Foundations of Estimation

What is a Model?

- A **model** is a simplified mathematical description of how we believe data are generated.
- Models connect:
  - **Inputs** (e.g. number of trials, predictors x_i)
  - **Parameters** (unknown quantities like p, μ, λ, β)
  - **Outputs** (observed data y_i)
- Models can be:
  - **Probability models:** specify a full probability distribution for outcomes
  - **Non-probability models:** focus on prediction or minimizing a loss without full distributional assumptions
- Key idea: models are not reality, but *useful simplifications*.

Probability Models

- Specify a probability distribution for the data, with unknown parameters.
- Examples: Binomial(p), Normal(μ, σ^2), Exponential(λ).
- Advantages:
  - Clear interpretation of parameters
  - Built-in methods for uncertainty (standard errors, confidence intervals)
  - Can compare across models using likelihood-based criteria
- Disadvantages:
  - Require assumptions about the data generating process
  - Sensitive to misspecification

Non-Probability Models

- Do not assume a full probability distribution.
- Examples: k-means clustering, regression trees, neural networks.
- Advantages:
  - Flexible and widely applicable
  - Fewer distributional assumptions
- Disadvantages:
  - Harder to quantify uncertainty
  - Comparisons between models are less formal

Estimators vs. Estimates

- **Parameter** θ: an unknown number describing a distribution or model
  - Binomial: p, Normal: (μ, σ^2), Exponential: λ
  - Regression: $\beta_0, \beta_1, \ldots$ and error spread σ^2
- **Estimator** $\hat{\theta}$: a rule or formula that uses data to produce a guess for θ
- **Estimate**: the numerical value you get from the estimator after plugging in your sample

Examples of Parameters

Model	Parameter(s)	Meaning
Bernoulli/Binomial	p	success probability
Normal	$\mu,\ \sigma^2$	center and spread
Exponential	λ	event rate
Simple Regression	β_0, β_1	intercept, slope

The Regression Equation

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \qquad \mathbb{E}[\varepsilon_i] = 0.$$

- The *parameters* are the coefficients (β_0, β_1) (and often error spread σ^2).
- Fitting regression means *estimating* these parameters from data (x_i, y_i).

What makes a good estimator?

- **Unbiasedness**: $\mathbb{E}[\hat{\theta}] = \theta$ (right on average)
- **Low variability**: estimates don't bounce around too much across samples
- **Mean Squared Error (MSE)**: $\mathrm{MSE}(\hat{\theta}) = \mathrm{Bias}(\hat{\theta})^2 + \mathrm{Var}(\hat{\theta})$
- **Consistency**: with more data, estimates hone in on the truth
- **Robustness**: not overreacting to a few unusual points

The Likelihood Function

- Suppose we have a distribution with parameter θ and data points $x_1, \ldots, x_n$.
- The probability of the entire dataset is

$$L(\theta) = f(x_1;\theta) \times f(x_2;\theta) \times \cdots \times f(x_n;\theta).$$

- This function of θ is called the **likelihood function**.
- **MLE idea:** pick the value of θ that makes the observed data most "plausible."

Likelihood: Data Fixed, Parameter Varies

- For i.i.d. data $x_1, \ldots, x_n$ from a model with parameter θ, the **likelihood** is

$$L(\theta \mid x) = \prod_{i=1}^{n} f(x_i; \theta).$$

- Think of x as *fixed* (already observed) and θ as the *variable*. We "slide" θ along its domain and see how plausible the observed data look.
- $L(\theta \mid x)$ is not a probability distribution over θ (it does not integrate to 1); it is a *score* of plausibility.
- We often maximize the **log-likelihood** $\ell(\theta) = \log L(\theta \mid x)$ for numerical stability; $\arg\max_\theta \ell(\theta) = \arg\max_\theta L(\theta)$.

