boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{\beta} \right\}$
- Conjugate prior on λ: $p(\lambda) = \mathrm{Ga}(a_\lambda, b_\lambda)$
- Posterior full conditional of λ:
- Conjugate prior on $\boldsymbol{\Sigma}_u$: $p(\boldsymbol{\Sigma}_u) = \mathrm{IW}\left(\boldsymbol{\Sigma}_u \mid \nu_0, \boldsymbol{\Sigma}_0\right)$
- Posterior full conditional of $\boldsymbol{\Sigma}_u$:
- Posterior full conditional of $\mathbf{u}_i$:

Bayesian Ridge for Linear Mixed Models

$$y_{i,j} = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \mathbf{z}_i^{\mathrm{T}}\mathbf{u}_i + \epsilon_{i,j}, \quad \mathbf{u}_i \overset{iid}{\sim} \mathrm{MVN}(\mathbf{0}, \boldsymbol{\Sigma}_u), \quad \epsilon_{i,j} \overset{iid}{\sim} \mathrm{Normal}(0, \sigma_\epsilon^2),$$
$$i = 1, \ldots, n, \quad j = 1, \ldots, m_i, \quad N = \textstyle\sum_{i=1}^n m_i.$$

- Semi-conjugate priors on $(\boldsymbol{\beta}, \sigma_\epsilon^2)$:
  $p(\boldsymbol{\beta}, \sigma_\epsilon^2) = \mathrm{MVN}(\boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta) \cdot \text{Inv-Ga}(a_\sigma, b_\sigma)$ with $\boldsymbol{\Sigma}_\beta = \lambda^{-1}\sigma_\epsilon^2\mathbf{I}_p$
- Posterior full conditionals of $\boldsymbol{\beta}, \sigma_\epsilon^2$:
  - $p(\boldsymbol{\beta} \mid -) = \mathrm{MVN}(\boldsymbol{\mu}_{\beta,N}, \boldsymbol{\Sigma}_{\beta,N})$,
    $$\boldsymbol{\Sigma}_{\beta,N} = (\boldsymbol{\Sigma}_\beta^{-1} + \sigma_\epsilon^{-2}\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}, \quad \boldsymbol{\mu}_{\beta,N} = \boldsymbol{\Sigma}_{\beta,N}\left(\boldsymbol{\Sigma}_\beta^{-1}\boldsymbol{\mu}_\beta + \sigma_\epsilon^{-2}\mathbf{X}^{\mathrm{T}}\widetilde{\mathbf{y}}\right)$$
  - $p(\sigma_\epsilon^2 \mid -) = \text{Inv-Ga}\left\{a_\sigma + \frac{N}{2}, b_\sigma + \frac{1}{2}(\widetilde{\mathbf{y}} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\widetilde{\mathbf{y}} - \mathbf{X}\boldsymbol{\beta}) + \frac{\lambda}{2}\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\beta}\right\}$
- Conjugate prior on λ: $p(\lambda) = \mathrm{Ga}(a_\lambda, b_\lambda)$
- Posterior full conditional of λ:
- Conjugate prior on $\boldsymbol{\Sigma}_u$: $p(\boldsymbol{\Sigma}_u) = \mathrm{IW}(\boldsymbol{\Sigma}_u \mid \nu_0, \boldsymbol{\Sigma}_0)$
- Posterior full conditional of $\boldsymbol{\Sigma}_u$:
- Posterior full conditional of $\mathbf{u}_i$:

Bayesian Ridge for Linear Mixed Models

$$y_{i,j} = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \mathbf{z}_i^{\mathrm{T}}\mathbf{u}_i + \epsilon_{i,j}, \quad \mathbf{u}_i \overset{iid}{\sim} \mathrm{MVN}(\mathbf{0}, \boldsymbol{\Sigma}_u), \quad \epsilon_{i,j} \overset{iid}{\sim} \mathrm{Normal}(0, \sigma_\epsilon^2),$$
$$i = 1, \ldots, n, \quad j = 1, \ldots, m_i, \quad N = \textstyle\sum_{i=1}^n m_i.$$

