Statistical Inference Review

Gonzalo Mateos

Dept. of ECE and Goergen Institute for Data Science
University of Rochester
gmateosb@ece.rochester.edu
http://www.ece.rochester.edu/~gmateosb/

Facultad de Ingenieria, UdelaR
Montevideo, Uruguay
December 30, 2020

Statistical inference and models

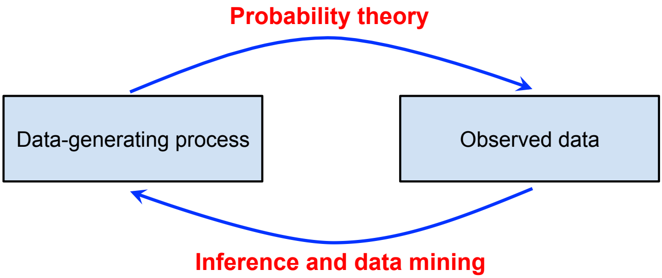
Statistical inference and models

Point estimates, confidence intervals and hypothesis tests

Tutorial on inference about a mean

Tutorial on linear regression inference

Probability and inference



- Probability theory is a formalism to work with uncertainty
  - Given a data-generating process, what are properties of outcomes?
- Statistical inference deals with the inverse problem
  - Given outcomes, what can we say on the data-generating process?

Statistical inference

- Statistical inference refers to the process whereby
  - $\Rightarrow$ Given observations $\mathbf{x} = [x_1, \ldots, x_n]^T$ from $X_1, \ldots, X_n \sim F$
  - $\Rightarrow$ We aim to extract information about the distribution F
- Ex: Infer a feature of F such as its mean
- Ex: Infer the CDF F itself, or the PDF $f = F'$
- Often observations are of the form (y_i, x_i), $i = 1, \ldots, n$
  - $\Rightarrow$ Y is the response or outcome. X is the predictor or feature
- Q: Relationship between the random variables (RVs) Y and X?
- Ex: Learn $\mathbb{E}\left[Y \,\middle|\, X = x\right]$ as a function of x
- Ex: Foretelling a yet-to-be observed value y_* from the input $X_* = x_*$

Models

- A statistical model specifies a set $\mathcal{F}$ of CDFs to which F may belong
- A common parametric model is of the form $\mathcal{F} = \{f(x;\theta) : \theta \in \Theta\}$
  - Parameter(s) θ are unknown, take values in parameter space Θ
  - Space Θ has $\dim(\Theta) < \infty$, not growing with the sample size n
- Ex: Data come from a Gaussian distribution

$$\mathcal{F}_N = \left\{ f(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}},\ \mu \in \mathbb{R},\ \sigma > 0 \right\}$$

$\Rightarrow$ A two-parameter model: $\boldsymbol{\theta} = [\mu, \sigma]^T$ and $\Theta = \mathbb{R} \times \mathbb{R}_+$

- A nonparametric model has $\dim(\Theta) = \infty$, or $\dim(\Theta)$ grows with n
- Ex: $\mathcal{F}_{All} = \{\text{All CDFs } F\}$

Models and inference tasks

- Given independent data $\mathbf{x} = [x_1, \ldots, x_n]^T$ from $X_1, \ldots, X_n \sim F$

  $\Rightarrow$ Statistical inference often conducted in the context of a model

Ex: One-dimensional parametric estimation

- Suppose observations are Bernoulli distributed with parameter p
- The task is to estimate the parameter p (i.e., the mean)

Ex: Two-dimensional parametric estimation

- Suppose the PDF $f \in \mathcal{F}_N$, i.e., data are Gaussian distributed
- The problem is to estimate the parameters μ and σ
- May only care about μ, and treat σ as a nuisance parameter

Ex: Nonparametric estimation of the CDF

- The goal is to estimate F assuming only $F \in \mathcal{F}_{All} = \{\text{All CDFs } F\}$

Regression models

- Suppose observations are from $(Y_1, X_1), \ldots, (Y_n, X_n) \sim F_{YX}$

  $\Rightarrow$ Goal is to learn the relationship between the RVs Y and X

- A typical approach is to model the regression function

$$r(x) := \mathbb{E}\left[Y \,\middle|\, X = x\right] = \int_{-\infty}^{\infty} y f_{Y|X}(y|x) dy$$

  $\Rightarrow$ Equivalent to the regression model $Y = r(X) + \epsilon$, $\mathbb{E}\left[\epsilon\right] = 0$

- Ex: Parametric linear regression model

$$r \in \mathcal{F}_{Lin} = \{r : r(x) = \beta_0 + \beta_1 x\}$$

- Ex: Nonparametric regression model, assuming only smoothness

$$r \in \mathcal{F}_{Sob} = \left\{r : \int_{-\infty}^{\infty} (r''(x))^2 dx < \infty\right\}$$

Regression, prediction and classification

- Given data $(y_1, x_1), \ldots, (y_n, x_n)$ from $(Y_1, X_1), \ldots, (Y_n, X_n) \sim F_{YX}$
  - Ex: x_i is the blood pressure of subject i, y_i how long she lived
- Model the relationship between Y and X via $r(x) = \mathbb{E}\left[Y \,\middle|\, X = x\right]$

  $\Rightarrow$ Q: What are classical inference tasks in this context?

