# ECON220B Discussion Section 3

# Linear Regression and Bayesian Inference

---

Lapo Bini

# Roadmap

1. Where is $\hat{\beta}^{\text{OLS}}$ going?
2. Linear Projection
3. Ridge Regression
4. Bayesian Inference

# Understanding The Assumptions

Linear regression: $y_i = x_i^T \beta + u_i$ with $\{u_1, \ldots, u_n\}$ iid, $E[u_i] = 0$.

1. If $u_i | x_i \sim \mathcal{N}(0, \sigma^2)$ then $\hat{\beta}^{OLS}$ is BLUE by Markov-Gauss theorem, $\hat{\beta}^{OLS} = \hat{\beta}^{MLE}$. We are estimating a causal effect $x \to y$, i.e.
$$\frac{\partial}{\partial x_i} E[y_i | x_i] = \beta$$

2. If $E[u_i | x_i] = 0$ and, $E[u_i^2 | x_i] = \sigma^2$, then $\hat{\beta}^{OLS}$ is BLUE by Markov-Gauss theorem. We are estimating a causal effect.

3. If $E[u_i | x_i] \neq 0$ but $E[u_i \, x_i] = 0$ still holds then $\hat{\beta}^{OLS} \xrightarrow{p} \beta$ but we are estimating correlation between $x$ and $y$, no partial effects.

# Example (1/2)

Model: $y_i = \beta x_i + u_i$ $u_i = x_i^2 + \eta_i$ with true parameter $\beta = 3$, and $x_i \sim \mathcal{N}(0,1)$, $\eta_i \sim \mathcal{N}(0,4)$, $x_i \perp \eta_i$. $E[x_i \eta_i] = 0 \Rightarrow Cov(x_i, \eta_i) = 0$

(1) Suppose we estimate the model by OLS, can we apply Markov-Gauss theorem?

(2) Is $\hat{\beta}^{OLS}$ consistent for the true $\beta$?

(1) $E[u_i | x_i] = x^2$ NO WE CAN'T. NOTE: $\frac{\partial}{\partial \beta} E[y_i | x_i] = \underbrace{\beta}_{\text{APE}} + \underbrace{2x_i}_{\text{HETEROG. EFFECT}}$

(2) $\hat{\beta} = \left(\frac{1}{n}\sum x_i x_i^T\right)^{-1}\left(\frac{1}{n}\sum x_i y_i\right) = \left(\frac{1}{n}\sum x_i x_i^T\right)^{-1}\left(\frac{1}{n}\sum x_i (x_i^T\beta + x_i^2 + \eta_i)\right)$

$= \left(\frac{1}{n}\sum x_i x_i^T\right)^{-1}\left(\frac{1}{n}\sum x_i x_i^T \beta\right) + \left(\frac{1}{n}\sum x_i x_i^T\right)^{-1}\left(\frac{1}{n}\sum x_i x_i^2\right) + \left(\frac{1}{n}\sum x_i x_i^T\right)^{-1}\left(\frac{1}{n}\sum x_i \eta_i\right)$

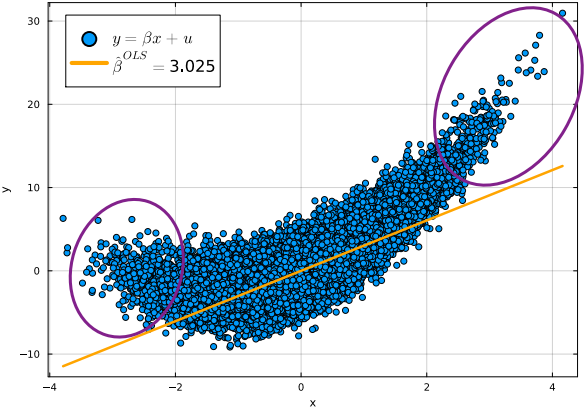
$= \beta + 0 + 0$

$\to E[X^3]$ but $X \sim N(0,1)$, $Skew(X) = 0$, $\therefore E[x^3] = 0$

$\xrightarrow{p} E[x_i \eta_i]$

# Example (1/2)

Model: $y_i = \beta x_i + u_i$, $u_i = x_i^2 + \eta_i$ with true parameter $\beta = 3$, and $x_i \sim \mathcal{N}(0, 1)$, $\eta_i \sim \mathcal{N}(0, 4)$, $x_i \perp \eta_i$.