- Semi-conjugate priors on $(\boldsymbol{\beta}, \sigma_\epsilon^2)$:
  $p(\boldsymbol{\beta}, \sigma_\epsilon^2) = \mathrm{MVN}(\boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta) \cdot \text{Inv-Ga}(a_\sigma, b_\sigma)$ with $\boldsymbol{\Sigma}_\beta = \lambda^{-1}\sigma_\epsilon^2\mathbf{I}_p$
- Posterior full conditionals of $\boldsymbol{\beta}, \sigma_\epsilon^2$:
  - $p(\boldsymbol{\beta} \mid -) = \mathrm{MVN}(\boldsymbol{\mu}_{\beta,N}, \boldsymbol{\Sigma}_{\beta,N})$,
    $$\boldsymbol{\Sigma}_{\beta,N} = (\boldsymbol{\Sigma}_\beta^{-1} + \sigma_\epsilon^{-2}\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}, \quad \boldsymbol{\mu}_{\beta,N} = \boldsymbol{\Sigma}_{\beta,N}\left(\boldsymbol{\Sigma}_\beta^{-1}\boldsymbol{\mu}_\beta + \sigma_\epsilon^{-2}\mathbf{X}^{\mathrm{T}}\widetilde{\mathbf{y}}\right)$$
  - $p(\sigma_\epsilon^2 \mid -) = \text{Inv-Ga}\left\{a_\sigma + \frac{N}{2}, b_\sigma + \frac{1}{2}(\widetilde{\mathbf{y}} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\widetilde{\mathbf{y}} - \mathbf{X}\boldsymbol{\beta}) + \frac{\lambda}{2}\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\beta}\right\}$
- Conjugate prior on λ: $p(\lambda) = \mathrm{Ga}(a_\lambda, b_\lambda)$
- Posterior full conditional of λ:
- Conjugate prior on $\boldsymbol{\Sigma}_u$: $p(\boldsymbol{\Sigma}_u) = \mathrm{IW}(\boldsymbol{\Sigma}_u \mid \nu_0, \boldsymbol{\Sigma}_0)$
- Posterior full conditional of $\boldsymbol{\Sigma}_u$:
- Posterior full conditional of $\mathbf{u}_i$:

Bayesian Ridge for Linear Mixed Models

$$y_{i,j} = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \mathbf{z}_i^{\mathrm{T}}\mathbf{u}_i + \epsilon_{i,j}, \quad \mathbf{u}_i \overset{iid}{\sim} \mathrm{MVN}(\mathbf{0}, \boldsymbol{\Sigma}_u), \quad \epsilon_{i,j} \overset{iid}{\sim} \mathrm{Normal}(0, \sigma_\epsilon^2),$$
$$i = 1, \ldots, n, \quad j = 1, \ldots, m_i, \quad N = \sum_{i=1}^n m_i.$$

- Semi-conjugate priors on $(\boldsymbol{\beta}, \sigma_\epsilon^2)$:
  $p(\boldsymbol{\beta}, \sigma_\epsilon^2) = \mathrm{MVN}(\boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta) \cdot \text{Inv-Ga}(a_\sigma, b_\sigma)$ with $\boldsymbol{\Sigma}_\beta = \lambda^{-1}\sigma_\epsilon^2\mathbf{I}_p$
- Posterior full conditionals of $\boldsymbol{\beta}, \sigma_\epsilon^2$:
  - $p(\boldsymbol{\beta} \mid -) = \mathrm{MVN}(\boldsymbol{\mu}_{\beta,N}, \boldsymbol{\Sigma}_{\beta,N})$,
    $$\boldsymbol{\Sigma}_{\beta,N} = (\boldsymbol{\Sigma}_\beta^{-1} + \sigma_\epsilon^{-2}\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}, \quad \boldsymbol{\mu}_{\beta,N} = \boldsymbol{\Sigma}_{\beta,N}\left(\boldsymbol{\Sigma}_\beta^{-1}\boldsymbol{\mu}_\beta + \sigma_\epsilon^{-2}\mathbf{X}^{\mathrm{T}}\widetilde{\mathbf{y}}\right)$$
  - $p(\sigma_\epsilon^2 \mid -) = \text{Inv-Ga}\left\{a_\sigma + \frac{N}{2}, b_\sigma + \frac{1}{2}(\widetilde{\mathbf{y}} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\widetilde{\mathbf{y}} - \mathbf{X}\boldsymbol{\beta}) + \frac{\lambda}{2}\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\beta}\right\}$
- Conjugate prior on λ: $p(\lambda) = \mathrm{Ga}(a_\lambda, b_\lambda)$
- Posterior full conditional of λ:
  - $p(\lambda \mid -) = \mathrm{Ga}\left\{a_\lambda + p/2, b_\lambda + \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\beta}/(2\sigma_\epsilon^2)\right\}$
- Conjugate prior on $\boldsymbol{\Sigma}_u$: $p(\boldsymbol{\Sigma}_u) = \mathrm{IW}(\boldsymbol{\Sigma}_u \mid \nu_0, \boldsymbol{\Sigma}_0)$
- Posterior full conditional of $\boldsymbol{\Sigma}_u$:
- Posterior full conditional of $\mathbf{u}_i$:

Bayesian Ridge for Linear Mixed Models

$$y_{i,j} = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \mathbf{z}_i^{\mathrm{T}}\mathbf{u}_i + \epsilon_{i,j}, \quad \mathbf{u}_i \overset{iid}{\sim} \mathrm{MVN}(\mathbf{0}, \boldsymbol{\Sigma}_u), \quad \epsilon_{i,j} \overset{iid}{\sim} \mathrm{Normal}(0, \sigma_\epsilon^2),$$
$$i = 1, \ldots, n, \quad j = 1, \ldots, m_i, \quad N = \textstyle\sum_{i=1}^n m_i.$$

- Semi-conjugate priors on $(\boldsymbol{\beta}, \sigma_\epsilon^2)$:
  $p(\boldsymbol{\beta}, \sigma_\epsilon^2) = \mathrm{MVN}(\boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta) \cdot \text{Inv-Ga}(a_\sigma, b_\sigma)$ with $\boldsymbol{\Sigma}_\beta = \lambda^{-1}\sigma_\epsilon^2\mathbf{I}_p$
- Posterior full conditionals of $\boldsymbol{\beta}, \sigma_\epsilon^2$:
  - $p(\boldsymbol{\beta} \mid -) = \mathrm{MVN}(\boldsymbol{\mu}_{\beta,N}, \boldsymbol{\Sigma}_{\beta,N})$,
    $$\boldsymbol{\Sigma}_{\beta,N} = (\boldsymbol{\Sigma}_\beta^{-1} + \sigma_\epsilon^{-2}\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}, \quad \boldsymbol{\mu}_{\beta,N} = \boldsymbol{\Sigma}_{\beta,N}\left(\boldsymbol{\Sigma}_\beta^{-1}\boldsymbol{\mu}_\beta + \sigma_\epsilon^{-2}\mathbf{X}^{\mathrm{T}}\widetilde{\mathbf{y}}\right)$$
  - $p(\sigma_\epsilon^2 \mid -) = \text{Inv-Ga}\left\{a_\sigma + \dfrac{N}{2}, b_\sigma + \dfrac{1}{2}(\widetilde{\mathbf{y}} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\widetilde{\mathbf{y}} - \mathbf{X}\boldsymbol{\beta}) + \dfrac{\lambda}{2}\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\beta}\right\}$
- Conjugate prior on λ: $p(\lambda) = \mathrm{Ga}(a_\lambda, b_\lambda)$
- Posterior full conditional of λ:
  - $p(\lambda \mid -) = \mathrm{Ga}\left\{a_\lambda + p/2, b_\lambda + \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\beta}/(2\sigma_\epsilon^2)\right\}$
- Conjugate prior on $\boldsymbol{\Sigma}_u$: $p(\boldsymbol{\Sigma}_u) = \mathrm{IW}(\boldsymbol{\Sigma}_u \mid \nu_0, \boldsymbol{\Sigma}_0)$
- Posterior full conditional of $\boldsymbol{\Sigma}_u$:
  - $p(\boldsymbol{\Sigma}_u \mid -) = \mathrm{IW}(n + \nu_0, \boldsymbol{\Sigma}_0 + \sum_{i=1}^n \mathbf{u}_i\mathbf{u}_i^{\mathrm{T}})$
- Posterior full conditional of $\mathbf{u}_i$:
  - $p(\mathbf{u}_i \mid -) = \mathrm{MVN}(\boldsymbol{\mu}_{i,u,N}, \boldsymbol{\Sigma}_{i,u,N})$,

