Humboldt-Universität zu Berlin

Methods of Statistics (M25)

Lecturer: Prof. Dr. Sven Wang
WS 24/25

Last Update: Friday 6th December, 2024

Contents

1 Intro and Disclaimer 5

2 Basic Statistical Concepts 6

3 Parameter Estimation 8
- 3.1 Maximum Likelihood Principle 10
- 3.2 Bayesian method 11

4 Decision Theory 15
- 4.1 Another optimality concept: Admissibility 18

5 Confidence Sets 20
- 5.1 Hypothesis Testing 22
  - 5.1.1 Basic Definitions 22
  - 5.1.2 Type I and Type II Errors 22
  - 5.1.3 Level and Uniformly Most Powerful Tests 22
  - 5.1.4 The Neyman-Pearson Lemma 23
  - 5.1.5 Proof of the Neyman-Pearson Lemma 23

6 Lecture 6: Neyman-Pearson Lemma and Likelihood Ratio Tests 25

7 Linear Models 30
- 7.1 Introduction to Linear Regression Models 30
- 7.2 Simple Linear Regression Model 30
  - 7.2.1 Statistical Model 30
  - 7.2.2 Likelihood Function 31
  - 7.2.3 Maximum Likelihood Estimation (MLE) 31
- 7.3 General Linear Models 32
- 7.4 Least Squares Estimation 33
  - 7.4.1 Geometric Interpretation 33
- 7.5 Representation for the Least Squares Estimator 34
- 7.6 Optimality of the Least Squares Estimator 34
- 7.7 Conclusion 37

8 Lecture 9: Ridge Regression and Its Properties 38

9 Lecture 10: Confidence Sets and Hypothesis Testing in Linear Models 46
9.0.1 The t- and F-Tests . . . 49
9.0.2 General Linear Hypothesis Testing Problems . . . 50
9.0.3 The t- and F-Tests . . . 52
9.0.4 General Linear Hypothesis Testing Problems . . . 52

10 Lecture 11 55
10.1 Introduction . . . 55
10.2 Hypothesis Testing in Linear Models . . . 55
10.2.1 Residual Sum of Squares (RSS) . . . 55
10.2.2 The F-Test Theorem . . . 56
10.3 Analysis of Variance (ANOVA) . . . 57
10.3.1 Motivation . . . 58
10.3.2 Model Specification . . . 58
10.3.3 Factor Model . . . 58
10.3.4 Hypothesis Testing in ANOVA . . . 58
10.3.5 Decomposition of RSS . . . 59
10.3.6 ANOVA F-Test . . . 60
10.4 Conclusion . . . 62

11 Lecture 12 63
11.1 Analysis of Variance (ANOVA) . . . 63
11.1.1 Linear Model for ANOVA . . . 63
11.1.2 Ordinary Least Squares (OLS) Estimation . . . 63
11.1.3 Theorem: Properties of the ANOVA Model . . . 63
11.2 Exponential Families . . . 66
11.2.1 General Model and Hierarchy . . . 66
11.2.2 Regularity Assumptions . . . 66
11.2.3 Score Function . . . 66
11.2.4 Fisher Information . . . 67
11.2.5 Additional Regularity Assumptions . . . 67
11.2.6 Properties of the Score and Fisher Information . . . 67
11.3 Uniformly Minimum Variance Unbiased Estimators (UMVUE) . . . 68
11.3.1 Definition . . . 68
11.3.2 Remarks . . . 68
11.3.3 Cramér-Rao Lower Bound . . . 69
11.3.4 Invertibility of the Fisher Information Matrix . . . 70
11.4 Conclusion . . . 70