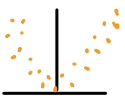

m = 5000

EFFECT QUADRATIC TERM

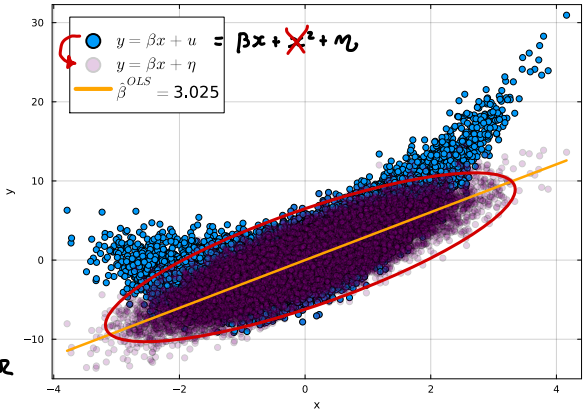
# Example (1/2)

Model: $y_i = \beta x_i + u_i$ $u_i = x_i^2 + \eta_i$ with true parameter $\beta = 3$, and $x_i \sim \mathcal{N}(0, 1)$, $\eta_i \sim \mathcal{N}(0, 4)$, $x_i \perp \eta_i$.

$\hat{\beta}^{OLS}$ CORR.



- OFFSET
- NON LINEAR
- $Cov(x, u) = 0$ LINEAR ESTIMATOR



# Example (2/2)

Now we have: $y_i = \beta x_i + \eta_i$ with true parameter $\beta = 3$, and $x_i \sim \mathcal{N}(0, 1)$, $\eta_i \sim \mathcal{N}(0, 4)$.

$\hookrightarrow \quad x_i = \frac{1}{\beta} y_i - \frac{1}{\beta} \eta_i$

(1) Suppose that instead of running a regression of $y_i$ on $x_i$, you run the regression of $x_i$ and $y_i$, that is you switch the dependent and independent variables:

$$x_i = \phi y_i + \nu_i$$

What is $\hat{\phi}^{OLS}$ estimating? $E[\nu_i y_i] = -\frac{1}{\beta} E[\eta_i y_i]$

$= -\frac{1}{\beta} E[\eta_i (\beta x_i + \eta_i)] = -\frac{\beta}{\beta} E[\eta_i x_i] - \frac{1}{\beta} E[\eta_i^2] = -\frac{\sigma^2}{\beta} \neq 0$

NOT CONSISTENT

# Example (2/2)

Now we have: $y_i = \beta x_i + \eta_i$ with true parameter $\beta = 3$, and $x_i \sim \mathcal{N}(0, 1)$, $\eta_i \sim \mathcal{N}(0, 4)$.

(1) Suppose that instead of running a regression of $y_i$ on $x_i$, you run the regression of $x_i$ and $y_i$, that is you switch the dependent and independent variables:

$$x_i = \phi y_i + \nu_i$$

What is $\hat{\phi}^{OLS}$ estimating? $\hat{\phi} = \left(\frac{1}{m}\sum y_i y_i^T\right)^{-1}\left(\frac{1}{m}\sum y_i x_i\right)$

$\sigma^2 = 4$

$\phi = 1/\beta = 0.33$

$$\left(\frac{1}{m}\sum y_i y_i^T\right)^{-1}\left(\frac{1}{m}\sum y_i (\phi y_i + \nu_i)\right) = \phi + \left(\frac{1}{m}\sum y_i^2\right)^{-1}\left(\frac{1}{m}\sum \nu_i y_i\right) = \phi + \frac{(-\sigma^2/\beta)}{\text{Var}(y_i) + E[y_i^2]}$$