$$\boldsymbol{\Sigma}_{i,u,N} = (\boldsymbol{\Sigma}_u^{-1} + \sigma_\epsilon^{-2}m_i\mathbf{z}_i\mathbf{z}_i^{\mathrm{T}})^{-1}, \quad \boldsymbol{\mu}_{i,u,N} = \boldsymbol{\Sigma}_{i,u,N}\left\{\sigma_\epsilon^{-2}\textstyle\sum_{j=1}^{m_i}(y_{i,j} - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})\mathbf{z}_i\right\}$$

Bayesian LASSO for Linear Mixed Models

$$y_{i,j} = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \mathbf{z}_i^{\mathrm{T}}\mathbf{u}_i + \epsilon_{i,j}, \quad \mathbf{u}_i \overset{iid}{\sim} \mathrm{MVN}(\mathbf{0}, \boldsymbol{\Sigma}_u), \quad \epsilon_{i,j} \overset{iid}{\sim} \mathrm{Normal}(0, \sigma_\epsilon^2),$$
$$i = 1, \ldots, n, \quad j = 1, \ldots, m_i, \quad N = \textstyle\sum_{i=1}^n m_i.$$

- Semi-conjugate LASSO priors on $\boldsymbol{\beta}$:
  $$p(\boldsymbol{\beta} \mid \lambda, \sigma_\epsilon^2) = \textstyle\prod_{j=1}^p \int_0^\infty \mathrm{Normal}(\beta_j \mid 0, \tau_j^2)\ \mathrm{Exp}\left(\tau_j^2 \mid \frac{\lambda^2}{2\sigma_\epsilon^2}\right) d\tau_j^2$$
- Posterior full conditionals for block-Gibbs sampler:
- Conjugate prior on λ^2: $p(\lambda^2) = \mathrm{Ga}(a_\lambda, b_\lambda)$
- Posterior full conditional of λ^2:
- Posterior full conditional of $\boldsymbol{\Sigma}_u$:
- Posterior full conditional of $\mathbf{u}_i$:

Bayesian LASSO for Linear Mixed Models

$$y_{i,j} = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \mathbf{z}_i^{\mathrm{T}}\mathbf{u}_i + \epsilon_{i,j}, \quad \mathbf{u}_i \overset{iid}{\sim} \mathrm{MVN}(\mathbf{0}, \boldsymbol{\Sigma}_u), \quad \epsilon_{i,j} \overset{iid}{\sim} \mathrm{Normal}(0, \sigma_\epsilon^2),$$
$$i = 1, \ldots, n, \quad j = 1, \ldots, m_i, \quad N = \textstyle\sum_{i=1}^n m_i.$$

- Semi-conjugate LASSO priors on $\boldsymbol{\beta}$:
$$p(\boldsymbol{\beta} \mid \lambda, \sigma_\epsilon^2) = \textstyle\prod_{j=1}^p \int_0^\infty \mathrm{Normal}(\beta_j \mid 0, \tau_j^2)\ \mathrm{Exp}\left(\tau_j^2 \mid \frac{\lambda^2}{2\sigma_\epsilon^2}\right) d\tau_j^2$$
- Posterior full conditionals for block-Gibbs sampler:
  - $p(\boldsymbol{\beta} \mid -) = \mathrm{MVN}(\boldsymbol{\mu}_{\beta,N}, \boldsymbol{\Sigma}_{\beta,N}), \qquad \mathbf{D}_\tau = \mathrm{diag}(\tau_1^2, \ldots, \tau_p^2),$
  $$\boldsymbol{\Sigma}_{\beta,N} = \sigma_\epsilon^2(\mathbf{X}^{\mathrm{T}}\mathbf{X} + \mathbf{D}_\tau^{-1})^{-1}, \quad \boldsymbol{\mu}_{\beta,N} = \boldsymbol{\Sigma}_{\beta,n}(\sigma_\epsilon^{-2}\mathbf{X}^{\mathrm{T}}\widetilde{\mathbf{y}}) = (\mathbf{X}^{\mathrm{T}}\mathbf{X} + \mathbf{D}_\tau^{-1})^{-1}\mathbf{X}^{\mathrm{T}}\widetilde{\mathbf{y}}$$
  - $p(\sigma_\epsilon^2 \mid -) = \text{Inv-Ga}\left\{a_\sigma + \frac{n+p}{2}, b_\sigma + \frac{(\tilde{\mathbf{y}} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\tilde{\mathbf{y}} - \mathbf{X}\boldsymbol{\beta})}{2} + \frac{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{D}_\tau^{-1}\boldsymbol{\beta}}{2}\right\}$
  - $\tau_j^2 \to w_j = \tau_j^{-2},\ p(w_j \mid -) = \text{Inv-Gs}(\mu', \lambda'),\ \ \mu' = \frac{\lambda}{\sigma_\epsilon |\beta_j|},\ \lambda' = \frac{\lambda^2}{\sigma_\epsilon^2}$
- Conjugate prior on λ^2: $p(\lambda^2) = \mathrm{Ga}(a_\lambda, b_\lambda)$
- Posterior full conditional of λ^2:
- Posterior full conditional of $\boldsymbol{\Sigma}_u$:
- Posterior full conditional of $\mathbf{u}_i$:

Bayesian LASSO for Linear Mixed Models

$$y_{i,j} = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \mathbf{z}_i^{\mathrm{T}}\mathbf{u}_i + \epsilon_{i,j}, \quad \mathbf{u}_i \overset{iid}{\sim} \mathrm{MVN}(\mathbf{0}, \boldsymbol{\Sigma}_u), \quad \epsilon_{i,j} \overset{iid}{\sim} \mathrm{Normal}(0, \sigma_\epsilon^2),$$

$$i = 1, \ldots, n, \quad j = 1, \ldots, m_i, \quad N = \textstyle\sum_{i=1}^{n} m_i.$$

- Semi-conjugate LASSO priors on $\boldsymbol{\beta}$:

$$p(\boldsymbol{\beta} \mid \lambda, \sigma_\epsilon^2) = \prod_{j=1}^{p} \int_0^\infty \mathrm{Normal}(\beta_j \mid 0, \tau_j^2)\ \mathrm{Exp}\left(\tau_j^2 \mid \frac{\lambda^2}{2\sigma_\epsilon^2}\right) d\tau_j^2$$

- Posterior full conditionals for block-Gibbs sampler:
  - $p(\boldsymbol{\beta} \mid -) = \mathrm{MVN}(\boldsymbol{\mu}_{\beta,N}, \boldsymbol{\Sigma}_{\beta,N}), \qquad \mathbf{D}_\tau = \mathrm{diag}(\tau_1^2, \ldots, \tau_p^2),$

    $\boldsymbol{\Sigma}_{\beta,N} = \sigma_\epsilon^2(\mathbf{X}^{\mathrm{T}}\mathbf{X} + \mathbf{D}_\tau^{-1})^{-1}, \quad \boldsymbol{\mu}_{\beta,N} = \boldsymbol{\Sigma}_{\beta,n}(\sigma_\epsilon^{-2}\mathbf{X}^{\mathrm{T}}\widetilde{\mathbf{y}}) = (\mathbf{X}^{\mathrm{T}}\mathbf{X} + \mathbf{D}_\tau^{-1})^{-1}\mathbf{X}^{\mathrm{T}}\widetilde{\mathbf{y}}$
  - $p(\sigma_\epsilon^2 \mid -) = \text{Inv-Ga}\left\{a_\sigma + \frac{n+p}{2}, b_\sigma + \frac{(\widetilde{\mathbf{y}} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\widetilde{\mathbf{y}} - \mathbf{X}\boldsymbol{\beta})}{2} + \frac{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{D}_\tau^{-1}\boldsymbol{\beta}}{2}\right\}$
  - $\tau_j^2 \to w_j = \tau_j^{-2},\ p(w_j \mid -) = \text{Inv-Gs}(\mu', \lambda'),\ \ \mu' = \frac{\lambda}{\sigma_\epsilon|\beta_j|},\ \lambda' = \frac{\lambda^2}{\sigma_\epsilon^2}$
- Conjugate prior on λ^2: $p(\lambda^2) = \mathrm{Ga}(a_\lambda, b_\lambda)$
- Posterior full conditional of λ^2:
- Posterior full conditional of $\boldsymbol{\Sigma}_u$:
- Posterior full conditional of $\mathbf{u}_i$:

Bayesian LASSO for Linear Mixed Models

$$y_{i,j} = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \mathbf{z}_i^{\mathrm{T}}\mathbf{u}_i + \epsilon_{i,j}, \quad \mathbf{u}_i \overset{iid}{\sim} \mathrm{MVN}(\mathbf{0}, \boldsymbol{\Sigma}_u), \quad \epsilon_{i,j} \overset{iid}{\sim} \mathrm{Normal}(0, \sigma_\epsilon^2),$$
$$i = 1, \ldots, n, \quad j = 1, \ldots, m_i, \quad N = \textstyle\sum_{i=1}^n m_i.$$

- Semi-conjugate LASSO priors on $\boldsymbol{\beta}$:
$$p(\boldsymbol{\beta} \mid \lambda, \sigma_\epsilon^2) = \textstyle\prod_{j=1}^p \int_0^\infty \mathrm{Normal}(\beta_j \mid 0, \tau_j^2)\ \mathrm{Exp}\left(\tau_j^2 \mid \frac{\lambda^2}{2\sigma_\epsilon^2}\right) d\tau_j^2$$
- Posterior full conditionals for block-Gibbs sampler:
  - $p(\boldsymbol{\beta} \mid -) = \mathrm{MVN}(\boldsymbol{\mu}_{\beta,N}, \boldsymbol{\Sigma}_{\beta,N}), \qquad \mathbf{D}_\tau = \mathrm{diag}(\tau_1^2, \ldots, \tau_p^2),$
  $$\boldsymbol{\Sigma}_{\beta,N} = \sigma_\epsilon^2(\mathbf{X}^{\mathrm{T}}\mathbf{X} + \mathbf{D}_\tau^{-1})^{-1}, \quad \boldsymbol{\mu}_{\beta,N} = \boldsymbol{\Sigma}_{\beta,n}(\sigma_\epsilon^{-2}\mathbf{X}^{\mathrm{T}}\widetilde{\mathbf{y}}) = (\mathbf{X}^{\mathrm{T}}\mathbf{X} + \mathbf{D}_\tau^{-1})^{-1}\mathbf{X}^{\mathrm{T}}\widetilde{\mathbf{y}}$$
  - $p(\sigma_\epsilon^2 \mid -) = \mathrm{Inv\text{-}Ga}\left\{a_\sigma + \frac{n+p}{2}, b_\sigma + \frac{(\widetilde{\mathbf{y}} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\widetilde{\mathbf{y}} - \mathbf{X}\boldsymbol{\beta})}{2} + \frac{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{D}_\tau^{-1}\boldsymbol{\beta}}{2}\right\}$
  - $\tau_j^2 \to w_j = \tau_j^{-2},\ p(w_j \mid -) = \mathrm{Inv\text{-}Gs}(\mu', \lambda'),\ \ \mu' = \frac{\lambda}{\sigma_\epsilon|\beta_j|},\ \lambda' = \frac{\lambda^2}{\sigma_\epsilon^2}$
- Conjugate prior on λ^2: $p(\lambda^2) = \mathrm{Ga}(a_\lambda, b_\lambda)$
- Posterior full conditional of λ^2:
  - $p(\lambda^2 \mid -) = \mathrm{Ga}\left\{a_\lambda + p, b_\lambda + \sum_{j=1}^p \tau_j^2/(2\sigma_\epsilon^2)\right\}.$
- Posterior full conditional of $\boldsymbol{\Sigma}_u$:
  - $p(\boldsymbol{\Sigma}_u \mid -) = \mathrm{IW}(n + \nu_0, \boldsymbol{\Sigma}_0 + \sum_{i=1}^n \mathbf{u}_i\mathbf{u}_i^{\mathrm{T}})$
- Posterior full conditional of $\mathbf{u}_i$:
  - $p(\mathbf{u}_i \mid -) = \mathrm{MVN}(\boldsymbol{\mu}_{i,u,N}, \boldsymbol{\Sigma}_{i,u,N}),$

$$\boldsymbol{\Sigma}_{i,u,N} = (\boldsymbol{\Sigma}_u^{-1} + \sigma_\epsilon^{-2}m_i\mathbf{z}_i\mathbf{z}_i^{\mathrm{T}})^{-1}, \quad \boldsymbol{\mu}_{i,u,N} = \boldsymbol{\Sigma}_{i,u,N}\left\{\sigma_\epsilon^{-2}\textstyle\sum_{j=1}^{m_i}(y_{i,j} - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})\mathbf{z}_i\right\}$$