12 Lecture 13 71

13 Lecture 14: Exponential Families and Generalized Linear Models 80

13.1 Exponential Families . . . 80

13.1.1 Definition . . . 80

13.1.2 Properties . . . 80

13.1.3 Lemma . . . 81

13.1.4 Examples . . . 82

13.2 Maximum Likelihood Estimators and UMVUE in Exponential Families . 83

13.2.1 Theorem . . . 83

13.3 Generalized Linear Models (GLMs) . . . 85

13.3.1 Motivation . . . 85

13.3.2 Definition . . . 85

13.3.3 Remarks . . . 85

13.3.4 Examples . . . 86

13.3.5 Link Functions . . . 87

13.3.6 Estimation in GLMs . . . 87

13.3.7 Inference in GLMs . . . 87

13.3.8 Advantages of GLMs . . . 87

13.4 Conclusion . . . 88

14 Lecture 15: GLMs & Model Selection 89

14.1 Model Selection . . . 92

1 Intro and Disclaimer

These lecture notes are based on the material presented by Professor Wang during class and are written by students. They may contain errors or omissions. Please refer to the in-person lectures and the literature on Moodle for accurate and authoritative information. If you find an error or want to help, please send an email to:

said.kassner@student.hu-berlin.de

stephensonmonroe@gmail.com

salihiad@hu-berlin.de

2 Basic Statistical Concepts

Lecture 1

Here is the literature that the class is based on.

Literature:

- WS 19/20 R. Altmeyer *"Gliederung Methoden der Statistik"*
- L. Wasserman, *All of Statistics*
- M. Trabs, K. Krenz, M. Jirak and M. Reiss. *Methoden der Statistik.*
- Hastie, Tibshirani, et al., *Elements of Statistical Learning*

Let's start with a (simplest possible) example:

Example 1 (Polling). Consider a poll with two answers A and B (representing political parties).

- N = total number of votes
- M = total number of votes supporting party A

Poll Definitions:

- n = size of the poll
- $x = (x_1, ..., x_n)$ = responses, where:

$$x_i = \begin{cases} 0 & \text{if the i-th person supports B} \\ 1 & \text{if the i-th person supports A} \end{cases}$$

Additional Assumptions:

- n-times, we select a person randomly from the set $\{1, ..., N\}$, and record their (truthful) response.
- Every asked person responds (i.e., no selection bias).
- People can be asked repeatedly.

Aim of the Poll: The aim of the poll is to estimate the fraction of party A supporters. This can be written as:

$$\theta = \frac{M}{N} \in [0,1]$$

An intuitive estimate of θ is:

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

Question: Is this a good (or best possible) estimator? What properties does it have?

To answer this, we formalize some statistical notions.

Definition 1 (Sample space). A *sample space* is a measurable space $(\mathcal{X}, \mathcal{F})$, i.e., a set $\mathcal{X}$ with a σ-algebra $\mathcal{F}$, in which our statistical observations take values.

Definition 2 (Statistical model). Let $(\mathcal{X}, \mathcal{F})$ be some sample space and let Θ be a set, which we call the *parameter space.* A *statistical model* is a family of probability measures $\{P_\theta : \theta \in \Theta\}$ on $(\mathcal{X}, \mathcal{F})$.

Remark 1. Often, $(\mathcal{X}, \mathcal{F})$ is a "product space." For example, in Example 1, $\mathcal{X} = \{0,1\}^n$, and each P_θ is a product distribution, i.e., $x_1, \ldots, x_n$ are independent, identically distributed. Then we say $\{P_\theta : \theta \in \Theta\}$ is an *iid statistical model.*

Remark 2 (Back to Example 1). Here:

- $\mathcal{X} = \{0,1\}^n$
- $\Theta = [0,1]$
- $\mathcal{F} = \mathcal{P}(\{0,1\}^n)$
- $P_\theta = (\text{Bernoulli}(\theta))^{\otimes n}$

Remark 3. If every person could only be asked once, we would have $P_\theta = \text{Hypergeometric}(N, M, n)$, which "converges" to the Bernoulli model as $N, M \to \infty$. We might have to discretize Θ and take $\theta = \frac{M}{N}$. (Exercise: Think about it!)