Ex: Regression or curve fitting

- The problem is to estimate the regression function $r \in \mathcal{F}$

Ex: Prediction

- The goal is to predict Y_* for a new patient based on their $X_* = x_*$
- If a regression estimate $\hat{r}$ is available, can do $y_* := \hat{r}(x_*)$

Ex: Classification

- Suppose RVs Y_i are discrete, e.g. live or die encoded as ± 1
- The prediction problem above is termed classification

Fundamental concepts in inference

Statistical inference and models

Point estimates, confidence intervals and hypothesis tests

Tutorial on inference about a mean

Tutorial on linear regression inference

Point estimators

- Point estimation refers to making a single "best guess" about F
- Ex: Estimate the parameter $\boldsymbol{\beta}$ in a linear regression model

$$\mathcal{F}_{Lin} = \left\{ r : r(\mathbf{x}) = \boldsymbol{\beta}^T \mathbf{x} \right\}$$

- **Def:** Given data $\mathbf{x} = [x_1, \ldots, x_n]^T$ from $X_1, \ldots, X_n \sim F$, a point estimator $\hat{\theta}$ of a parameter θ is some function

$$\hat{\theta} = g(X_1, \ldots, X_n)$$

  $\Rightarrow$ The estimator $\hat{\theta}$ is computed from the data, hence it is a RV

  $\Rightarrow$ The distribution of $\hat{\theta}$ is called sampling distribution

- The estimate is the specific value for the given data sample $\mathbf{x}$

  $\Rightarrow$ May write $\hat{\theta}_n$ to make explicit reference to the sample size

Bias, standard error and mean squared error

- **Def:** The bias of an estimator $\hat{\theta}$ is given by $\text{bias}(\hat{\theta}) := \mathbb{E}\left[\hat{\theta}\right] - \theta$
- **Def:** The standard error is the standard deviation of $\hat{\theta}$

$$\text{se} = \text{se}(\hat{\theta}) := \sqrt{\text{var}\left[\hat{\theta}\right]}$$

$\Rightarrow$ Often, se depends on the unknown F. Can form an estimate $\hat{\text{se}}$

- **Def:** The mean squared error (MSE) is a measure of quality of $\hat{\theta}$

$$\text{MSE} = \mathbb{E}\left[(\hat{\theta} - \theta)^2\right]$$

- Expected values are with respect to the data distribution

$$f(x_1, \ldots, x_n; \theta) = \prod_{i=1}^{n} f(x_i; \theta)$$

The bias-variance decomposition of the MSE

Theorem

The MSE $= \mathbb{E}\left[(\hat{\theta}-\theta)^2\right]$ *can be written as*

$$MSE = bias^2(\hat{\theta}) + \text{var}\left[\hat{\theta}\right]$$

Proof.

- Let $\bar{\theta} = \mathbb{E}\left[\hat{\theta}\right]$. Then

$$\begin{aligned}
\mathbb{E}\left[(\hat{\theta}-\theta)^2\right] &= \mathbb{E}\left[(\hat{\theta}-\bar{\theta}+\bar{\theta}-\theta)^2\right] \\
&= \mathbb{E}\left[(\hat{\theta}-\bar{\theta})^2\right] + 2(\bar{\theta}-\theta)\mathbb{E}\left[\hat{\theta}-\bar{\theta}\right] + (\bar{\theta}-\theta)^2 \\
&= \text{var}\left[\hat{\theta}\right] + \text{bias}^2(\hat{\theta})
\end{aligned}$$

- The last equality follows since $\mathbb{E}\left[\hat{\theta}-\bar{\theta}\right] = \mathbb{E}\left[\hat{\theta}\right] - \bar{\theta} = 0$

□

Desirable properties of point estimators

- Q: Desiderata for an estimator $\hat{\theta}$ of the parameter θ?
- **Def:** An estimator is unbiased if $\text{bias}(\hat{\theta}) = 0$, i.e., if $\mathbb{E}\left[\hat{\theta}\right] = \theta$

  $\Rightarrow$ An unbiased estimator is "on target" on average
- **Def:** An estimator is consistent if $\hat{\theta}_n \xrightarrow{p} \theta$, i.e. for any $\epsilon > 0$

$$\lim_{n\to\infty} \text{P}\left[|\hat{\theta}_n - \theta| < \epsilon\right] = 1$$

  $\Rightarrow$ A consistent estimator converges to θ as we collect more data
- **Def:** An unbiased estimator is asymptotically Normal if

$$\lim_{n\to\infty} \text{P}\left[\frac{\hat{\theta}_n - \theta}{\text{se}} \leq x\right] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-u^2/2} du$$

  $\Rightarrow$ Equivalently, for large enough sample size then $\hat{\theta}_n \sim \mathcal{N}(\theta, \text{se}^2)$

Coin tossing example

Ex: Consider tossing the same coin n times and record the outcomes

- Model observations as $X_1, \ldots, X_n \sim \text{Ber}(p)$. Estimate of p?
- A natural choice is the sample mean estimator

$$\hat{p} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

- Recall that for $X \sim \text{Ber}(p)$, then $\mathbb{E}[X] = p$ and $\text{var}[X] = p(1-p)$
- The estimator $\hat{p}$ is unbiased since

$$\mathbb{E}[\hat{p}] = \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^{n} X_i\right] = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[X_i] = p$$

$\Rightarrow$ Also used that the expected value is a linear operator

Coin tossing example (continued)

- The standard error is

$$\text{se} = \sqrt{\text{var}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right]} = \sqrt{\frac{1}{n^2}\sum_{i=1}^{n}\text{var}\left[X_i\right]} = \sqrt{\frac{p(1-p)}{n}}$$

$\Rightarrow$ Unknown p. Estimated standard error is $\hat{\text{se}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

