

ECON220B Discussion Section 3

Consistency and Ridge Regression

Lapo Bini

UC San Diego

March 18, 2026

ROADMAP

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

WHAT ARE WE ESTIMATING?

Suppose you have a collection of data $\{y_i, w_i, \mathbf{x}_i\}_{i=1}^n$ and we estimate a linear regression. How do we **interpret** the estimated OLS coefficients?

► Causal Estimate

Economic assumption $E[u_i | w_i, \mathbf{x}_i] = 0$ + true **DGP assumption** $E[y_i | w_i, \mathbf{x}_i] = \beta w_i + \mathbf{x}_i' \boldsymbol{\gamma}$ guarantees $E[\hat{\beta}] = \beta = \frac{\partial}{\partial w_i} E[y_i | w_i, \mathbf{x}_i]$, that is β reflects **ceteris paribus causal effects**.

► Covariance

If $E[u_i w_i] = 0$ (**economic assumption**) then $\hat{\beta} \xrightarrow{p} \beta$ which express the **covariance** between w and y .

► Best Predictor

If $E[u_i | \mathbf{x}_i] \neq 0$, then $\hat{\beta} \xrightarrow{p} \delta \equiv \beta + \Delta$ it converges to the coefficient of the **linear projection**.

$$\delta := \arg \min_{b \in \mathbb{R}} E[(y_i - b w_i - \mathbf{x}_i' \boldsymbol{\gamma})^2]$$

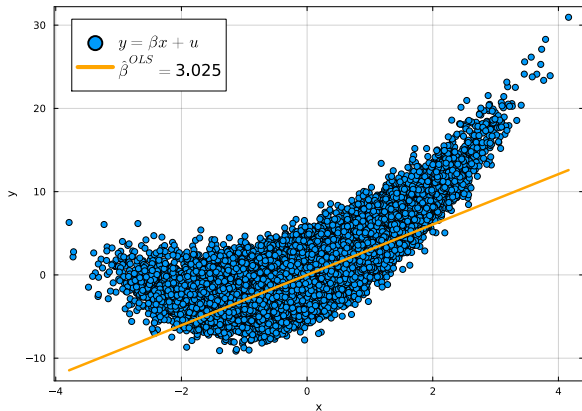
EXERCISE: MODEL MISSPECIFICATION

Suppose the **true DGP** for y_i is $y_i = \beta x_i + x_i^2 + \eta_i$, with $\beta = 3$, and $x_i \sim \mathcal{N}(0, 1)$, $\eta_i \sim \mathcal{N}(0, 4)$, $E[\eta_i|x_i] = 0$. Now suppose that instead we run the following linear regression: $y_i = \beta x_i + u_i$

- (1) What is the direct causal effect of x on y ?
- (2) Suppose we estimate the model by OLS. Can we apply the Markov-Gauss theorem?
- (3) Is $\hat{\beta}^{OLS}$ consistent for the true β ? If yes, propose an intuition.
- (4) Is $\hat{\beta}^{OLS}$ consistent for the true β , if we assume that the true model is $y_i = \beta x_i + x_i^3 + \eta_i$?
- (5) Now suppose the true model is $y_i = \beta x_i + \eta_i$, but instead of running a regression of y_i on x_i , you run the regression of x_i and y_i : $x_i = \phi y_i + \nu_i$, that is, you switch the dependent and independent variables. What is $\hat{\phi}^{OLS}$ estimating? Find the value for the probability limit of $\hat{\phi}^{OLS}$.

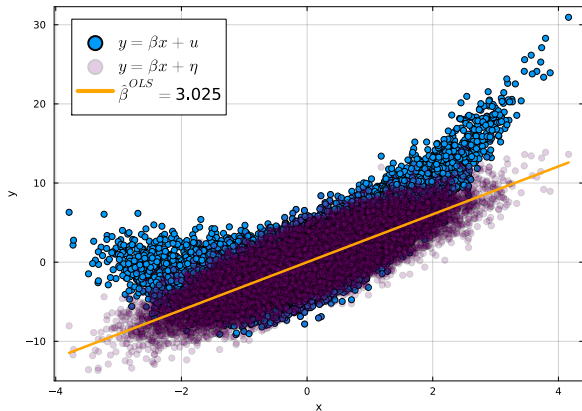
EXAMPLE (1/4)

Suppose the **true DGP** for y_i is $y_i = \beta x_i + x_i^2 + \eta_i$, with $\beta = 3$, and $x_i \sim \mathcal{N}(0, 1)$, $\eta_i \sim \mathcal{N}(0, 4)$, $E[\eta_i|x_i] = 0$.



