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 $\{u_1, \dots, u_n\}$ 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 $x \rightarrow y$, 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)

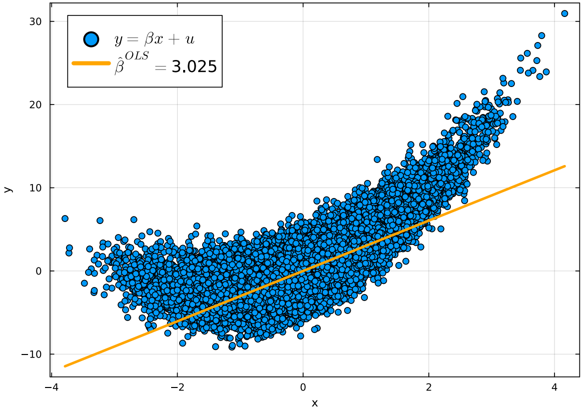
Model: $y_i = \beta x_i + u_i \ u_i = x_i^2 + \eta_i$ with true parameter $\beta = 3$, and $x_i \sim \mathcal{N}(0,1)$, $\eta_i \sim \mathcal{N}(0,4)$, $x_i \perp \eta_i$.

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

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

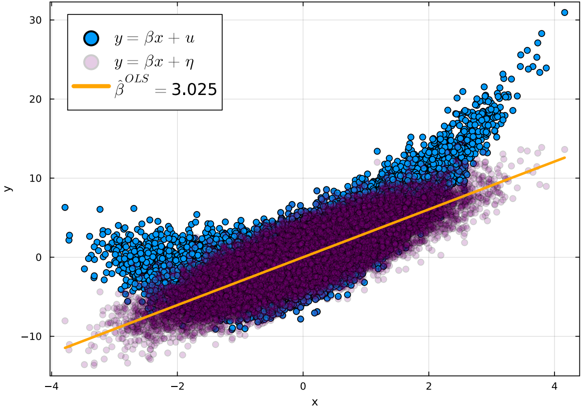
Example (1/2)

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



Example (1/2)

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



Example (2/2)

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

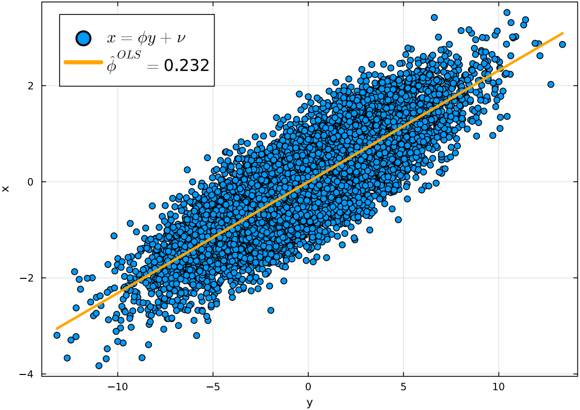
(1) Suppose that instead of running a regression of y_i on x_i, you run the regression of x_i and y_i, that is you switch the dependent and independent variables:

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

What is $\hat{\phi}^{OLS}$ estimating?

Example (2/2)

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



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 δ 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 $E[u_i] = 0$.

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 β_1?

Two Religions: Frequentists vs Bayesians

Given $\{y_1, \ldots, y_n\}$ iid sample with $y_i \sim \mathcal{N}(\mu, \sigma^2)$ we are interested in the population mean μ. 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 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 treates 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.

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 I_d is the $(d\mathrm{x}d)$ identity matrix.
- Lastly, assume that u_i, x_i, β are mutually independent.

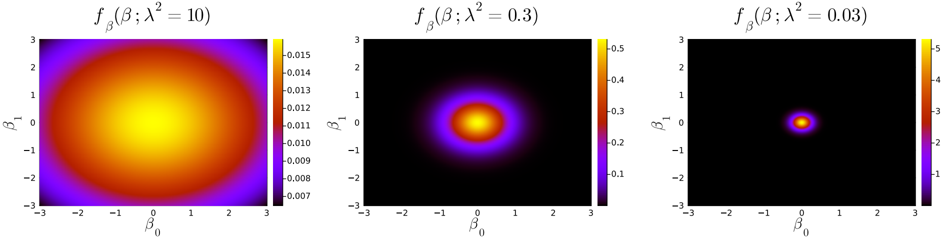
(1) Prove that $f_\beta(\beta) = \lambda^{-d} \prod_{j=1}^{d} \phi\big(\beta_j/\lambda\big)$.
(2) Show that $f_{\mathbf{Y}|\beta,\mathbf{X}}(y_1, \ldots, y_n|\beta, \mathbf{X}) = \prod_{i=1}^{n} \phi(y_i - x_i^T \beta)$.
(3) Derive the Maximum Likelihood Estimator $\hat{\beta}^{MLE}$.
(4) Find the posterior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

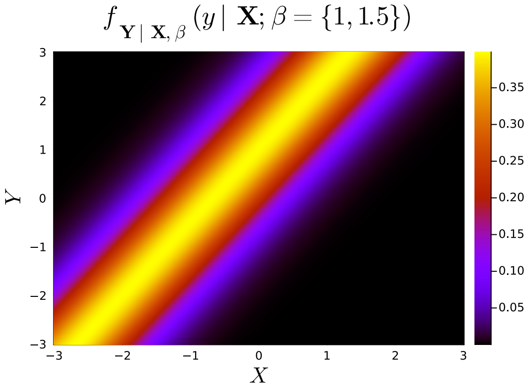
Before observing the data, our prior belief is that the parameters are most likely to be close to zero. The parameter λ^2 represents the uncertainty of our guess, i.e. $\beta \sim \mathcal{N}(0, \lambda^2 \mathrm{I}_2)$.

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



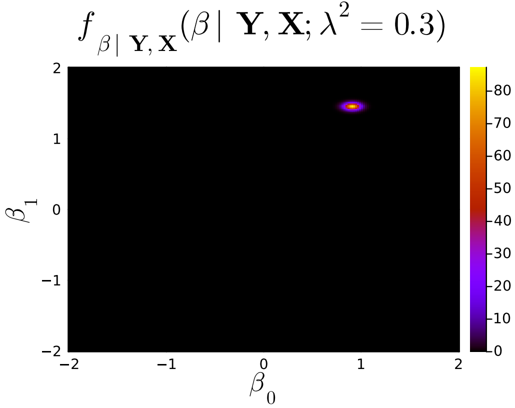
Ridge Regression - Likelihood Function

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



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 σ^2 known.
- $\mu \sim \mathcal{N}(m, Q)$
- $\mu|\mathbf{Y} \sim ?$

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

Posterior distribution:

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

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

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

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

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

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

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

Posterior moments:

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

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

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 μ that minimizes the posterior expected loss is the mean of the posterior distribution $\dot{m}$:

$E_{\mu|\mathbf{Y}}[(\mu - \hat{\mu})^2|\mathbf{Y}] =$

Link Bayesian and Frequentist Inference

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

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

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

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

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

  (iii) $\hat{Q}_n = \left(\sum_{i=1}^n x_i x_i^T\right)^{-1}$

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