- Linear models are regression models linear in parameters.
  - Ordinary least squares minimizes the L_2 distance between the observed and the model hypothesized response values.
  - MLE under a normal likelihood is naturally connected to OLS.
  - Ridge regression is useful when the model matrix is ill-conditioned.
  - LASSO is useful when the model is sparse.
  - Bayesian MCMC based inference is straightforward under (semi)conjugate priors.
  - Ridge estimates can be obtained as Bayesian MAP under a type of independent normal priors on the regression coefficients.
  - LASSO estimates can be obtained as Bayesian MAP under a type of independent Laplace priors on the regression coefficients.
  - Mixed models are useful for including population level fixed effects as well as individual level random effects.
  - Mixed models are generally computation intensive but can usually be relatively easily handled using Bayesian hierarchies.

- Linear models are regression models linear in parameters.
  - Ordinary least squares minimizes the L_2 distance between the observed and the model hypothesized response values.
  - MLE under a normal likelihood is naturally connected to OLS.
  - Ridge regression is useful when the model matrix is ill-conditioned.
  - LASSO is useful when the model is sparse.
  - Bayesian MCMC based inference is straightforward under (semi)conjugate priors.
  - Ridge estimates can be obtained as Bayesian MAP under a type of independent normal priors on the regression coefficients.
  - LASSO estimates can be obtained as Bayesian MAP under a type of independent Laplace priors on the regression coefficients.
  - Mixed models are useful for including population level fixed effects as well as individual level random effects.
  - Mixed models are generally computation intensive but can usually be relatively easily handled using Bayesian hierarchies.

Module IX Summary

- Linear models are regression models linear in parameters.
  - Ordinary least squares minimizes the L_2 distance between the observed and the model hypothesized response values.
  - MLE under a normal likelihood is naturally connected to OLS.
  - Ridge regression is useful when the model matrix is ill-conditioned.
  - LASSO is useful when the model is sparse.
  - Bayesian MCMC based inference is straightforward under (semi)conjugate priors.
  - Ridge estimates can be obtained as Bayesian MAP under a type of independent normal priors on the regression coefficients.
  - LASSO estimates can be obtained as Bayesian MAP under a type of independent Laplace priors on the regression coefficients.
  - Mixed models are useful for including population level fixed effects as well as individual level random effects.
  - Mixed models are generally computation intensive but can usually be relatively easily handled using Bayesian hierarchies.

Module IX Summary

- Linear models are regression models linear in parameters.
    - Ordinary least squares minimizes the L_2 distance between the observed and the model hypothesized response values.
    - MLE under a normal likelihood is naturally connected to OLS.
    - Ridge regression is useful when the model matrix is ill-conditioned.
    - LASSO is useful when the model is sparse.
    - Bayesian MCMC based inference is straightforward under (semi)conjugate priors.
    - Ridge estimates can be obtained as Bayesian MAP under a type of independent normal priors on the regression coefficients.
    - LASSO estimates can be obtained as Bayesian MAP under a type of independent Laplace priors on the regression coefficients.
    - Mixed models are useful for including population level fixed effects as well as individual level random effects.
    - Mixed models are generally computation intensive but can usually be relatively easily handled using Bayesian hierarchies.

Module IX Summary

- Linear models are regression models linear in parameters.
  - Ordinary least squares minimizes the L_2 distance between the observed and the model hypothesized response values.
  - MLE under a normal likelihood is naturally connected to OLS.
  - Ridge regression is useful when the model matrix is ill-conditioned.
  - LASSO is useful when the model is sparse.
  - Bayesian MCMC based inference is straightforward under (semi)conjugate priors.
  - Ridge estimates can be obtained as Bayesian MAP under a type of independent normal priors on the regression coefficients.
  - LASSO estimates can be obtained as Bayesian MAP under a type of independent Laplace priors on the regression coefficients.
  - Mixed models are useful for including population level fixed effects as well as individual level random effects.
  - Mixed models are generally computation intensive but can usually be relatively easily handled using Bayesian hierarchies.