3 Parameter Estimation

Assume that $\Theta \subseteq \mathbb{R}^p$, for $p \geq 1$. This is the setting of parametric statistics. [Assume Θ is measurable.]

Definition 3 (Estimator). An estimator for $\theta \in \Theta$ is any measurable function:

$$\hat{\theta} : (\mathcal{X}, \mathcal{F}) \to \Theta.$$

Any function that, based on some data $x \in \mathcal{X}$, outputs a guess / estimate $\hat{\theta}(x) \in \Theta$.

Lecture 2

Last time: Statistical model = family of probability measures on $(\mathcal{X}, \mathcal{F})$ indexed by $\theta \in \Theta$.

Sample space: $(\mathcal{X}, \mathcal{F})$

Estimator: = measurable function $(\mathcal{X}, \mathcal{F}) \to \Theta$

Now, what are some desirable properties we would like to have?

Definition 4 (Unbiased estimator). Let $\Theta = \mathbb{R}^p$ (measurable), $p \geq 1$. An estimator $\hat{\theta}$ is unbiased if

$$\mathbb{E}_\theta[\hat{\theta}] = \mathbb{E}_{\mathbb{P}_\theta}[\hat{\theta}] = \theta, \text{ for all } \theta \in \Theta.$$

Where $\mathbb{E}_\theta[\cdot] = \mathbb{E}_{\mathbb{P}_\theta}[\cdot]$ denotes expectation under the law $\mathbb{P}_\theta$.

In more explicit terms:

$$\mathbb{E}_{x \sim \mathbb{P}_\theta}[\hat{\theta}(x)] = \theta \quad \forall \theta$$

Remark 4 (Unbiasedness). Unbiasedness means ”no systematic errors.” However, we'd also like a ”good” $\hat{\theta}$ to be concentrated around the data-generating parameter.

Definition 5 (Consistent estimator). Let $(\mathbb{P}_\theta^n)_{\theta \in \Theta}$ be a sequence of statistical models ($n \geq 1$), on the same parameter space Θ not depending on $n \geq 1$.

Let $\hat{\theta}_n$ be a sequence of estimators. Then $\hat{\theta}_n$ is called consistent if for every $\theta \in \Theta$,

$$\hat{\theta}_n \xrightarrow{\mathbb{P}_\theta^n} \theta$$

or explicitly, for every $\epsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}_\theta^n(|\hat{\theta}_n - \theta|> \epsilon) = 0.$$

Back to Example 1:

- $X_i = \{0,1\}^n$
- $\Theta = [0,1]$
- $\mathbb{P}_\theta^n = \text{Bernoulli}(\theta)^{\otimes n}$
- $\hat{\theta}_n = \frac{1}{n}\sum_{i=1}^n X_i$

Unbiasedness:

Let $\theta \in \Theta$, then

$$\mathbb{E}_\theta\left[\frac{1}{n}\sum_{i=1}^n X_i\right] = \frac{1}{n}\sum_{i=1}^n \mathbb{E}_\theta[X_i] = \frac{1}{n}\sum_{i=1}^n \theta = \theta.$$

Thus, $\mathbb{E}_\theta[\hat{\theta}_n] = \theta$.

Consistency:

- We could use the Weak Law of Large Numbers (WLLN).
- Alternatively,

$$\text{Var}_\theta(\hat{\theta}_n) = \frac{1}{n^2}\sum_{i=1}^n \text{Var}_\theta(X_i) = \frac{1}{n^2}\sum_{i=1}^n \theta(1-\theta) = \frac{\theta(1-\theta)}{n}$$

  which tends to zero as $n \to \infty$.