Binomial Likelihood: Setup

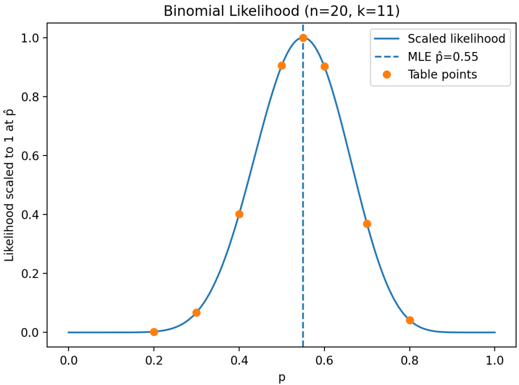
- Data summary: n trials, k successes.
- Binomial model: $X \sim \text{Binomial}(n, p)$ with unknown $p \in (0, 1)$.
- Likelihood (treating k as observed):

$$L(p \mid k) \;=\; \binom{n}{k} p^k (1-p)^{n-k} \;\propto\; p^k (1-p)^{n-k}.$$

- The combinatorial factor $\binom{n}{k}$ does not depend on p, so we can drop it when maximizing or plotting *up to scale*.

Plotting $L(p)$ for the Binomial

- Example: $n = 20$, $k = 11$.
- The MLE is $\hat{p} = \dfrac{k}{n} = 0.55$ (can be shown by calculus or inspection).
- We can plot $L(p)$ (or a scaled version) across $p \in (0, 1)$ and mark the peak at $\hat{p}$.



Numerical Check: Evaluate at Several p

To make the "maximizes" idea concrete, evaluate the likelihood at a few p's. We report *scaled likelihood* $L(p)/L(\hat{p})$ and *log-likelihood difference* $\Delta\ell(p) = \ell(p) - \ell(\hat{p}) \leq 0$.

p	$L(p)/L(\hat{p})$	$\Delta\ell(p)$
0.20	0.0026	-5.949
0.30	0.0678	-2.691
0.40	0.4010	-0.914
0.50	0.9047	-0.100
0.55	**1.0000**	**0.000**
0.60	0.9022	-0.103
0.70	0.3692	-0.996
0.80	0.0417	-3.177

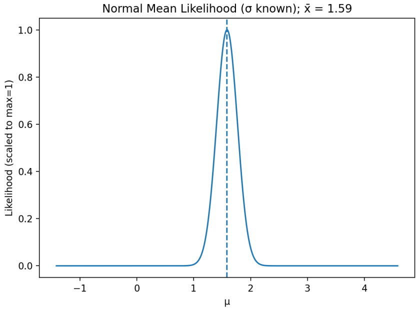
Here $n = 20$, $k = 11$, $\hat{p} = k/n = 0.55$. Scaling by $L(\hat{p})$ makes the table numerically stable (the maximum becomes 1).

Normal Mean Example

- Data: $x_1, \ldots, x_n$ from Normal(μ, σ^2), with σ known.
- Likelihood:

$$L(\mu) \propto \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right).$$

- This peaks at the sample mean $\hat{\mu} = \bar{x}$.

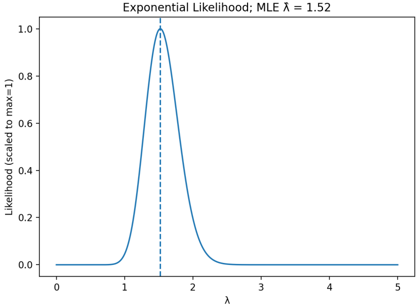


Exponential Rate Example

- Data: $x_1, \ldots, x_n$ from Exponential(λ).
- Likelihood:

$$L(\lambda) \propto \lambda^n \exp(-\lambda \sum_{i=1}^{n} x_i).$$

- This peaks at $\hat{\lambda} = \dfrac{n}{\sum x_i}$.



A Gentle Taste of Calculus

- We often maximize $\ell(\theta) = \log L(\theta)$ (log-likelihood).
- Example (Binomial):

$$\ell(p) = k \log p + (n-k)\log(1-p).$$

- Differentiate, set derivative $= 0$:

$$\frac{k}{p} - \frac{n-k}{1-p} = 0 \quad \Rightarrow \quad \hat{p} = \frac{k}{n}.$$

- Similar calculations show $\hat{\mu} = \bar{x}$ and $\hat{\lambda} = n/\sum x_i$.