$\xrightarrow{P}$ $E[y_i^2] = \text{Var}(y_i) + E[y_i]^2$

$\text{Var}(\beta x_i) + \text{Var}(\eta_i) = 9 \cdot 1 + 4 = 13$

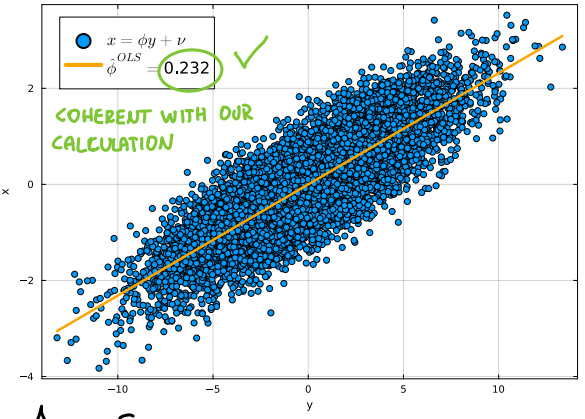
$(E[y_i])^2 = (E[\beta x_i] + E[\eta_i])^2 = 0^2$

$\therefore \hat{\phi}^{OLS} \xrightarrow{P} \phi + \Delta$

$\hat{\phi}^{OLS} \xrightarrow{P} 0.33 + \frac{(-4/3)}{13} = 0.33 - 0.10$

# Example (2/2)

Now we have: $y_i = \beta x_i + \eta_i$ with true parameter $\beta = 3$, and $x_i \sim \mathcal{N}(0, 1)$, $\eta_i \sim \mathcal{N}(0, 4)$.




$\hat{\phi}^{OLS} \xrightarrow{p} \phi + \Delta := \delta$

# Linear Projection

- If $E[u_i x_i] \neq 0$ then $\hat{\beta}^{OLS} \xrightarrow{p} \delta \equiv \beta + \Delta$ it converges to the coefficient of the linear projection.

- The linear projection $y_i = x_i^T \delta + u_i$ is also called the **minimum mean square linear predictor** since $\delta$ solves the following problem:

$$\min_{\mathbf{d} \in \mathbb{R}^k} E\left[(y_i - x_i^T \mathbf{d})^2\right]$$

- The linear projection **always** satisfies $E[x_i u_i] = 0$ and $E[u_i] = 0$.

$\hat{u}_i = y_i - x_i^T \hat{\beta} = x_i^T \beta + u_i - x_i^T \hat{\beta} = u_i + x_i^T (\beta - \hat{\beta})$ TAKE SUM AND DIVIDE BOTH TERMS BY $n$: $\frac{1}{n} \sum \hat{u}_i = \frac{1}{n} \sum u_i + \frac{1}{n} \sum x_i^T (\beta - \hat{\beta})$

CAN WE APPLY WLLN? IS THIS ZERO?

$$\hat{u}_i = y_i - x_i^T \hat{\beta} = x_i^T \delta + u_i - x_i^T \hat{\beta}$$

$$\hat{u}_i = u_i + x_i^T (\delta - \hat{\beta})$$

(a) $\frac{1}{m} \sum \hat{u}_i = \frac{1}{m} \sum u_i + \frac{1}{m} \sum x_i^T (\delta - \hat{\beta})$

$\hat{\beta} \xrightarrow{p} \delta$

(b) $x_i \hat{u}_i = x_i u_i + x_i x_i^T (\delta + \hat{\beta})$

$\frac{1}{m} \sum x_i \hat{u}_i = \frac{1}{m} \sum x_i u_i + \frac{1}{m} \sum x_i x_i^T (\delta + \beta)$

$o_p(1)$

$\frac{1}{m} \sum x_i \hat{u}_i \xrightarrow{p} E[x_i u_i]$

$= 0$

# When Does $E[u_i x_i] = 0$ Fail?

- Omitted variable bias: consider the following linear regression model $y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + u_i$ where $y_i, x_i, z_i, u_i$ are all scalars and $E[u_i x_i] = E[u_i z_i] = 0$

- Suppose we regress $y_i$ on $x_i$ only: what is the probability limit of $\hat{\beta}_1^{OLS}$? When does the limit coincide with the true parameter $\beta_1$?

