Microeconometrics

Introduction to non-linear models

Alex Armand
alex.armand@novasbe.pt

Summary

1. **Nonlinear models and M-estimators**
   1. Introduction
   2. Consistency
2. **Maximum Likelihood**
   1. Consistency of MLE
   2. Asymptotic Distribution
   3. MLE Testing

Estimators as closed form solutions

- In previous topics, all estimators can be written as **closed form** functions of the data.
    - Observed data $\Rightarrow$ mathematical rule for estimates
    - Example: OLS

$$\hat{\beta}_{OLS} = \left(\sum_{i=1}^{N} x_i' x_i\right)^{-1} \left(\sum_{i=1}^{N} x_i' y_i\right)$$

- **Estimators in closed form do not cover all cases of interest**
    - **CASE 1**: we are interested in other conditional moments of y
    - **CASE 2**: the conditional mean of y is a non-linear function

Conditional median instead of conditional mean

CASE 1: we are interested in conditional medians

$$\begin{aligned} y &= \alpha + \mathrm{X}\beta + u \\ Med(y|\mathrm{x}) &= \mathrm{x}\beta = \beta_1 + \beta_2 x_2 + ... + \beta_K x_K \end{aligned}$$

- OLS does not consistently estimate β_j unless we make further assumptions
- **Example**: assume $D[u|x]$ is symmetric about zero

$$E[y|x] = med[y|x] = \alpha + \mathrm{x}\beta$$

OLS vs LAD

Consider a linear population model

$$y = \alpha + \mathrm{X}\beta + u$$

1. **OLS** estimator solves

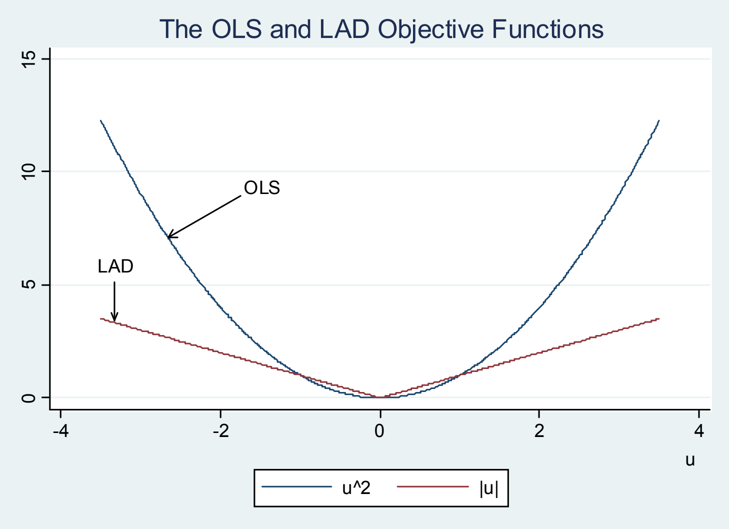
$$\min_{a,\mathrm{b}} \sum_{i=1}^{N} (y_i - a - \mathrm{x}_i \mathrm{b})^2$$

2. **LAD** (least absolute deviations) estimator solves

$$\min_{a,\mathrm{b}} \sum_{i=1}^{N} |y_i - a - \mathrm{x}_i \mathrm{b}|$$

   - A solution cannot generally be written in closed form

Minimization functions



Beyond LAD: quantile regression

Want to know the effect of changing a covariate on **features of the distribution** other than the mean

- Mean effects may mask very different effects in different parts of the distribution of the outcome variable
- **Example**: effect of a particular kind of pension plan policy intervention
  - **OLS**: the effect of the pension plan on the (conditional) mean of total wealth
  - **LAD**: the effect of the pension plan on the (conditional) median of total wealth
  - **Quantile**: the effect of the pension plan on the (conditional) quantile of total wealth