Formulas for Common MLEs

- **Binomial(n, p)**: $\hat{p} = \frac{k}{n}$
- **Normal(μ, σ^2)**:

$$\hat{\mu} = \bar{x}, \quad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

- **Exponential(λ)**: $\hat{\lambda} = \frac{n}{\sum_{i=1}^{n} x_i}$

Example: Normal Data

Suppose we have data that follows a normal distribution with an unknown μ and σ. The collected data is

$$x = (1.9, 2.4, 1.8, 2.1, 2.2)$$

What are the maximum likelihood estimates of μ and σ?

Example: Normal Data

Suppose we have data that follows a normal distribution with an unknown μ and σ. The collected data is

$$x = (1.9, 2.4, 1.8, 2.1, 2.2)$$

What are the maximum likelihood estimates of μ and σ?

Example: Normal Data

Suppose that the data instead follows an exponential distribution with unknown parameter λ.

$$x = (1.9, 2.4, 1.8, 2.1, 2.2)$$

What are the maximum likelihood estimates of λ?

Example: Normal Data

Suppose that the data instead follows an exponential distribution with unknown parameter λ.

$$x = (1.9, 2.4, 1.8, 2.1, 2.2)$$

What are the maximum likelihood estimates of λ?

- $\bar{x} = 2.08$, so $\hat{\mu} = 2.08$.
- Sample variance: $\hat{\sigma}^2 = 0.05$.

Comparing Probability Models

- Use AIC:
$$AIC = -2 \log L(\hat{\theta}) + 2k$$
where k = number of parameters.
- Lower AIC means better fit (balance between goodness-of-fit and parsimony).

Example: Data Comparison

- Data: 20 values (times between arrivals, say).
- Fit Binomial, Normal, Exponential models.
- Compute log-likelihoods at $\hat{\theta}$, then AIC.
- Suppose we get:
  - Binomial: AIC $= 62$
  - Normal: AIC $= 58$
  - Exponential: AIC $= 65$
- Best model here is Normal.

How to Decide in Practice

- By data type:
  - Number of successes out of fixed trials $\Rightarrow$ Binomial.
  - Continuous, symmetric $\Rightarrow$ Normal.
  - Waiting times, skewed $\Rightarrow$ Exponential.
- Graphically: histograms, skewness, symmetry.
- Formally: compare likelihoods/AIC.

Example: Normal Model (MLE and Numbers)

Suppose $x_1, \ldots, x_n$ are i.i.d. Normal(μ, σ^2) with both μ, σ unknown.

$$x = (1.9,\ 2.4,\ 1.8,\ 2.1,\ 2.2), \quad n = 5.$$

MLEs (closed form):

$$\hat{\mu} = \bar{x} = \frac{1}{n}\sum x_i, \qquad \hat{\sigma}^2 = \frac{1}{n}\sum (x_i - \bar{x})^2$$

Values for our data:

$$\bar{x} = 2.08, \qquad \hat{\sigma}^2 = 0.0456, \quad \hat{\sigma} = 0.2135.$$

Log-likelihood at the MLE:

$$\ell_{\text{Norm}}(\hat{\mu}, \hat{\sigma}) = -n\log(\hat{\sigma}\sqrt{2\pi}) - \tfrac{1}{2}\sum \frac{(x_i - \hat{\mu})^2}{\hat{\sigma}^2} = -5\log(0.2135\sqrt{2\pi}) - \tfrac{1}{2}\cdot 5 = 0.6249.$$

(We used $\sum (x_i - \hat{\mu})^2 = n\hat{\sigma}^2$.)

Example: Exponential Model (MLE and Numbers)

Suppose $x_1, \ldots, x_n$ are i.i.d. Exponential(λ) with density $f(x;\lambda) = \lambda e^{-\lambda x}$ for $x \geq 0$.

MLE (closed form):

$$\hat{\lambda} = \frac{n}{\sum_{i=1}^{n} x_i}$$

Values for our data:

$$\sum x_i = 10.4, \quad n = 5 \quad \Rightarrow \quad \hat{\lambda} = \frac{5}{10.4} = 0.4808.$$

Log-likelihood at the MLE:

$$\ell_{\text{Exp}}(\hat{\lambda}) = n \log(\hat{\lambda}) - \hat{\lambda} \sum x_i = 5 \log(0.4808) - 0.4808 \times 10.4 = -8.6618.$$

AIC: Full Calculation and Comparison

$$AIC = -2 \log L(\hat{\theta}) + 2k,$$

where k = number of free parameters.