# Two Religions: Frequentists vs Bayesians

WE ARE GOING TO START FROM AN EXAMPLE BEFORE FORMAL ANALISYS

Given $\{y_1, \ldots, y_n\}$ iid sample with $y_i \sim \mathcal{N}(\mu, \sigma^2)$ we are interested in the population mean $\mu$. We already know that MLE estimator is $\hat{\mu}^{MLE} = n^{-1} \sum_{i=1}^{n} y_i \sim \mathcal{N}(\mu, \sigma^2/n)$. Two different approaches:

1. Frequentist: the data is the result of sampling from a random process. Frequentists see the data as varying and the parameter $\mu$ of this random process that generates the data as being fixed. $\mathcal{N}(\mu, \sigma^2/n)$ describes a distribution across different samples.

2. Bayesian: $\mu$ treates as a random variable. Bayesians have prior beliefs about $\mu$ (**prior distribution**), which is updated after observing the data (**likelihood function**) using **Bayes' Rule**. The **posterior distribution** summarises the uncertainty about credible values of $\mu$.

# Ridge Regression

- Consider the follow linear regression model $y_i = x_i^T \beta + u_i$, $u_i \sim \mathcal{N}(0, 1)$.
- Assume that the parameters $\beta \in \mathbb{R}^d$ follow the distribution $\beta \sim \mathcal{N}(0, \lambda^2 \mathrm{I}_d)$ where $\lambda > 0$ and $\mathrm{I}_d$ is the $(d \mathrm{x} d)$ identity matrix.
- Lastly, assume that $u_i, x_i, \beta$ are mutually independent.

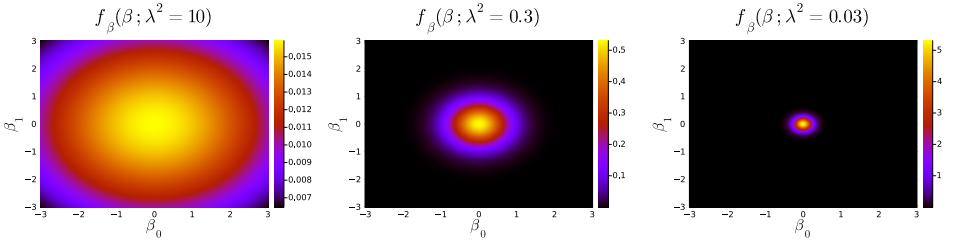
(1) Prove that $f_\beta(\beta) = \lambda^{-d} \prod_{j=1}^{d} \phi(\beta_j/\lambda)$.
(2) Show that $f_{\mathbf{Y}|\beta,\mathbf{X}}(y_1, \ldots, y_n|\beta, \mathbf{X}) = \prod_{i=1}^{n} \phi(y_i - x_i^T \beta)$.
(3) Derive the Maximum Likelihood Estimator $\hat{\beta}^{MLE}$.
(4) Find the posterior distribution $f_{\beta|\mathbf{Y},\mathbf{X}}(\beta|\mathbf{Y},\mathbf{X})$ and derive the Bayes estimator defined as

$$\hat{\beta}^{Bayes} \equiv \arg\max_{\beta} f_{\beta|\mathbf{Y},\mathbf{X}}(\beta|\mathbf{Y},\mathbf{X})$$

# Ridge Regression - Prior Distribution

Before observing the data, our prior belief is that the parameters are most likely to be close to zero. The parameter $\lambda^2$ represents the uncertainty of our guess, i.e. $\beta \sim \mathcal{N}(0, \lambda^2 I_2)$.

Figure: Prior distribution for different values of $\lambda^2$.



(1) $\beta \sim N(0, \lambda^2 I_d)$

UNCORRELATED $\Rightarrow$ INDEP.

$\therefore \beta_i \sim N(0, \lambda^2)$

$$f_\beta = \prod_{i=1}^{m} f_{\beta_i} = \prod_{i}^{m} \frac{1}{\sqrt{2\pi}\lambda} e^{\left\{-\frac{1}{2}(\beta/\lambda)^2\right\}}$$

$$= \frac{1}{\lambda^d} \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi}} e^{\left\{-\frac{1}{2}(\beta/\lambda)^2\right\}}$$

$\frac{\beta}{\lambda} \sim N(0,1)$

$$\therefore f_\beta(b) = \frac{1}{\lambda^d} \prod_{i=1}^{m} \phi\left(\frac{\beta_i}{\lambda_i}\right)$$