- Since $\hat{p}_n$ is unbiased, then $\text{MSE} = \mathbb{E}\left[(\hat{p}_n - p)^2\right] = \frac{p(1-p)}{n} \to 0$
  - Thus $\hat{p}$ converges in the mean square sense, hence also $\hat{p}_n \xrightarrow{p} p$
  - Establishes $\hat{p}$ is a consistent estimator of the parameter p

- Also, $\hat{p}$ is asymptotically Normal by the Central Limit Theorem

Confidence intervals

- Set estimates specify regions of Θ where θ is likely to lie on
- **Def:** Given i.i.d. data $X_1, \ldots, X_n \sim F$, a $1-\alpha$ confidence interval of a parameter θ is an interval $C_n = (a, b)$, where $a = a(X_1, \ldots, X_n)$ and $b = b(X_1, \ldots, X_n)$ are functions of the data such that

$$\mathrm{P}\left[\theta \in C_n\right] = 1 - \alpha, \quad \text{for all } \theta \in \Theta$$

$\Rightarrow$ In words, $C_n = (a, b)$ traps θ with probability $1-\alpha$

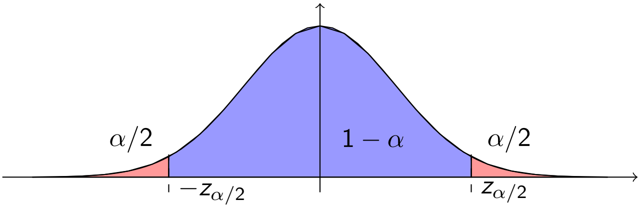
$\Rightarrow$ The interval C_n is computed from the data, hence it is random

- We call $1-\alpha$ the coverage of the confidence interval
- Ex: It is common to report 95% confidence intervals, i.e., $\alpha = 0.05$

Aside on the standard Normal distribution

- Let X be a standard Normal RV, i.e., $X \sim \mathcal{N}(0,1)$ with CDF $\Phi(x)$

$$\Phi(x) = \mathrm{P}\left[X \leq x\right] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{u^2}{2}} du$$



- Define $z_{\alpha/2} = \Phi^{-1}(1-(\alpha/2))$, i.e., the value such that

$$\mathrm{P}\left[X > z_{\alpha/2}\right] = \alpha/2 \text{ and } \mathrm{P}\left[-z_{\alpha/2} < X < z_{\alpha/2}\right] = 1-\alpha$$

Normal-based confidence intervals

- Nice point estimators $\hat{\theta}_n$ are Normal as $n \to \infty$, i.e., $\hat{\theta}_n \sim \mathcal{N}(\theta, \hat{\mathsf{se}}^2)$

  $\Rightarrow$ Useful property in constructing confidence intervals for θ

Theorem

Suppose that $\hat{\theta}_n \sim \mathcal{N}(\theta, \hat{\mathsf{se}}^2)$ as $n \to \infty$. Let Φ be the CDF of a standard Normal and define $z_{\alpha/2} = \Phi^{-1}(1 - (\alpha/2))$. Consider the interval

$$C_n = (\hat{\theta}_n - z_{\alpha/2}\hat{se}, \hat{\theta}_n + z_{\alpha/2}\hat{se}).$$

Then $\mathrm{P}[\theta \in C_n] \to 1 - \alpha$, as $n \to \infty$

- These intervals only have approximately (large n) correct coverage

Proof

Proof.

- Consider the normalized (centered and scaled) RV

$$X_n = \frac{\hat{\theta}_n - \theta}{\hat{\mathsf{se}}}$$

- By assumption $X_n \to X \sim \mathcal{N}(0, 1)$ as $n \to \infty$. Hence,

$$\begin{aligned} \mathsf{P}\left[\theta \in C_n\right] &= \mathsf{P}\left[\hat{\theta}_n - z_{\alpha/2}\hat{\mathsf{se}} < \theta < \hat{\theta}_n + z_{\alpha/2}\hat{\mathsf{se}}\right] \\ &= \mathsf{P}\left[-z_{\alpha/2} < \frac{\hat{\theta}_n - \theta}{\hat{\mathsf{se}}} < z_{\alpha/2}\right] \\ &\to \mathsf{P}\left[-z_{\alpha/2} < X < z_{\alpha/2}\right] = 1 - \alpha \end{aligned}$$

- The last equality follows by definition of $z_{\alpha/2}$

□

Coin tossing example (encore)

Ex: Given observations $X_1, \dots, X_n \sim \text{Ber}(p)$. Estimate of p?

- We studied properties of the sample mean estimator

$$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

- By the Central Limit Theorem, it follows that

$$\hat{p} \sim \mathcal{N}\left(p, \frac{\hat{p}(1-\hat{p})}{n}\right) \text{ as } n \to \infty$$

- Therefore, an approximate $1-\alpha$ confidence interval for p is

$$C_n = \left(\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$$

Hypothesis testing

- In hypothesis testing we start with some default theory
  - Ex: The data come from a zero-mean Gaussian distribution
- Q: Do the data provide sufficient evidence to reject the theory?
- The hypothesized theory is called null hypothesis, written as H_0

  $\Rightarrow$ Specify also an alternative hypothesis to the null, H_1
- Formally, given i.i.d. data $\mathbf{x} = [x_1, \ldots, x_n]^T$ from $X_1, \ldots, X_n \sim F$
  (i) Form a test statistic $T(\mathbf{x})$, i.e., a function of the data
  (ii) Define a rejection region $\mathcal{R}$ of the form

$$\mathcal{R} = \{\mathbf{x} : T(\mathbf{x}) > c\}$$

- If data $\mathbf{x} \in \mathcal{R}$ we reject H_0, otherwise we retain (do not reject) H_0
- The problem is to select the test statistic T and the critical value c