Normal has $k = 2$ (μ, σ):

$$AIC_{\text{Norm}} = -2 \cdot \ell_{\text{Norm}}(\hat{\mu}, \hat{\sigma}) + 2 \cdot 2 = -2(0.6249) + 4 = 2.7501.$$

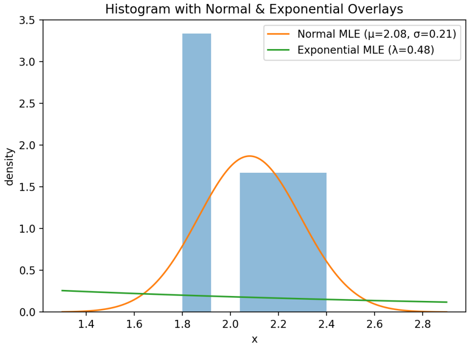
Exponential has $k = 1$ (λ):

$$AIC_{\text{Exp}} = -2 \cdot \ell_{\text{Exp}}(\hat{\lambda}) + 2 \cdot 1 = -2(-8.6618) + 2 = 19.3237.$$

Result: $AIC_{\text{Norm}} \ll AIC_{\text{Exp}} \Rightarrow$ Normal fits much better for this dataset.

Visual Comparison: Overlaid Fits

- Histogram of the data with the **Normal (MLE)** and **Exponential (MLE)** densities overlaid.
- This matches the AIC result: Normal aligns closely; Exponential decays too slowly for these values.



How to Decide in Practice

- By data type:
  - Number of successes out of fixed trials $\Rightarrow$ Binomial.
  - Continuous, symmetric $\Rightarrow$ Normal.
  - Waiting times, skewed $\Rightarrow$ Exponential.
- Graphically: histograms, skewness, symmetry.
- Formally: compare likelihoods/AIC.

Empirical Risk (Loss) Minimization

- Pick a **loss** $\ell(\widehat{y}, y)$ to measure how bad a prediction $\widehat{y}$ is when the truth is y.
- For a model $f_\theta(x)$, define the **empirical risk** (total/average loss)

$$Q(\theta) = \sum_{i=1}^{n} \ell\big(f_\theta(x_i),\, y_i\big) \quad \text{or} \quad \frac{1}{n}\sum_{i=1}^{n} \ell(\cdot).$$

- Estimation by **minimization**:

$$\hat{\theta} = \arg\min_{\theta} Q(\theta).$$

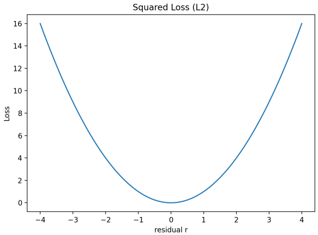
- ERM is very general: many classical estimators are solutions of $\min Q(\theta)$.

Loss Choice ⇒ What You Estimate

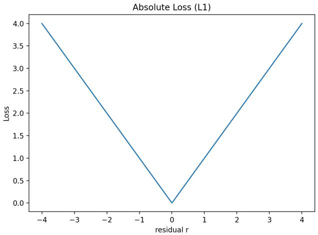
- **Squared loss** $\ell(r) = r^2$ with $r = y - \widehat{y}$ ⇒ estimates the **mean**.
- **Absolute loss** $\ell(r) = |r|$ ⇒ estimates the **median**.
- **0–1 loss** $\ell(r) = \mathbf{1}\{r \neq 0\}$ ⇒ estimates the **mode**.
- **Pinball (quantile) loss** for $\tau \in (0, 1)$:

$$\ell_\tau(r) = \begin{cases} \tau\, r, & r \geq 0, \\ (\tau - 1)\, r, & r < 0, \end{cases} \quad \Rightarrow \quad \text{estimates the } \tau\text{-quantile.}$$