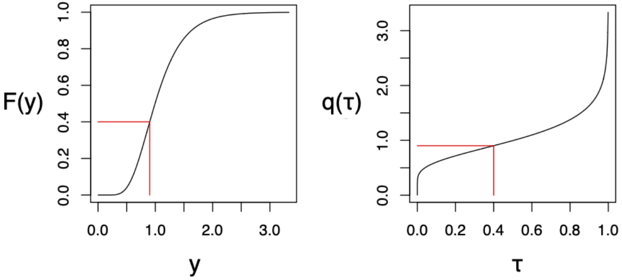
Quantile of a random variable y

Definition

For $0 < \tau < 1$, $q(\tau)$ is the τ^{th} quantile of y_i if

$$P[y_i \leq q(\tau)] \geq \tau \text{ and } P[y_i \geq q(\tau)] \geq 1 - \tau$$

- Assume y_i is continuous with strictly increasing cdf $\Rightarrow q(\tau)$ is a unique value



Estimation of quantiles

1. Order statistic $y_{(1)} < .. < y_{(n)}$
2. The τ^{th} quantile $q_\tau(y)$ is given by

$$q_\tau(y) = arg\ min_a E[\rho_\tau(y-a)]$$

where $\rho_\tau(u) = (\tau - \mathbb{1}[u < 0])u$ is the **check function**

- **Example**: find the median of y

$$\begin{aligned} q_{0.5}(y) &= arg\ min_a E[\rho_{0.5}(y-a)] \\ &= arg\ min_a E[0.5 - \mathbb{1}[y < a]] \end{aligned}$$

Proof of check function

Check function.

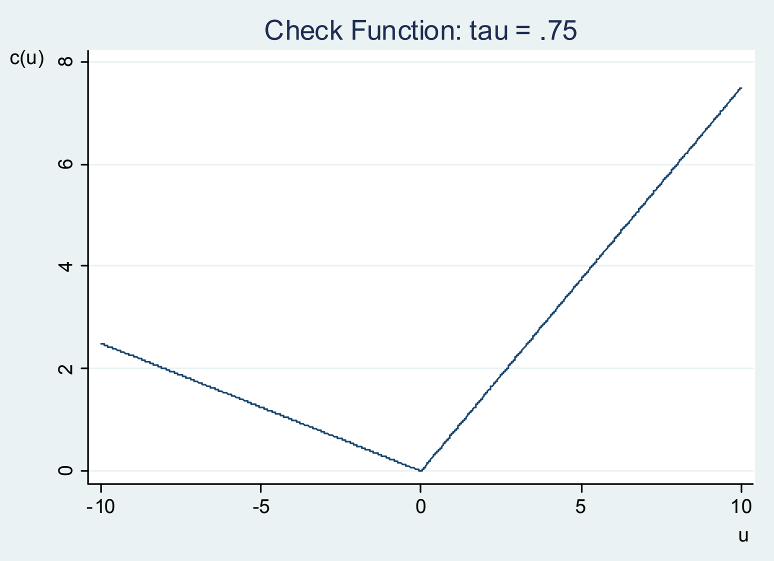
$$E[\rho_\tau(y-a)] = \tau[E(y)-a] - \int_{-\infty}^{b} (y-a) f_y(y)\, dy$$

This function is differentiable with

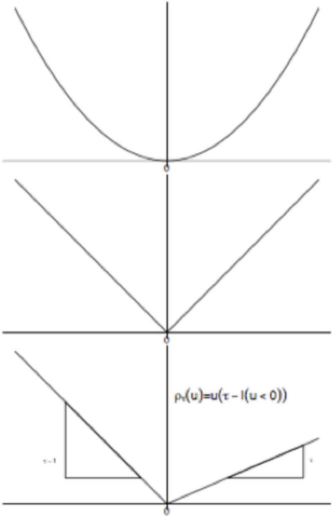
$$\begin{aligned} \frac{\partial E[\rho_\tau(y-a)]}{\partial a} &= -\tau - (a-a) f_y(a) + \int_{-\infty}^{b} f_y(y)\, dy \\ &= F_y(a) - \tau \end{aligned}$$