Testing if a coin is fair

Ex: Consider tossing the same coin n times and record the outcomes

- Model observations as $X_1, \ldots, X_n \sim \text{Ber}(p)$. Is the coin fair?
- Let H_0 be the hypothesis that the coin is fair, and H_1 the alternative
  $\Rightarrow$ Can write the hypotheses as

$$H_0 : p = 1/2 \quad \text{versus} \quad H_1 : p \neq 1/2$$

- Consider the test statistic given by

$$T(X_1, \ldots, X_n) = \left| \hat{p}_n - \frac{1}{2} \right| = \left| \frac{1}{n} \sum_{i=1}^{n} X_i - \frac{1}{2} \right|$$

- It seems reasonable to reject H_0 if $(X_1, \ldots, X_n) \in \mathcal{R}$, where

$$\mathcal{R} = \{(X_1, \ldots, X_n) : T(X_1, \ldots, X_n) > c\}$$

- Will soon see this is a Wald's test, hence $c = z_{\alpha/2}\hat{\text{se}}$. More later

Tutorial on inference about a mean

Inference about a mean

- Consider a sample of n i.i.d. observations $X_1, \ldots, X_n \sim F$
- Q: How can we perform inference about the mean $\mu = \mathbb{E}[X_1]$?

  $\Rightarrow$ Practical and canonical problem in statistical inference

- A natural estimator of μ is the sample mean estimator

$$\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

  $\Rightarrow$ Well motivated since by the strong law of large numbers

$$\lim_{n\to\infty} \hat{\mu}_n = \mu \quad \text{almost surely}$$

- It is a simple example of a method of moments estimator (MME)...
- ...and also a maximum likelihood estimator (MLE)

Moments and sample moments

- In parametric inference we wish to estimate $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^p$ in

$$\mathcal{F} = \{f(x; \boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}$$

- For $1 \leq j \leq p$, define the j-th moment of $X \sim F$ as

$$\alpha_j \equiv \alpha_j(\boldsymbol{\theta}) = \mathbb{E}\left[X^j\right] = \int_{-\infty}^{\infty} x^j f(x; \boldsymbol{\theta}) dx$$

- Likewise, the j-th sample moment is an estimate of α_j, namely

$$\hat{\alpha}_j = \frac{1}{n} \sum_{i=1}^{n} X_i^j$$

$\Rightarrow$ The j-th moment $\alpha_j(\boldsymbol{\theta})$ depends on the unknown $\boldsymbol{\theta}$

$\Rightarrow$ But $\hat{\alpha}_j$ does not, a function of the data only

Method of moments estimator

- A first method for parametric estimation is the method of moments
  - $\Rightarrow$ MMEs are not optimal, yet typically easy to compute
- **Def:** The method of moments estimator (MME) $\hat{\boldsymbol{\theta}}_n$ is the solution to

$$\begin{array}{ccc} \alpha_1(\hat{\boldsymbol{\theta}}_n) & = & \hat{\alpha}_1 \\ \alpha_2(\hat{\boldsymbol{\theta}}_n) & = & \hat{\alpha}_2 \\ \vdots & \vdots & \vdots \\ \alpha_p(\hat{\boldsymbol{\theta}}_n) & = & \hat{\alpha}_p \end{array}$$

  - $\Rightarrow$ This is a system of p (nonlinear) equations with p unknowns
- Ex: Back to estimating a mean μ, $p = 1$ and $\mu = \theta = \alpha_1(\theta)$ so

$$\hat{\mu}_n^{MM} = \hat{\alpha}_1 = \frac{1}{n} \sum_{i=1}^{n} X_i$$

Example: Gaussian data model

Ex: Suppose now $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$, i.e., the model is $F \in \mathcal{F}_N$

- Q: What is the MME of the parameter vector $\boldsymbol{\theta} = [\mu, \sigma^2]^T$?
- The first $p = 2$ moments are given by

$$\alpha_1(\boldsymbol{\theta}) = \mathbb{E}\left[X_1\right] = \mu, \quad \alpha_2(\boldsymbol{\theta}) = \mathbb{E}\left[X_1^2\right] = \sigma^2 + \mu^2$$

- The MME $\hat{\boldsymbol{\theta}}_n$ is the solution to the following system of equations

$$\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$\hat{\sigma}_n^2 + \hat{\mu}_n^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2$$

- The solution is

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

Maximum likelihood estimator

- Often "the" method for parametric estimation is maximum likelihood

- Consider i.i.d. data $X_1, \ldots, X_n$ from a PDF $f(x;\theta)$
- The likelihood function $\mathcal{L}_n(\theta) : \Theta \to \mathbb{R}_+$ is defined by

$$\mathcal{L}_n(\theta) := \prod_{i=1}^{n} f(X_i;\theta)$$

$\Rightarrow$ $\mathcal{L}_n(\theta)$ is the joint PDF of the data, treated as a function of θ

$\Rightarrow$ The log-likelihood function is $\ell_n(\theta) := \log \mathcal{L}_n(\theta)$

- **Def:** The maximum likelihood estimator (MLE) $\hat{\boldsymbol{\theta}}_n$ is given by

$$\hat{\boldsymbol{\theta}}_n = \arg\max_{\theta} \mathcal{L}_n(\theta)$$

- Very useful: The maximizer of $\mathcal{L}_n(\theta)$ coincides with that of $\ell_n(\theta)$