(2) $y_i = x_i^T\beta + u_i$

$y_i | x_i^T\beta = K + u_i = K + N(0,1)$

$= N(K,1)$

$\{y_1, \dots, y_m\}$ IID

$\therefore f_{Y|\beta X} = \prod_{i=1}^{m} f_{Y_i|\beta X} = \prod_{i}^{m} \frac{1}{\sqrt{2\pi}} e^{\{-\frac{1}{2}(y_i - K)^2\}}$

$= \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi}} e^{\{-\frac{1}{2}(y_i - x_i^T\beta)^2\}}$

PDF STANDARD NORMAL

$= \prod_{i=1}^{m} \phi(y_i - x_i^T\beta)$

(3) $\hat{\beta} = \arg\max \prod_{i=1}^{m} \phi(y_i - x_i^T\beta)$

$LIK = \prod_{i=1}^{m} \phi(y_i - x_i^T\beta)$

$LIK = (2\pi)^{-\frac{m}{2}} \exp\{-\frac{1}{2}\sum (y_i - x_i^T\beta)^2\}$

NOW I TAKE LOG

$\log(LIK) = \log(2\pi)^{-\frac{m}{2}} - \frac{1}{2}\sum (y_i - x_i^T\beta)^2$

max $\log(LIK)$ IS EQUIVALENT TO

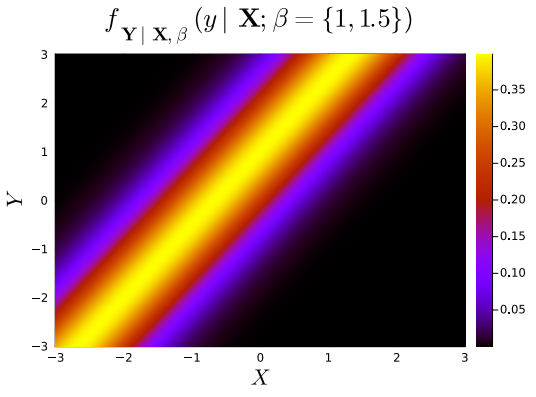
min $\frac{1}{2}\sum (y_i - x_i^T\hat{\beta})^2$

FOC $-\sum x_i (y_i - x_i^T\hat{\beta}) = 0$

$\hat{\beta} = \left(\frac{1}{m}\sum x_i x_i^T\right)^{-1} \left(\frac{1}{m}\sum x_i y_i\right) = \hat{\beta}^{OLS}$

# Ridge Regression - Likelihood Function

The likelihood describes the probability of the data that has already been observed given certain parameter values $\beta$. Given different values of $x_i$ and $y_i$, the points with highest probability lies on $y_i = 1 + 1.5x_i$.



## ④ POSTERIOR DISTRIBUTION

INDEPENDENT

a. $f_{Y\beta X}(Y,\beta,X) = f_{Y|\beta X}(y|\beta X)\, f_{\beta X}(b,x)$

$= f_{Y|\beta X}(y|\beta X)\, f_{\beta}(b) \cdot f_X(x)$

b. $f_{Y\beta X}(Y,\beta,X) = f_{\beta|YX}(b|YX) \cdot f_{YX}(y,x)$

$$f_{\beta|YX} = \frac{f_{Y\beta X}(Y,\beta,X)}{f_{YX}(y,x)} =$$

$$f_{\beta|YX} = \frac{f_{Y|\beta X}(y|\beta X)\, f_{\beta}(b) \cdot f_X(x)}{f_{Y|X}(y|x) \cdot f_k(x)} \propto f_{Y|\beta X} \cdot f_{\beta}$$