This function is increasing and reaches its maximum at $q_\tau(y)$. □

Check function: example 75th percentile



Comparison OLS vs LAD vs quantile



Given a sample $\{Y_1, \ldots, Y_n\}$ from a single distribution F, it can be shown that:

Sample mean

$\overline{Y} = \arg\min_\xi \sum_i (Y_i - \xi)^2$

Sample median

$\hat{Q}_Y(0.5) = \arg\min_\xi \sum_i |Y_i - \xi|$

Sample τth quantile

$\hat{Q}_Y(\tau) = \arg\min_\xi \sum_i \rho_\tau(Y_i - \xi)$

Quantile regression

- Assume linearity
- Let covariates affect quantiles $\Rightarrow$ estimates depend on τ

$$Quant_\tau(y_i|x_i) = \alpha(\tau) + \mathrm{x_i}\beta(\tau)$$

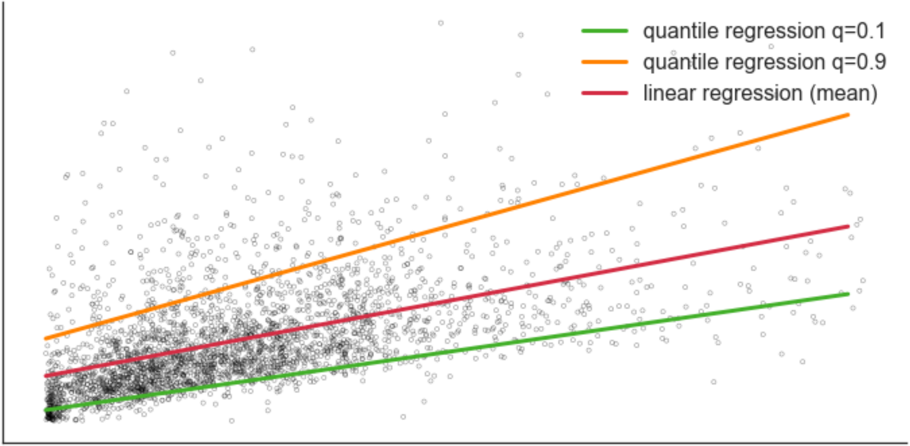
- $\alpha(\tau)$ and $\beta(\tau)$ are obtained by minimizing the *check* function

$$\min_{\alpha,\beta} \sum_{i=1}^{N} \rho_\tau(y_i - \alpha - \mathrm{x}_i\beta) = \min_{\alpha,\beta} \sum_{i=1}^{N} (\tau - \mathbb{1}[y_i - \alpha - \mathrm{x}_i\beta < 0])$$

- How to interpret $\alpha(\tau)$ and $\beta(\tau)$?

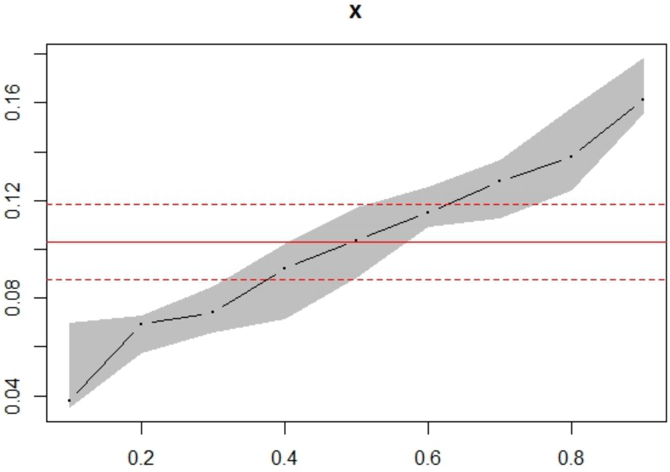
Quantile regression: an example

Apply quantile regression at different quantiles



Quantile regression: an example

Plot results to **compare quantile regression with OLS**



Non-linear least squares (NLS)