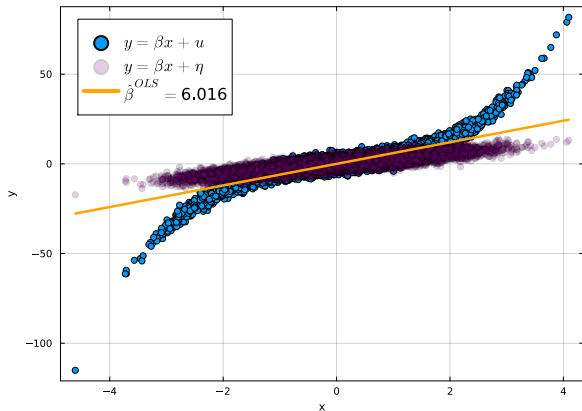
EXAMPLE (2/4)

Suppose the **true DGP** for y_i is $y_i = \beta x_i + x_i^2 + \eta_i$, with $\beta = 3$, and $x_i \sim \mathcal{N}(0, 1)$, $\eta_i \sim \mathcal{N}(0, 4)$, $E[\eta_i|x_i] = 0$



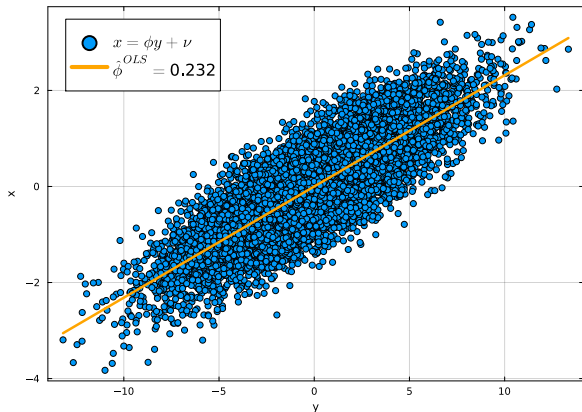
EXAMPLE (3/4)

Suppose the **true DGP** for y_i is $y_i = \beta x_i + x_i^3 + \eta_i$, with $\beta = 3$, and $x_i \sim \mathcal{N}(0, 1)$, $\eta_i \sim \mathcal{N}(0, 4)$, $E[\eta_i|x_i] = 0$



EXAMPLE (4/4)

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



FREQUENTISTS VS BAYESIANS

Given $\{y_i\}_{i=1}^n$, $y_i \sim \text{iid } \mathcal{N}(\mu, \sigma^2)$ we are interested in the population mean μ . We know that:

$$\hat{\mu}^{MLE} = n^{-1} \sum_{i=1}^n y_i \sim \mathcal{N}(\mu, \sigma^2/n)$$

1. Frequentist

The data is the result of sampling from a random process. Frequentists see the data as varying and the parameter μ 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

μ treated as a random variable. Bayesians have prior beliefs about μ (**prior distribution**), which is updated after observing the data (**likelihood function**) using **Bayes' Rule**. The **posterior distribution** summarises the uncertainty about credible values of μ .

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:

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

POSTERIOR DISTRIBUTION $\mu|Y$

Posterior distribution:

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

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

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

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

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

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

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

Posterior moments:

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

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

RIDGE REGRESSION

Framework

- ▶ Consider the following **linear regression model** $y_i = \mathbf{x}_i' \boldsymbol{\beta} + u_i$, $u_i \sim \mathcal{N}(0, 1)$.
- ▶ We have the **prior beliefs** that the parameters $\boldsymbol{\beta} \in \mathbb{R}^d$ follow the distribution $\boldsymbol{\beta} \sim \mathcal{N}(0, \lambda^2 \mathbb{I}_d)$ where $\lambda > 0$ and $\mathbb{I}_d$ is the $(d \times d)$ identity matrix.
- ▶ Lastly, assume that $u_i, \mathbf{x}_i, \boldsymbol{\beta}$ are mutually independent.

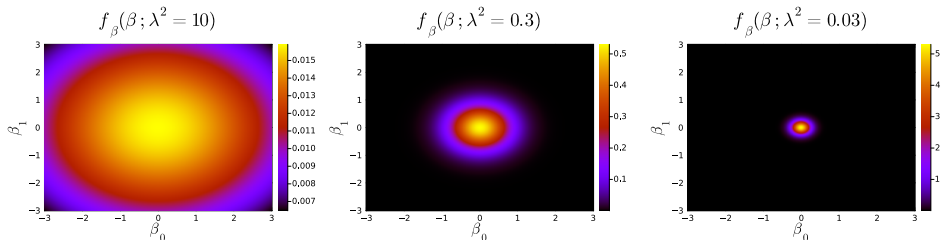
Questions

- (1) Prove that $f_{\boldsymbol{\beta}}(\boldsymbol{\beta}) = \lambda^{-d} \prod_{j=1}^d \phi(\beta_j/\lambda)$, where $\phi(z)$ is the pdf of a random variable z distributed as a standard normal.

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 λ^2 represents the uncertainty of our guess, i.e. $\boldsymbol{\beta} \sim \mathcal{N}(0, \lambda^2 \mathbb{I}_2)$.

Figure: Prior distribution for different values of λ^2 .