Example: Bernoulli data model

- Suppose $X_1, \ldots, X_n \sim \text{Ber}(p)$. MLE of $\mu = p$?
  $\Rightarrow$ The data PMF is $f(x;p) = p^x(1-p)^{1-x}$, $x \in \{0,1\}$
- The likelihood function is (define $S_n = \sum_{i=1}^n X_i$)

$$\mathcal{L}_n(p) = \prod_{i=1}^n f(X_i;p) = \prod_{i=1}^n p^{X_i}(1-p)^{1-X_i} = p^{S_n}(1-p)^{n-S_n}$$

  $\Rightarrow$ The log-likelihood is $\ell_n(p) = S_n \log(p) + (n - S_n)\log(1-p)$
- The MLE $\hat{p}_n$ is the solution to the equation

$$\left.\frac{\partial \ell_n(p)}{\partial p}\right|_{p=\hat{p}_n} = \frac{S_n}{\hat{p}_n} - \frac{n - S_n}{1 - \hat{p}_n} = 0$$

- The solution is

$$\hat{\mu}_n^{ML} = \hat{p}_n = \frac{S_n}{n} = \frac{1}{n}\sum_{i=1}^n X_i$$

Example: Gaussian data model

- Suppose $X_1, \ldots, X_n \sim \mathcal{N}(\mu, 1)$. MLE of μ?

  $\Rightarrow$ The data PDF is $f(x;\mu) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{(x-\mu)^2}{2}\right\}$, $x \in \mathbb{R}$

- The likelihood function is (up to constants independent of μ)

$$\mathcal{L}_n(\mu) = \prod_{i=1}^{n} f(X_i;\mu) \propto \exp\left\{-\sum_{i=1}^{n} \frac{(X_i-\mu)^2}{2}\right\}$$

  $\Rightarrow$ The log-likelihood is $\ell_n(\mu) \propto -\sum_{i=1}^{n}(X_i-\mu)^2$

- The MLE $\hat{\mu}_n$ is the solution to the equation

$$\left.\frac{\partial \ell_n(\mu)}{\partial \mu}\right|_{\mu=\hat{\mu}_n} = 2\sum_{i=1}^{n}(X_i - \hat{\mu}_n) = 0$$

- The solution is, once more, the sample mean estimator

$$\hat{\mu}_n^{ML} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Properties of the MLE

- MLEs have desirable properties under loose conditions on $f(x;\theta)$

P1) Consistency: $\hat{\theta}_n \xrightarrow{p} \theta$ as the sample size n increases

P2) Equivariance: If $\hat{\theta}_n$ is the MLE of θ, then $g(\hat{\theta}_n)$ is the MLE of $g(\theta)$

P3) Asymptotic Normality: For large n, one has $\hat{\theta}_n \sim \mathcal{N}(\theta, \hat{\text{se}}^2)$

P4) Efficiency: For large n, $\hat{\theta}_n$ attains the Cramér-Rao lower bound

- Efficiency means no other unbiased estimator has smaller variance

- Ex: Can use the MLE to create a confidence interval for μ, i.e.,

$$C_n = \left(\hat{\mu}_n^{ML} - z_{\alpha/2}\hat{\text{se}}, \hat{\mu}_n^{ML} + z_{\alpha/2}\hat{\text{se}}\right)$$

$\Rightarrow$ By asymptotic Normality, $\text{P}\left[\mu \in C_n\right] \approx 1 - \alpha$ for large n

$\Rightarrow$ For the $\mathcal{N}(\mu, 1)$ model, $\hat{\mu}_n^{ML} \pm \frac{z_{\alpha/2}}{\sqrt{n}}$ has exact coverage

The Wald test

- Consider the following hypothesis test regarding the mean μ

$$H_0 : \mu = \mu_0 \quad \text{versus} \quad H_1 : \mu \neq \mu_0$$

- Let $\hat{\mu}_n$ be the sample mean, with estimated standard error $\hat{\text{se}}$
- **Def:** Given $\alpha \in (0, 1)$, the Wald test rejects H_0 when

$$T(X_1, \ldots, X_n) := \left| \frac{\hat{\mu}_n - \mu_0}{\hat{\text{se}}} \right| > z_{\alpha/2}$$

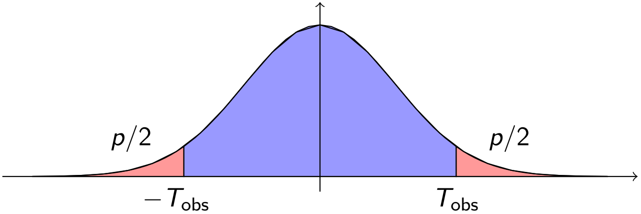
- If H_0 is true, $\frac{\hat{\mu}_n - \mu_0}{\hat{\text{se}}} \sim \mathcal{N}(0, 1)$ by the Central Limit Theorem

  $\Rightarrow$ Probability of incorrectly rejecting H_0 is no more than α

- The value of α is called the significance level of the test

The p-value

- Reporting "reject H_0" or "retain H_0" is not too informative
  - $\Rightarrow$ Could ask, for each α, whether the test rejects at that level
- Let $T_{\text{obs}} := T(\mathbf{x})$ be the test statistic value for the observed sample



- The probability $p := \mathrm{P}_{H_0}(|T(\mathbf{X})| \geq T_{\text{obs}})$ is called the p-value
  - $\Rightarrow$ Smallest level at which we would reject H_0
- A small p-value (< 0.05) indicates reduced evidence supporting H_0