CASE 2: the conditional mean of y is a non-linear function

$$E(y|\mathrm{x}) = m(\mathrm{x}, \theta_o),\ \mathrm{x} \in \mathcal{X}$$

- θ_o is the $P \times 1$ vector of numbers we are trying to learn about
- Θ is the **parameter space** $\Rightarrow$ set of all parameters values that are candidates for the population value
- In error form:

$$\begin{aligned} y &= m(\mathrm{x}, \theta_o) + u \\ E(u|\mathrm{x}) &= 0 \end{aligned}$$

- ***Analogy principle***: use the sample analog of the population problem

$$\min_{\theta \in \Theta} N^{-1} \sum_{i=1}^{N} [y_i - m(\mathrm{x}_i, \theta)]^2$$

M-estimation principle

- Assume that $\theta_o \in \Theta$ *uniquely* solves

$$\min_{\theta \in \Theta} \; E[q(\mathrm{w}, \theta)]$$

  - w and θ: vector of observable variables and of parameters to estimate
  - $q : \mathcal{W} \times \Theta \to \mathbb{R}$ is a real valued function

- **Consistency of M-estimator:**

$$N^{-1} \sum_{i=1}^{N} q(\mathrm{w}_i, \theta) \xrightarrow{p} E[q(\mathrm{w}, \theta)]$$

$\hat{\theta}$ minimizes (sample average) $\quad$ θ_o minimizes (population average)

- The solution must be **unique**:

$$E\{[q(\mathrm{w}, \theta_0) - q(\mathrm{w}, \theta)]^2\} > 0$$

for all $\theta \in \Theta$, $\theta \neq \theta_o$

Asymptotic distribution: definitions

$$\sqrt{N}(\hat{\theta} - \theta_o) = \left(N^{-1} \sum_{i=1}^{N} \ddot{\mathrm{H}}_i \right)^{-1} \left[-N^{-1/2} \sum_{i=1}^{N} \mathrm{s}(\mathrm{w}_i, \theta_o) \right]$$

1. ***Score*** is the transpose of the gradient

$$\mathrm{s}(\mathrm{w}, \theta) = \nabla_\theta q(\mathrm{w}, \theta)'$$

where

$$\nabla_\theta q(\mathrm{w}, \theta) = \begin{pmatrix} \frac{\partial q(\mathrm{w},\theta)}{\partial \theta_1} & \frac{\partial q(\mathrm{w},\theta)}{\partial \theta_2} & \cdots & \frac{\partial q(\mathrm{w},\theta)}{\partial \theta_P} \end{pmatrix}$$

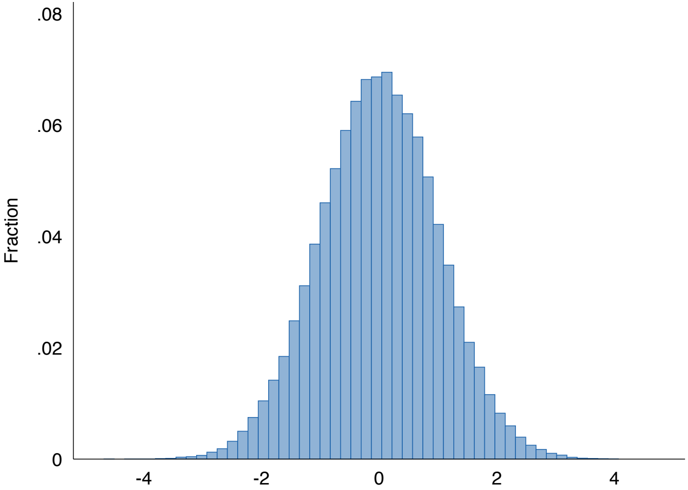
   - Condition $E[\mathrm{s}_i(\theta_o)] = 0$ is often referred to as **Fisher consistency**

2. **Hessian** is the matrix with second derivatives

$$\mathrm{H}(\mathrm{w}, \theta) = \nabla_\theta \mathrm{s}(\mathrm{w}, \theta)$$