RIDGE REGRESSION

Framework

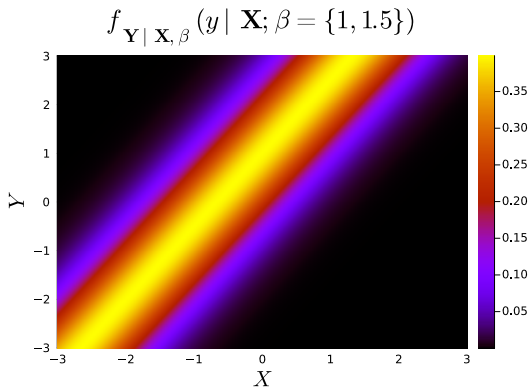
- ▶ Consider the following **linear regression model** $y_i = \mathbf{x}_i' \boldsymbol{\beta} + u_i$, $u_i \sim \mathcal{N}(0, 1)$.
- ▶ We have the **prior beliefs** that the parameters $\boldsymbol{\beta} \in \mathbb{R}^d$ follow the distribution $\boldsymbol{\beta} \sim \mathcal{N}(0, \lambda^2 \mathbb{I}_d)$ where $\lambda > 0$ and $\mathbb{I}_d$ is the $(d \times d)$ identity matrix.
- ▶ Lastly, assume that $u_i, \mathbf{x}_i, \boldsymbol{\beta}$ are mutually independent.

Questions

(2) Show that $f_{\mathbf{Y}|\boldsymbol{\beta}, \mathbf{X}}(y_1, \dots, y_n | \boldsymbol{\beta}, \mathbf{X}) = \prod_{i=1}^n \phi(y_i - \mathbf{x}_i' \boldsymbol{\beta})$.

RIDGE REGRESSION - LIKELIHOOD FUNCTION

The **likelihood** describes the probability of the data already observed given certain parameter values β . Given different values of x_i and y_i , the points with the highest probability lie on $y_i = \beta_0 + \beta_1 x_i$.



RIDGE REGRESSION

Framework

- ▶ Consider the following **linear regression model** $y_i = \mathbf{x}_i' \boldsymbol{\beta} + u_i$, $u_i \sim \mathcal{N}(0, 1)$.
- ▶ We have the **prior beliefs** that the parameters $\boldsymbol{\beta} \in \mathbb{R}^d$ follow the distribution $\boldsymbol{\beta} \sim \mathcal{N}(0, \lambda^2 \mathbb{I}_d)$ where $\lambda > 0$ and $\mathbb{I}_d$ is the ($d \times d$) identity matrix.
- ▶ Lastly, assume that $u_i, \mathbf{x}_i, \boldsymbol{\beta}$ are mutually independent.

Questions

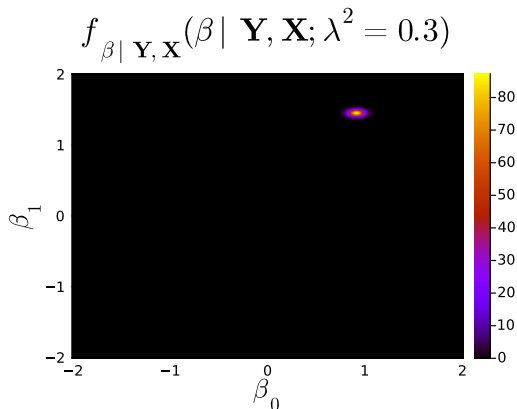
- (3) Derive the Maximum Likelihood Estimator $\hat{\boldsymbol{\beta}}^{MLE}$.
- (4) Find the posterior distribution $f_{\boldsymbol{\beta}|\mathbf{Y},\mathbf{X}}(\boldsymbol{\beta}|\mathbf{Y},\mathbf{X})$ and derive the Bayes estimator defined as

$$\hat{\boldsymbol{\beta}}^{Bayes} := \arg \max_b f_{b|\mathbf{Y},\mathbf{X}}(b|\mathbf{Y},\mathbf{X})$$

- (5) Is $\hat{\boldsymbol{\beta}}^{Bayes}$ consistent?

RIDGE REGRESSION - POSTERIOR DISTRIBUTION

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



EXTRA: LINK BAYESIAN AND FREQUENTIST INFERENCE

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

$$\begin{aligned}\tilde{\theta} &\xrightarrow{P} \hat{\theta}^{MLE} \\ \sqrt{N}(\tilde{\theta} - \hat{\theta}^{MLE}) &\xrightarrow{d} \mathcal{N}(0, \text{Var}(\hat{\theta}^{MLE}))\end{aligned}$$

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

EXTRA: 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 σ^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) $\dot{Q} = (Q^{-1} + \hat{Q}_n^{-1})^{-1}$
 - (ii) $\dot{m} = \dot{Q} (Q^{-1} m + \hat{Q}_n^{-1} \hat{\beta}^{OLS})$
 - (iii) $\hat{Q}_n = (\sum_{i=1}^n x_i x_i^T)^{-1}$
- ▶ Now compare $\dot{m}$ with the result from the ridge regression exercise.