Module IX Summary

- Linear models are regression models linear in parameters.
    - Ordinary least squares minimizes the L_2 distance between the observed and the model hypothesized response values.
    - MLE under a normal likelihood is naturally connected to OLS.
    - Ridge regression is useful when the model matrix is ill-conditioned.
    - LASSO is useful when the model is sparse.
    - Bayesian MCMC based inference is straightforward under (semi)conjugate priors.
    - Ridge estimates can be obtained as Bayesian MAP under a type of independent normal priors on the regression coefficients.
    - LASSO estimates can be obtained as Bayesian MAP under a type of independent Laplace priors on the regression coefficients.
    - Mixed models are useful for including population level fixed effects as well as individual level random effects.
    - Mixed models are generally computation intensive but can usually be relatively easily handled using Bayesian hierarchies.

- Linear models are regression models linear in parameters.
  - Ordinary least squares minimizes the L_2 distance between the observed and the model hypothesized response values.
  - MLE under a normal likelihood is naturally connected to OLS.
  - Ridge regression is useful when the model matrix is ill-conditioned.
  - LASSO is useful when the model is sparse.
  - Bayesian MCMC based inference is straightforward under (semi)conjugate priors.
  - Ridge estimates can be obtained as Bayesian MAP under a type of independent normal priors on the regression coefficients.
  - LASSO estimates can be obtained as Bayesian MAP under a type of independent Laplace priors on the regression coefficients.
  - Mixed models are useful for including population level fixed effects as well as individual level random effects.
  - Mixed models are generally computation intensive but can usually be relatively easily handled using Bayesian hierarchies.

Module IX Summary

- Linear models are regression models linear in parameters.
  - Ordinary least squares minimizes the L_2 distance between the observed and the model hypothesized response values.
  - MLE under a normal likelihood is naturally connected to OLS.
  - Ridge regression is useful when the model matrix is ill-conditioned.
  - LASSO is useful when the model is sparse.
  - Bayesian MCMC based inference is straightforward under (semi)conjugate priors.
  - Ridge estimates can be obtained as Bayesian MAP under a type of independent normal priors on the regression coefficients.
  - LASSO estimates can be obtained as Bayesian MAP under a type of independent Laplace priors on the regression coefficients.
  - Mixed models are useful for including population level fixed effects as well as individual level random effects.
  - Mixed models are generally computation intensive but can usually be relatively easily handled using Bayesian hierarchies.

Module IX Summary

- Linear models are regression models linear in parameters.
  - Ordinary least squares minimizes the L_2 distance between the observed and the model hypothesized response values.
  - MLE under a normal likelihood is naturally connected to OLS.
  - Ridge regression is useful when the model matrix is ill-conditioned.
  - LASSO is useful when the model is sparse.
  - Bayesian MCMC based inference is straightforward under (semi)conjugate priors.
  - Ridge estimates can be obtained as Bayesian MAP under a type of independent normal priors on the regression coefficients.
  - LASSO estimates can be obtained as Bayesian MAP under a type of independent Laplace priors on the regression coefficients.
  - Mixed models are useful for including population level fixed effects as well as individual level random effects.
  - Mixed models are generally computation intensive but can usually be relatively easily handled using Bayesian hierarchies.

Module IX Summary

- Linear models are regression models linear in parameters.
  - Ordinary least squares minimizes the L_2 distance between the observed and the model hypothesized response values.
  - MLE under a normal likelihood is naturally connected to OLS.
  - Ridge regression is useful when the model matrix is ill-conditioned.
  - LASSO is useful when the model is sparse.
  - Bayesian MCMC based inference is straightforward under (semi)conjugate priors.
  - Ridge estimates can be obtained as Bayesian MAP under a type of independent normal priors on the regression coefficients.
  - LASSO estimates can be obtained as Bayesian MAP under a type of independent Laplace priors on the regression coefficients.
  - Mixed models are useful for including population level fixed effects as well as individual level random effects.
  - Mixed models are generally computation intensive but can usually be relatively easily handled using Bayesian hierarchies.