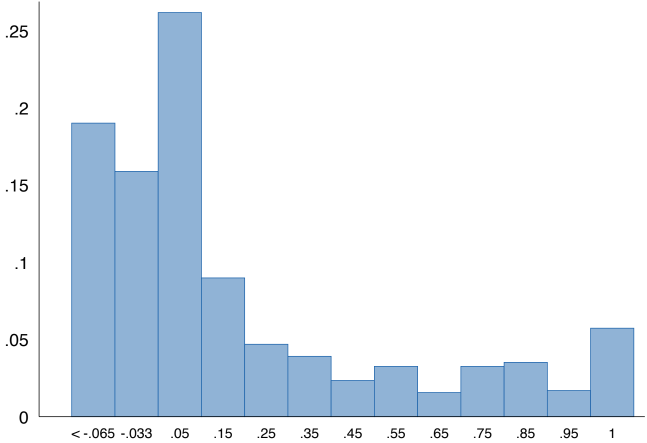
Maximum Likelihood (MLE) approach

- Let's start thinking about an unconditional distribution: population data are generate by a *Normal*(μ, σ^2)? How can we estimate μ, σ^2



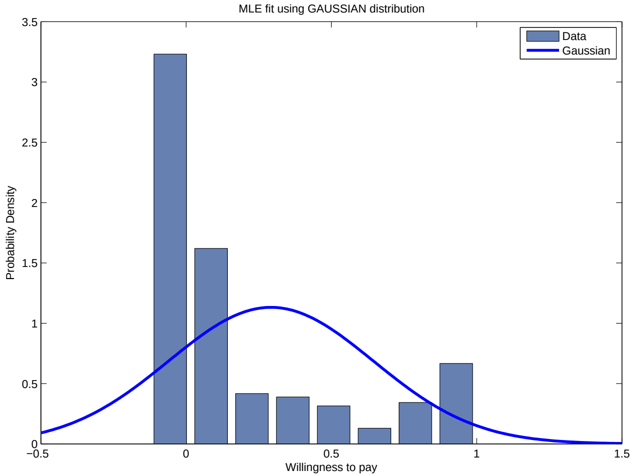
Example of application

- From **empirical distribution** to **theoretical distribution**



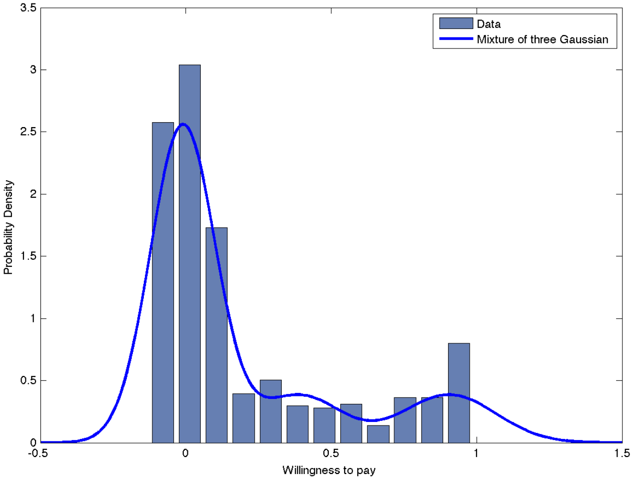
Example of application: normal distribution

- Start assuming a Normal distribution



Example of application: mixture of three Normal distributions

- Use different distributional assumption



Conditional MLE approach

(x_i, y_i) denote a random draw from a population

- Interested in the distribution of y_i conditional on x_i

$$D(y_i|x_i)$$

- We need to assume a conditional density for y

$$f(y|x;\theta),\ y \in \mathcal{Y},\ x \in \mathcal{X},\ \theta \in \Theta$$

- We will allow y_i to have **any characteristic**
    - continuous, discrete, or possibly both features

Conditional MLE estimation

Objective is to maximize the probability of observing the sample as drawn from the assumed density

$$\max_{\theta \in \Theta} \prod_{i=1}^{N} f(\mathrm{y_i} | \mathrm{x_i}; \theta)$$