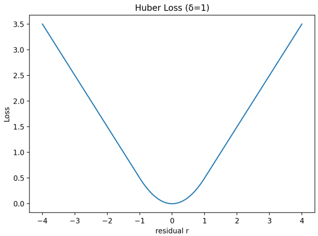
Common Loss Functions



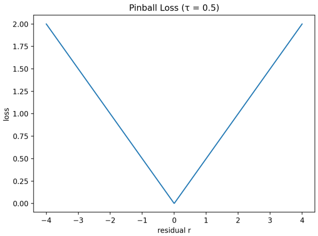
Squared (L2)



Absolute (L1)



Huber



Quantile (τ=0.5)

Regression as Loss Minimization

- **L2/OLS:**

$$\hat{\beta}_{\text{L2}} = \arg\min_{\beta_0,\beta_1} \sum_i (y_i - \beta_0 - \beta_1 x_i)^2.$$

- **L1/LAD (least absolute deviations):**

$$\hat{\beta}_{\text{L1}} = \arg\min_{\beta_0,\beta_1} \sum_i \left| y_i - \beta_0 - \beta_1 x_i \right|.$$

- **Huber regression:** replace r^2 by Huber loss to reduce outlier influence.
- If errors are Normal, maximizing likelihood $\Leftrightarrow$ minimizing SSE (L2).
- Different losses $\Rightarrow$ different solutions and robustness.

Why OLS = MLE (Normal Errors)

Assume the regression model

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \qquad \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2).$$

- Likelihood:

$$L(\beta_0, \beta_1, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2} \right).$$

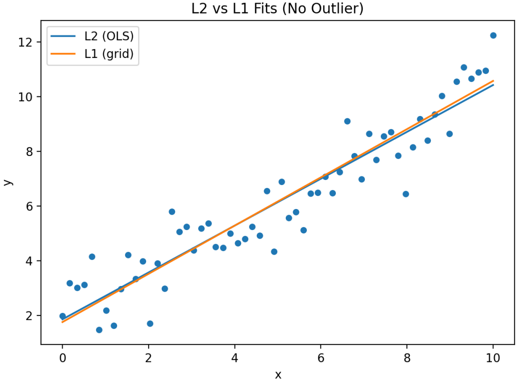
- Log-likelihood (up to constants):

$$\ell(\beta_0, \beta_1, \sigma^2) = -\frac{n}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2.$$

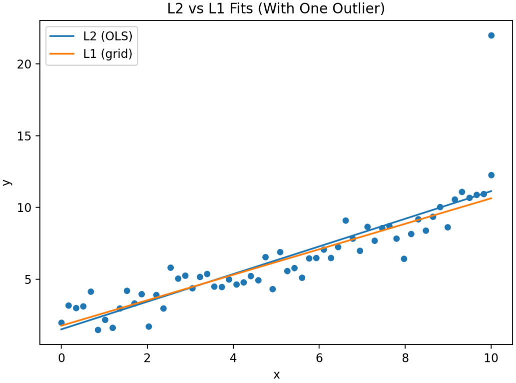
- For fixed σ^2, maximizing ℓ in (β_0, β_1) is

$$\arg\min_{\beta_0, \beta_1} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 .$$

Visual: Outliers Move L2 More Than L1



No Outlier



With Outlier

L2 and L1 fits are similar on clean data. With one extreme point, the L2 line tilts more; the L1 line is more stable.

Regularization = Penalized Loss (Shrinkage)

- Add a penalty to trade a little bias for lower variance (helps generalization):

$$\min_{\beta} \underbrace{\sum_i (y_i - x_i^\top \beta)^2}_{\text{fit}} + \lambda \underbrace{\|\beta\|_2^2}_{\text{Ridge}} \quad \text{or} \quad \min_{\beta} \sum_i (y_i - x_i^\top \beta)^2 + \lambda \|\beta\|_1 \text{ (Lasso).}$$

- Ridge shrinks coefficients smoothly; Lasso can set some exactly to zero (feature selection).
- Same ERM template: "fit" + "complexity penalty."

Choosing a Loss: A Quick Checklist

- Symmetric errors, few outliers $\Rightarrow$ L2 (efficient under Normality).
- Heavy tails / outliers $\Rightarrow$ L1 or Huber.
- Want a specific quantile (e.g., 90% service time) $\Rightarrow$ Pinball loss.
- Care about large errors more $\Rightarrow$ higher powers (but beware instability).