BAYES ESTIMATOR

$$\hat{\beta}^{BAYES} = \operatorname{argmax} f_{Y|\beta X} \cdot f_{\beta}$$

$$\left(\prod_{i=1}^{m} \phi(y_i - x_i^T\beta)\right)\left(\frac{1}{\lambda^d}\prod_{j=1}^{d} \phi\left(\frac{\beta_j}{\lambda}\right)\right)$$

$$\left(\prod_{i=1}^{m} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(y_i - x_i^T\beta)}\right)\left(\frac{1}{\lambda^d}\prod_{j=1}^{d} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\left(\beta_j/\lambda\right)^2}\right) \quad \prod \to \Sigma$$

$$\left(\frac{1}{\sqrt{2\pi}}\right)^m \exp\left\{-\frac{1}{2}\Sigma(y_i - x_i^T\beta)^2\right\}\left(\frac{1}{\sqrt{2\pi}\lambda}\right)^d \exp\left\{-\frac{1}{2}\sum_{j=1}^{d}\left(\beta_j/\lambda\right)^2\right\}$$

$$\log\left(\frac{1}{\sqrt{2\pi}}\right)^n - \frac{1}{2}\sum(y_i - x_i^T\beta)^2 + \log\left(\frac{1}{\sqrt{2\pi}\lambda}\right)^d - \frac{1}{2}\sum_{j=1}^{d}(\beta_j/\lambda)^2$$

$\max$ LIK $= \max(\log(\text{lik}))$ WHY? LIKELIHOOD POSITIVE

$$= \max\left\{-\frac{1}{2}\sum(y_i - x_i^T\beta)^2 - \frac{1}{2}\sum_{j=1}^{d}(\beta_j/\lambda)^2\right\}$$

$$= \min\left\{\frac{1}{2}\sum(y_i - x_i^T\beta)^2 + \frac{1}{2}\sum_{j=1}^{d}(\beta_j/\lambda)^2\right\}$$

$$= \min\left\{\frac{1}{2}\sum(y_i - x_i^T\beta)^2 + \frac{1}{2\lambda^2}\beta^T\beta\right\}$$ PENALIZATION

FOC $\quad \sum x_i(y_i - x_i^T\beta) + \frac{1}{2\lambda^2}2I_d\beta = 0$

$\frac{x^TAx}{dx} = 2Ax$

$$\left(\sum x_i x_i^T + \frac{1}{\lambda^2}I_d\right)\beta = \sum x_i y_i$$

$$\hat{\beta}^{\text{BAYES}} = \left(\sum x_i x_i^T + \frac{1}{\lambda^2}I_d\right)^{-1}\left(\sum x_i y_i\right)$$

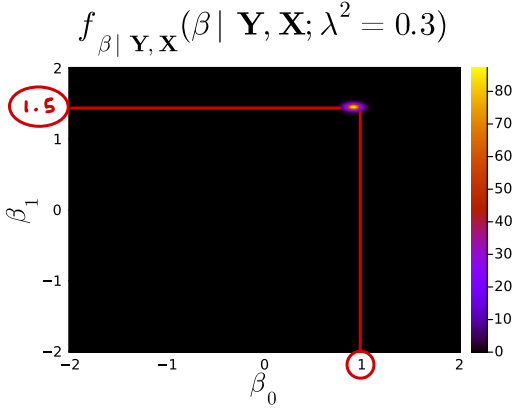
$\lambda\downarrow$, PENALIZATION TERM BIGGER. $\lambda\downarrow$, MORE CONFIDENT ABOUT MY PRIOR WHICH SHRINKS PARAMETERS TO ZERO

$\therefore$ I WILL PENALIZE MORE DEVIATIONS FROM ZEROS.

$\beta^T\beta = \|\beta\|_{L^2}^2$ PENALIZING ITS SQUARE EUCLIDEAN NORM.

# Ridge Regression - Posterior Distribution

The posterior distribution, $\beta \mid \mathbf{Y}, \mathbf{X} \sim \mathcal{N}(\dot{m}, \dot{Q})$, belongs to the same family of probability distributions as the prior when combined with the likelihood function $\Longrightarrow$ the prior and posterior distributions are known as conjugate distributions.