Bayesian inference

- Methods discussed so far are termed frequentist, where:
  - F1: Probability refers to limiting relative frequencies
  - F2: Parameters are fixed, unknown constants
  - F3: Statistical procedures offer guarantees on long-run performance
- Alternatively, Bayesian inference is based on these postulates:
  - B1: Probability describes degree of belief, not limiting frequency
  - B2: We can make probability statements about parameters
  - B3: A probability distribution for θ is produced to make inferences
- Controversial? Inherently embraces a subjective notion of probability
  - Bayesian methods do not offer long-run performance guarantees
  - Very useful to combine prior beliefs with data in a principled way

The Bayesian method

- Bayesian inference is usually carried out in the following way

Step 1: Choose a probability density $f(\theta)$ called the prior distribution

- The prior expresses our beliefs about θ, before seeing any data

Step 2: Choose a statistical model $f(x \,|\, \theta)$ (compare with $f(x;\theta)$)

- Reflects our beliefs about the data-generating process, i.e., X given θ

Step 3: Given data $\mathbf{X} = [X_1, \ldots, X_n]^T$, we update our beliefs and calculate the posterior distribution $f(\theta|\mathbf{X})$ using Bayes' rule

$$f(\theta|\mathbf{X}) \propto \prod_{i=1}^{n} f(X_i \,|\, \theta) f(\theta) = \mathcal{L}_n(\theta) f(\theta)$$

$\Rightarrow$ Point estimates, confidence intervals obtained from $f(\theta|\mathbf{X})$

- Ex: A maximum a posteriori (MAP) estimator $\hat{\theta}_n = \arg\max_\theta f(\theta|\mathbf{X})$

Example: Gaussian data model and prior

- Consider $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$. Suppose σ^2 is known
  $\Rightarrow$ To estimate θ we adopt the prior $\theta \sim \mathcal{N}(a, b^2)$
- Using Bayes' rule, can show the posterior is also Gaussian where

$$\hat{\theta}_n^{MAP} = \mathbb{E}\left[\theta \,\middle|\, \mathbf{X}\right] = \frac{w}{n}\sum_{i=1}^{n} X_i + (1-w)a, \text{ with } w = \frac{\text{se}^{-2}}{\text{se}^{-2} + b^{-2}}$$

  $\Rightarrow$ Weighted average of the sample mean $\hat{\theta}_n^{ML}$ and the prior mean a
  $\Rightarrow$ Here, $\text{se} = \sigma/\sqrt{n}$ is the standard error for the sample mean
- Asymptotics: Note that $w \to 1$ as the sample size $n \to \infty$
  $\Rightarrow$ For large n the posterior is approximately $\mathcal{N}(\hat{\theta}_n^{ML}, \text{se}^2)$
  $\Rightarrow$ Same holds if n is fixed but $b \to \infty$, i.e., prior is uninformative

Tutorial on linear regression inference

Statistical inference and models

Point estimates, confidence intervals and hypothesis tests

Tutorial on inference about a mean

Tutorial on linear regression inference

Linear regression

- Suppose observations are from $(Y_1, X_1), \ldots, (Y_n, X_n) \sim F_{YX}$

  $\Rightarrow$ Goal is to learn the relationship between the RVs Y and X

- A workhorse approach is to model the regression function

$$r(x) = \mathbb{E}\left[Y \,\middle|\, X = x\right] = \int_{-\infty}^{\infty} y f_{Y|X}(y|x) dy$$

- The simple linear regression model specifies that given $X_i = x_i$

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, \ldots, n$$

  - The y_i's are modeled as noisy samples of the line $r(x) = \beta_0 + \beta_1 x$
  - Errors ϵ_i are i.i.d., with $\mathbb{E}\left[\epsilon_i | X_i = x_i\right] = 0$ and $\text{var}\left[\epsilon_i | X_i = x_i\right] = \sigma^2$

- With the linear model, regression amounts to parametric inference

$$\hat{r}(x) \Leftrightarrow [\hat{\beta}_0, \hat{\beta}_1]^T$$

Multiple linear regression

- More generally, suppose we observe data $(y_1, \mathbf{x}_1), \ldots, (y_n, \mathbf{x}_n)$

  $\Rightarrow$ Each input $\mathbf{x}_i = [x_{i1}, \ldots, x_{ip}]^T$ is a $p \times 1$ feature vector

- The multiple linear regression model specifies

$$y_i = \sum_{j=1}^{p} x_{ij}\beta_j + \epsilon_i = \boldsymbol{\beta}^T \mathbf{x}_i + \epsilon_i, \quad i = 1, \ldots, n$$

  - Typically $x_{i1} = 1$ for all i, providing an intercept term
  - Errors ϵ_i are i.i.d., with $\mathbb{E}\left[\epsilon_i | \mathbf{X}_i = \mathbf{x}_i\right] = 0$ and $\mathrm{var}\left[\epsilon_i | \mathbf{X}_i = \mathbf{x}_i\right] = \sigma^2$

- Can be compactly represented as $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, defining

$$\mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \mathbf{X} = \begin{bmatrix} x_{11} & \ldots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \ldots & x_{np} \end{bmatrix}, \boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}, \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

Least-squares estimator

- A sound estimate $\hat{\boldsymbol{\beta}}$ minimizes the residual sum of squares (RSS)

$$\text{RSS}(\boldsymbol{\beta}) = \sum_{i=1}^{n} (y_i - \boldsymbol{\beta}^T \mathbf{x}_i)^2 = \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2$$