1. **Log-likelihood function**: $\ell_i(\theta) \equiv \log f(\mathrm{y}_i | \mathrm{x}_i, \theta)$
2. **M-estimation**

$$q(\mathrm{w}_i, \theta) = -\log f(\mathrm{y}_i | \mathrm{x}_i; \theta)$$

   1. Estimator of θ_o solves $\min_{\theta \in \Theta} \; N^{-1} \sum_{i=1}^{N} - \log f(\mathrm{y}_i | \mathrm{x}_i; \theta)$
   2. Unique solution (*Kullback-Leibler Information Inequality*)

$$E[\log f(\mathrm{y}_i | \mathrm{x}_i; \theta_o) | \mathrm{x}_i] \geq E[\log f(\mathrm{y}_i | \mathrm{x}_i; \theta) | \mathrm{x}_i], \text{ all } \theta \in \Theta$$

MLE Testing: Likelihood Ratio Test

Test whether two models are the same

1. **Unconstrained**
2. **Constrained**: imposes some conditions to the unconstrained model (example: one parameter is equal to zero)

- Under correct specification of the density, the LR statistic is:

$$LR = 2(\mathcal{L}_{ur} - \mathcal{L}_r) = 2\left[\sum_{i=1}^{N} \ell_i(\hat{\theta}) - \sum_{i=1}^{N} \ell_i(\tilde{\theta})\right]$$

  - $\hat{\theta}$ is the unrestricted estimator
  - $\tilde{\theta}$ is the estimator with Q restrictions imposed.

- Under H_0, the statistic follows

$$LR \xrightarrow{d} \chi^2_Q$$

An example of MLE application

We saw how to apply conditional MLE, now let's do a step back and apply unconditional MLE

- Unconditional MLE is a simpler version
  - you do not condition on control variables x
- Assume $y \sim Normal(\mu, \sigma^2)$

$$f(y|\mu, \sigma) = [2\pi\sigma^2]^{-1/2} \exp\{-\frac{[y-\mu]^2}{2\sigma^2}\}$$

- Can we use MLE to estimate these parameters and their variance-covariance matrix?

An example of MLE application: procedure

1. Compute log-likelihood function

$$\ell_i(\theta) \equiv -\frac{1}{2}\log\left[2\pi\sigma^2\right] - \frac{[y-\mu]^2}{2\sigma^2}$$

2. Solve minimization problem where $\theta = (\mu, \sigma^2)$

$$\min_{\theta\in\Theta} \; -N^{-1}\sum_{i=1}^{N}\ell_i(\theta)$$

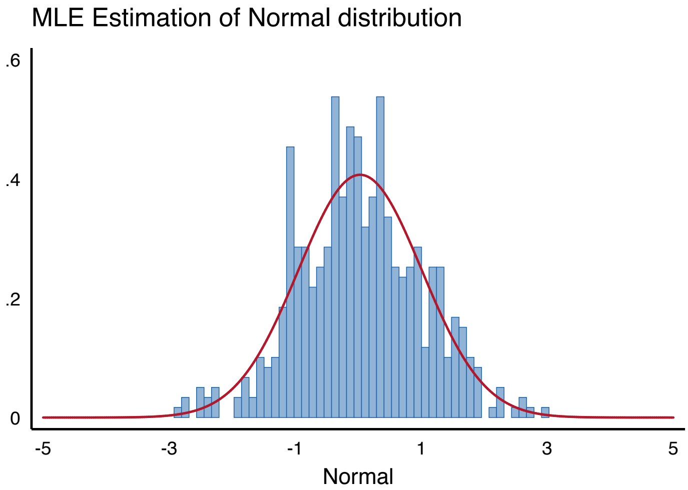
   1. Compute first order conditions to find estimates – **how many?**
   2. Compute second derivatives to compute standard errors