# Formalization Bayesian Inference

Chain's Rule:

$$f_{\mu\mathbf{Y}}(\mu, \mathbf{Y}) = f_{\mu|\mathbf{Y}}(\mu|\mathbf{Y})\ f_{\mathbf{Y}}(\mathbf{y})$$
$$f_{\mu\mathbf{Y}}(\mu, \mathbf{Y}) = f_{\mathbf{Y}|\mu}(\mathbf{Y}|\mu)\ f_{\mu}(\mu)$$

Bayes' Rule:

$$f_{\mu|\mathbf{Y}}(\mu|\mathbf{Y}) = \frac{f_{\mathbf{Y}|\mu}(\mathbf{y}|\mu)}{f_{\mathbf{Y}}(\mathbf{y})}\ f_{\mu}(\mu) \propto f_{\mathbf{Y}|\mu}(\mathbf{y}|\mu)\ f_{\mu}(\mu)$$

Sample mean case:

- $\{y_1, \ldots, y_n\}$ iid sample with $y_i \sim \mathcal{N}(\mu, \sigma^2)$ and $\sigma^2$ known.
- $\mu \sim \mathcal{N}(m, Q)$
- $\mu|\mathbf{Y} \sim$ ?

# Posterior Distribution $\mu|\mathbf{Y}$

Posterior distribution:

$$f_{\mu|\mathbf{Y}}(\mu|\mathbf{Y}) \propto \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} exp\Big\{-\frac{1}{2\sigma^2}(y_i-\mu)^2\Big\} \cdot \frac{1}{\sqrt{2\pi Q}} exp\Big\{-\frac{1}{2Q}(\mu-m)^2\Big\}$$

$$f_{\mu|\mathbf{Y}}(\mu|\mathbf{Y}) \propto (2\pi\sigma^2)^{-\frac{n}{2}} exp\Big\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\bar{y}+\bar{y}-\mu)^2\Big\} \cdot \frac{1}{\sqrt{2\pi Q}} exp\Big\{-\frac{1}{2Q}(\mu-m)^2\Big\}$$

$$f_{\mu|\mathbf{Y}}(\mu|\mathbf{Y}) \propto (2\pi\sigma^2)^{-\frac{n}{2}} exp\Big\{-\frac{1}{2\sigma^2}\Big[n(\bar{y}-\mu)^2+\sum_{i=1}^{n}(y_i-\bar{y})^2+2(\bar{y}-\mu)\sum_{i=1}^{n}(y_i-\bar{y})\Big]\Big\} \cdot \frac{1}{\sqrt{2\pi Q}} exp\Big\{-\frac{1}{2Q}(\mu-m)^2\Big\}$$

$$f_{\mu|\mathbf{Y}}(\mu|\mathbf{Y}) \propto (2\pi\sigma^2)^{-\frac{n}{2}} exp\Big\{-\frac{n}{2\sigma^2}(\bar{y}-\mu)^2\Big\} \cdot (2\pi\sigma^2)^{-\frac{n}{2}} exp\Big\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\bar{y})^2\Big\} \cdot \frac{1}{\sqrt{2\pi Q}} exp\Big\{-\frac{1}{2Q}(\mu-m)^2\Big\}$$

$$f_{\mu|\mathbf{Y}}(\mu|\mathbf{Y}) \propto exp\Big\{-\frac{n}{2\sigma^2}(\bar{y}-\mu)^2-\frac{1}{2Q}(\mu-m)^2\Big\} = exp\Big\{-\frac{n}{2\sigma^2}(\bar{y}^2+\mu^2-2\bar{y}\mu)-\frac{1}{2Q}(\mu^2+m^2-2\mu m)\Big\}$$

$$f_{\mu|\mathbf{Y}}(\mu|\mathbf{Y}) \propto exp\Big\{-\frac{1}{2}\Big[\mu^2\Big(\frac{n}{\sigma^2}+\frac{1}{Q}\Big)+m^2\Big(\frac{1}{Q}\Big)-2\mu\Big(\frac{n}{\sigma^2}\bar{y}+\frac{1}{Q}m\Big)\Big]\Big\} \cdot exp\Big\{-\frac{1}{2}\Big(\bar{y}^2\frac{n}{\sigma^2}\Big)\Big\}$$

$$f_{\mu|\mathbf{Y}}(\mu|\mathbf{Y}) \propto exp\Big\{-\frac{1}{2\dot{Q}}(\mu-\dot{m})^2\Big\} \implies \mu|\mathbf{Y} \sim \mathcal{N}(\dot{m},\dot{Q})$$