  $\Rightarrow$ Residuals are the distances from y_i to hyperplane $r(\mathbf{x}) = \boldsymbol{\beta}^T \mathbf{x}$

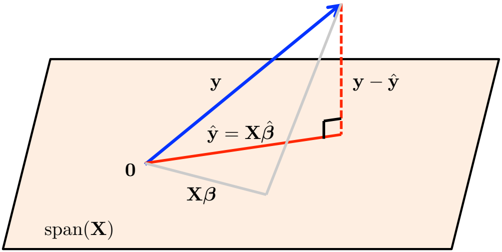
- **Def:** The least-squares estimator (LSE) $\hat{\boldsymbol{\beta}}_n$ is the solution to

$$\hat{\boldsymbol{\beta}}_n = \arg\min_{\boldsymbol{\beta}} \text{RSS}(\boldsymbol{\beta})$$

- Carrying out the optimization yields the LSE $\hat{\boldsymbol{\beta}}_n = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$
  $\Rightarrow$ Only defined if $\mathbf{X}^T\mathbf{X}$ invertible $\Leftrightarrow$ $\mathbf{X}$ has full column rank p

Geometry of the LSE

- In least squares we seek the vector $\hat{\mathbf{y}} = \mathbf{X}\hat{\beta} \in \text{span}(\mathbf{X})$ closest to $\mathbf{y}$



- Solution: Orthogonal projection of $\mathbf{y}$ onto $\text{span}(\mathbf{X})$, i.e., (let $\mathbf{X} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T$)

$$\hat{\mathbf{y}} = P_{\mathbf{X}}(\mathbf{y}) = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} = \mathbf{U}\mathbf{U}^T\mathbf{y}$$

- The residual $\mathbf{y} - \hat{\mathbf{y}}$ lies in the orthogonal complement $(\text{span}(\mathbf{X}))^{\perp}$
  - $\Rightarrow$ This way $\text{RSS}(\hat{\beta}) = \|\mathbf{y} - \hat{\mathbf{y}}\|^2$ is minimum

Properties of the LSE

- LSE $\hat{\boldsymbol{\beta}}_n = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$ is a linear combination of the random $\mathbf{y}$

P1) Unbiasedness: $\mathbb{E}\left[\hat{\boldsymbol{\beta}}_n \,\middle|\, \mathbf{X}\right] = \boldsymbol{\beta}$ with $\text{var}\left[\hat{\boldsymbol{\beta}}_n \,\middle|\, \mathbf{X}\right] = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$

P2) Consistency: $\hat{\boldsymbol{\beta}}_n \xrightarrow{p} \boldsymbol{\beta}$ as the sample size n increases

P3) Asymptotic Normality: For large n, one has $\hat{\boldsymbol{\beta}}_n \sim \mathcal{N}(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1})$

P4) If errors $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$, then $\hat{\boldsymbol{\beta}}_n \sim \mathcal{N}(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1})$ exactly; and Efficiency: No other unbiased estimator of $\boldsymbol{\beta}$ has smaller variance

- Ex: Can use the LSE to create confidence intervals for each β_j, i.e.,

$$C_n = \left(\hat{\beta}_j - z_{\alpha/2}\hat{\text{se}}(\hat{\beta}_j), \hat{\beta}_j + z_{\alpha/2}\hat{\text{se}}(\hat{\beta}_j)\right)$$

$\Rightarrow$ By asymptotic (or exact) Normality, $\text{P}\left[\beta_j \in C_n\right] \approx 1 - \alpha$

$\Rightarrow$ Note that $\hat{\text{se}}(\hat{\beta}_j) = \hat{\sigma}\sqrt{[(\mathbf{X}^T\mathbf{X})^{-1}]_{jj}}$, where $\hat{\sigma}^2 = \frac{RSS(\hat{\boldsymbol{\beta}})}{n-p}$

Hypothesis testing and prediction

Ex: Consider the hypothesis test regarding the parameter β_j

$$H_0 : \beta_j = \beta_j^{(0)} \quad \text{versus} \quad H_1 : \beta_j \neq \beta_j^{(0)}$$

- By asymptotic (or exact) Normality of the LSE, an α-level test is

$$\text{Reject } H_0 \text{ if } T_j := \left| \frac{\hat{\beta}_j - \beta_j^{(0)}}{\hat{\text{se}}(\hat{\beta}_j)} \right| > z_{\alpha/2}$$

Ex: Can predict an unobserved value $Y_* = y_*$ from a given $\mathbf{x}_*$ via

$$y_* = \mathbf{x}_*^T \hat{\boldsymbol{\beta}}$$

- May define a notion of standard error for y_*, and predictive intervals
  $\Rightarrow$ Should account for the variability in estimating $\boldsymbol{\beta}$ and in ϵ_*

The LSE as a MLE

- Suppose that conditioned on $\mathbf{X}_i = \mathbf{x}_i$, the errors ϵ_i are i.i.d. Normal

  $\Rightarrow$ The conditional PDF is $f(\epsilon_i \,|\, \mathbf{x}_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{\epsilon_i^2}{2\sigma^2} \right\}$

- Assume σ^2 is known. The (conditional) likelihood function is

$$\mathcal{L}_n(\boldsymbol{\beta}) = \prod_{i=1}^{n} f(y_i \,|\, \mathbf{x}_i; \boldsymbol{\beta}) \propto \exp\left\{ -\sum_{i=1}^{n} \frac{(y_i - \boldsymbol{\beta}^T \mathbf{x}_i)^2}{2\sigma^2} \right\}$$