An example of MLE application

- Estimated result and comparison with data generating process

An example of MLE application

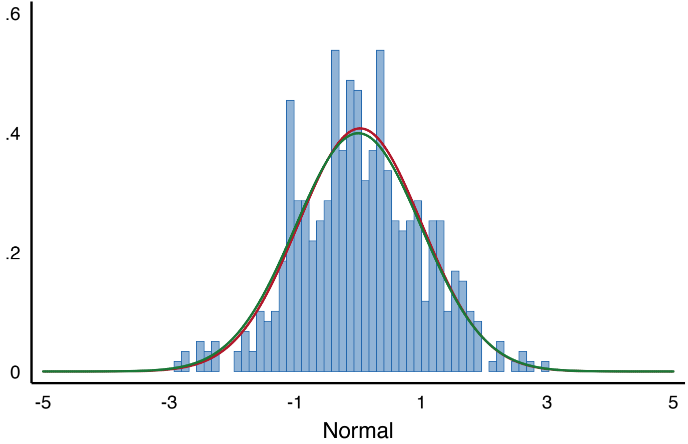
- Estimated result and comparison with data generating process



An example of MLE application

- Estimated result and comparison with data generating process

MLE Estimation of Normal distribution

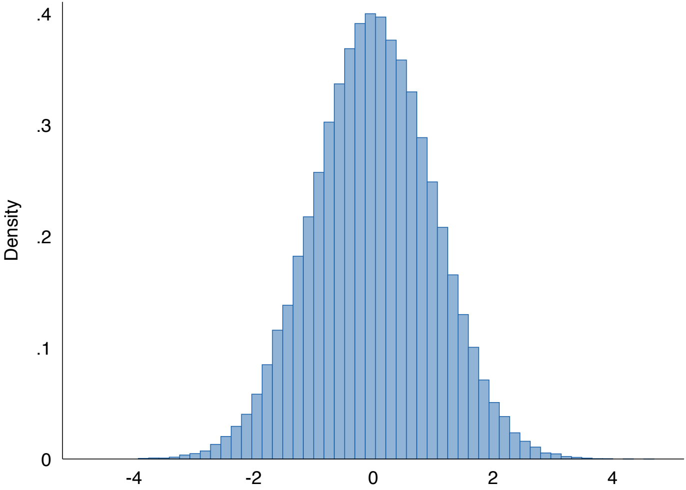


An example of MLE application

- Imagine now a larger sample of y_i , $i = 1, .., 100000$

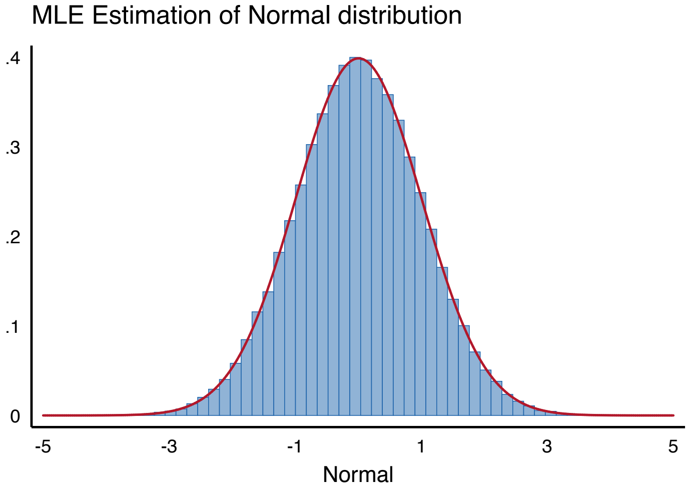
An example of MLE application

- Imagine now a larger sample of y_i , $i = 1, .., 100000$



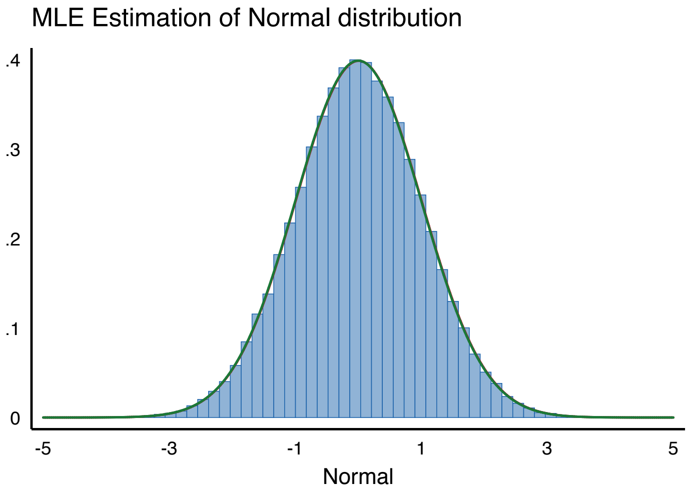
An example of MLE application

- Imagine now a larger sample of y_i , $i = 1, .., 100000$



An example of MLE application

- Imagine now a larger sample of y_i , $i = 1, .., 100000$



OLS - MLE comparison

Assume simple model with 1 regressor and two parameters (α and β)

$$y = \alpha + x\beta + u$$

1. **OLS**: no need to assume a distribution
2. **MLE**: assume a distribution
   - Example: $u \sim Normal(0, \sigma^2)$

$$\ell_i(\theta) \equiv -\frac{1}{2}\log\left[2\pi\sigma^2\right] - \frac{[y - \alpha - x\beta - 0]^2}{2\sigma^2}$$

   - Apply M-estimation to find α and β

EXTRA: Asymptotic distribution for NLS

- By the mean value theorem (for each element m of the score)