KERNEL

Posterior moments:

$$-\frac{1}{2\dot{Q}}\mu^2 = -\frac{1}{2}\mu^2\Big(\frac{n}{\sigma^2}+\frac{1}{Q}\Big) \implies \frac{1}{\dot{Q}} = -\frac{1}{2}\mu^2\Big(\frac{n}{\sigma^2}+\frac{1}{Q}\Big) \implies \dot{Q} = [(\sigma^2/n)^{-1}+Q^{-1}]^{-1}$$

$$\frac{1}{2\dot{Q}}2\mu\dot{m} = \frac{1}{2}2\mu\Big(\frac{n}{\sigma^2}\bar{y}+\frac{1}{Q}m\Big) \implies \frac{\dot{m}}{\dot{Q}} = \frac{n}{\sigma^2}\bar{y}+\frac{1}{Q}m \implies \dot{m} = \dot{Q}[(\sigma^2/n)^{-1}\bar{y}+Q^{-1}m]$$

# Bayesian Inference

$$\dot{m} = \left(\frac{Q^{-1}}{Q^{-1} + (\sigma^2/n)^{-1}}\right) m + \left(\frac{(\sigma^2/n)^{-1}}{Q^{-1} + (\sigma^2/n)^{-1}}\right) \bar{y}$$

What happens when $n \to \infty$? And when $Q \to \infty$?

Under a quadratic loss function, the bayesian estimate of $\mu$ that minimizes the posterior expected loss is the mean of the posterior distribution $\dot{m}$:

$$E_{\mu|\mathbf{Y}}[(\mu - \hat{\mu})^2|\mathbf{Y}] = E_{\mu|Y}\left[(\mu - \dot{m} + \dot{m} - \hat{\mu})^2|Y\right]$$

$$= E_{\mu|Y}\left[(\mu - \dot{m})^2|Y\right] + E_{\mu|Y}\left[(\dot{m} - \hat{\mu})^2|Y\right] + 2(\dot{m} - \hat{\mu}) E_{\mu|Y}\left[(\mu - \dot{m})|Y\right]$$

$$= E_{\mu|Y}\left[(\mu - \dot{m})^2|Y\right] + E_{\mu|Y}\left[(\dot{m} - \hat{\mu})^2|Y\right]$$

MINIMIZED FOR $\hat{\mu} = \dot{m} = E_{\mu|Y}[\mu|Y]$

# Link Bayesian and Frequentist Inference

Bernstein-von Mises Theorem: under some regularity conditions, given $\tilde{\theta}$ with the posterior distribution, we have:

$$\tilde{\theta} \xrightarrow{p} \hat{\theta}^{MLE}$$

$$\sqrt{N}(\tilde{\theta} - \hat{\theta}^{MLE}) \xrightarrow{d} \mathcal{N}(0, Var(\hat{\theta}^{MLE}))$$

The most important implication of the Bernstein–von Mises theorem is that the Bayesian inference is asymptotically correct from a frequentist point of view.

# Bayesian Linear Regression

- Previous result generalizes to linear regression case: $y_i = x_i^T \beta + u_i$ with $u_i \sim \mathcal{N}(0, \sigma^2)$ and $\sigma^2$ assumed to be known.

- Assume gaussian prior distribution: $f_\beta(\beta; \sigma^2) = \mathcal{N}(m, \sigma^2 Q)$.

- We get posterior distribution: $f_{\beta|\mathbf{Y},\mathbf{X}}(\beta|\mathbf{Y},\mathbf{X}; \sigma^2) = \mathcal{N}(\dot{m}, \sigma^2 \dot{Q})$ where the moments of posterior distribution are:

  (i) $\dot{Q} = \left(Q^{-1} + \hat{Q}_n^{-1}\right)^{-1}$ : INVERSE OF THE SUM OF THE PRECISION MATRICES

  (ii) $\dot{m} = \dot{Q}\left(Q^{-1} m + \hat{Q}_n^{-1} \hat{\beta}^{OLS}\right)$

  (iii) $\hat{Q}_n = \left(\sum_{i=1}^n x_i x_i^T\right)^{-1}$ $\approx$ $\sigma^2 X'X$

- Now compare $\dot{m}$ with the result from the ridge regression exercise.