  $\Rightarrow$ The log-likelihood is $\ell_n(\boldsymbol{\beta}) \propto -\text{RSS}(\boldsymbol{\beta})$

- The MLE $\hat{\boldsymbol{\beta}}_n^{ML}$ maximizes the log-likelihood function, thus

$$\hat{\boldsymbol{\beta}}_n^{ML} = \arg\max_{\boldsymbol{\beta}} \ell_n(\boldsymbol{\beta}) = \arg\min_{\boldsymbol{\beta}} \text{RSS}(\boldsymbol{\beta}) = \hat{\boldsymbol{\beta}}_n^{LS}$$

- **Take-home:** Under a linear-Gaussian model the LSE is also a MLE

MAP with Gaussian data model and prior

- Consider again Gaussian errors, i.e., $f(\epsilon_i \,|\, \mathbf{x}_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{\epsilon_i^2}{2\sigma^2}\right\}$
  - $\Rightarrow$ Gaussian prior to model the parameters: $\boldsymbol{\beta} \sim \mathcal{N}(\mathbf{0}, \tau^2\mathbf{I})$
  - $\Rightarrow$ Variances σ^2 and τ^2 assumed known. Define $\lambda := (\frac{\sigma}{\tau})^2$
- Bayesian approach: posterior $F_{\boldsymbol{\beta}|\mathbf{Y},\mathbf{X}}$ is Gaussian, with log-density

$$\log f(\boldsymbol{\beta} \,|\, \mathbf{Y}, \mathbf{X}) \propto -\sum_{i=1}^{n}(y_i - \boldsymbol{\beta}^T\mathbf{x}_i)^2 - \lambda \sum_{j=1}^{p} \beta_j^2$$

- MAP estimator $\hat{\boldsymbol{\beta}}_n^{MAP} := \arg\max_{\boldsymbol{\beta}} f(\boldsymbol{\beta} \,|\, \mathbf{Y}, \mathbf{X})$ is thus the solution to

$$\hat{\boldsymbol{\beta}}_n^{MAP} = \arg\min_{\boldsymbol{\beta}} \text{RSS}(\boldsymbol{\beta}) + \lambda\|\boldsymbol{\beta}\|_2^2$$

- Carrying out the optimization yields $\hat{\boldsymbol{\beta}}_n^{MAP} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}$
  - $\Rightarrow$ Recover the LSE as $\lambda \to 0 \Leftrightarrow$ Uninformative prior when $\tau^2 \to \infty$

Ridge regression

- Non-Bayesian, ℓ_2-norm penalized LSE also known as ridge regression

$$\hat{\boldsymbol{\beta}}^{ridge} = \arg\min_{\boldsymbol{\beta}} \text{RSS}(\boldsymbol{\beta}) + \lambda \|\boldsymbol{\beta}\|_2^2$$

- For $\lambda > 0$, the ridge estimator $\hat{\boldsymbol{\beta}}^{ridge} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}$
  - Differs from the LSE $\hat{\boldsymbol{\beta}}^{LS} := \arg\min_{\boldsymbol{\beta}} \text{RSS}(\boldsymbol{\beta})$
  - Is biased, and $\text{bias}(\hat{\boldsymbol{\beta}}^{ridge})$ increases with λ
  - Is well defined even when $\mathbf{X}$ is not of full rank

- In exchange for bias, potential to reduce variance below $\text{var}\left[\hat{\boldsymbol{\beta}}^{LS}\right]$
  - Ex: Large $\text{var}\left[\hat{\boldsymbol{\beta}}^{LS}\right]$ when $\mathbf{X}$ nearly rank-deficient, unstable $(\mathbf{X}^T\mathbf{X})^{-1}$

- From bias-variance MSE decomposition, fruitful tradeoff may yield

$$\text{MSE}(\hat{\boldsymbol{\beta}}^{ridge}) < \text{MSE}(\hat{\boldsymbol{\beta}}^{LS})$$

$\Rightarrow$ Tradeoff depends on λ, chosen subjectively or via cross validation

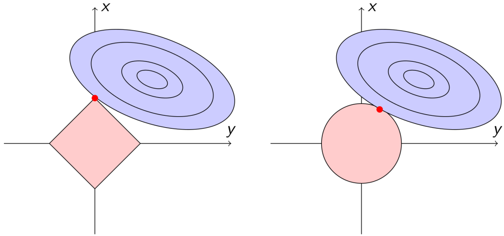
Complexity-penalized LSE

- Ridge an instance from the general class of complexity-penalized LSE

$$\hat{\boldsymbol{\beta}}^{J} = \arg\min_{\boldsymbol{\beta}} \text{RSS}(\boldsymbol{\beta}) + \lambda J(\boldsymbol{\beta})$$

  - Function $J(\cdot)$ penalizes (i.e., constrains) the parameters in $\boldsymbol{\beta}$
  - Constrained parameter space Θ effects 'less complex' models
  - Tuning λ balances goodness-of-fit and model complexity

- Ex: ℓ_1-norm penalized LSE for sparsity, i.e., variable selection



Glossary

- Statistical inference
- Outcome or response
- Predictor, feature or regressor
- (Non) parametric model
- Nuisance parameter
- Regression function
- Prediction
- Classification
- Point and set estimation
- Estimator and estimate
- Standard error
- Consistent estimator
- Confidence interval
- Hypothesis test
- Null hypothesis
- Test statistic and critical value
- Method of moments estimator
- Maximum likelihood estimator
- Likelihood function
- Significance level and $p - value$
- Prior and posterior distribution
- Multiple linear regression
- Least-squares estimator