$$\sum_{i=1}^{N} s_m(\mathrm{w}_i, \hat{\theta}) = \sum_{i=1}^{N} s_m(\mathrm{w}_i, \theta_o) + \left(\sum_{i=1}^{N} \nabla_\theta s_m(\mathrm{w}_i, \ddot{\theta}_m) \right) (\hat{\theta} - \theta_o)$$

where $\ddot{\theta}_m$ is on the line segment between $\hat{\theta}$ and θ_o

- Stack all P elements to get

$$\sum_{i=1}^{N} \mathrm{s}(\mathrm{w}_i, \hat{\theta}) = \sum_{i=1}^{N} \mathrm{s}(\mathrm{w}_i, \theta_o) + \left(\sum_{i=1}^{N} \ddot{\mathrm{H}}_i \right) (\hat{\theta} - \theta_o)$$

- By multiplying by $N^{-1/2}$ and applying Fisher consistency

$$0 = N^{-1/2} \sum_{i=1}^{N} \mathrm{s}(\mathrm{w}_i, \theta_o) + \left(N^{-1} \sum_{i=1}^{N} \ddot{\mathrm{H}}_i \right) \sqrt{N}(\hat{\theta} - \theta_o)$$

Asymptotic distribution

$$\sqrt{N}(\hat{\theta} - \theta_o) = \left(N^{-1} \sum_{i=1}^{N} \ddot{\mathrm{H}}_i \right)^{-1} \left[-N^{-1/2} \sum_{i=1}^{N} \mathrm{s}(\mathrm{w}_i, \theta_o) \right]$$

- By the central limit theorem

$$N^{-1/2} \sum_{i=1}^{N} \mathrm{s}(\mathrm{w}_i, \theta_o) \xrightarrow{d} Normal(0, \mathrm{B}_o)$$

$$\mathrm{B}_o = Var[\mathrm{s}(\mathrm{w}_i, \theta_o)] = E[\mathrm{s}(\mathrm{w}_i, \theta_o)\mathrm{s}(\mathrm{w}_i, \theta_o)']$$

- We can then write the asymptotic distribution of the estimator $\hat{\theta}$

$$\sqrt{N}(\hat{\theta} - \theta_o) \xrightarrow{d} Normal(0, \mathrm{A}_o^{-1}\mathrm{B}_o\mathrm{A}_o^{-1}).$$

  - **Does this remind you of something?**