It follows: For every $\epsilon > 0$,

$$\mathbb{P}_\theta^n(|\hat{\theta}_n - \theta|> \epsilon) = \mathbb{P}_\theta^n(|\hat{\theta}_n - \mathbb{E}[\hat{\theta}_n]|> \epsilon)$$

By Markov's inequality:

$$\mathbb{P}_\theta^n(|\hat{\theta}_n - \mathbb{E}[\hat{\theta}_n]|> \epsilon) \leq \frac{\mathbb{E}_\theta\left[(\hat{\theta}_n - \mathbb{E}[\hat{\theta}_n])^2\right]}{\epsilon^2} = \frac{\text{Var}_\theta(\hat{\theta}_n)}{\epsilon^2} = \frac{\theta(1-\theta)}{n\epsilon^2}$$

which tends to zero as $n \to \infty$. Thus,

$$(\hat{\theta}_n : n \geq 1) \text{ is consistent.} \qquad \square$$

3.1 Maximum Likelihood Principle

Is there another way to motivate $\hat{\theta}_n = \frac{1}{n}\sum_{i=1}^n X_i$? Yes, it turns out it is the maximum likelihood estimator, i.e.,

MLE = "parameter which assigns the highest probability to the observed data."

In our example, each $\mathbb{P}_\theta^n$ has a probability density (likelihood)

$$\begin{aligned}\mathbb{P}_\theta^n(x) = \prod_{i=1}^n \mathbb{P}_\theta(x_i) = \prod_{i=1}^n \theta^{x_i}(1-\theta)^{1-x_i} \\ = \theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i}.\end{aligned}$$

Fixing $x \in \{0,1\}^n$ and maximizing in $\theta \in [0,1]$ gives the following:

- If $\sum_{i=1}^n x_i = 0$, then $\hat{\theta}_n = 0$ is the maximizer.
- If $\sum_{i=1}^n x_i = n$, then $\hat{\theta}_n = 1$ is the maximizer.
- If $\hat{\theta}_n \in \{1, \dots, n-1\}$, then writing $S_n = \sum_{i=1}^n x_i$ gives:

$$\begin{aligned}\frac{\partial}{\partial\theta}\mathbb{P}_\theta^n(x) = S_n\theta^{S_n-1}(1-\theta)^{n-S_n-1} - (n-S_n)\theta^{S_n}(1-\theta)^{n-S_n-1} = 0 \\ \Leftrightarrow S_n(1-\theta) - \theta(n-S_n) \\ \Leftrightarrow \theta = \frac{S_n}{n}. \quad \square\end{aligned}$$

Definition 6 (Dominated statistical model & MLE). A model $(\mathbb{P}_\theta^n)_{\theta\in\Theta}$ is called dominated if there exists a measure μ on $(\mathcal{X}, \mathcal{F})$ such that for every $\theta \in \Theta$, $\mathbb{P}_\theta \ll \mu$ or equivalently (by Radon-Nikodym), for all $\theta \in \Theta$, there is a probability density $\frac{\mathrm{d}\mathbb{P}_\theta}{\mathrm{d}\mu}$ of $\mathbb{P}_\theta$ with respect to μ.

The MLE is defined as any $\hat{\theta} \in \Theta$ that maximizes the function

$$\theta \mapsto \frac{\mathrm{d}\mathbb{P}_\theta}{\mathrm{d}\mu}(x) = \mathbb{P}_\theta(x).$$

Remark 5 (Caveats).

- MLE might not be unique.
- MLE might not exist.
- It's not always clear that some selection $\hat{\theta}(x) \in \arg\max_\theta \mathbb{P}_\theta(x)$ is a measurable function of $x \in \mathcal{X}$. However, there are measurable selection theorems that permit a measurable choice of $\hat{\theta}$ under very general conditions.

Remark 6. In all the models we study, we will work with the Lebesgue measure (for continuous data) or the counting measure (for discrete data).

Example 2 (Normal model). Consider random samples $X_1, \dots, X_n \sim N(\mu, \sigma^2)$ for some unknown $\mu \in \mathbb{R}$, $\sigma^2 > 0$, and let λ denote the Lebesgue measure on $\mathbb{R}^n$.

$$\theta = (\mu, \sigma^2) \in \mathbb{R} \times (0, \infty).$$

Then the likelihood is:

$$\begin{aligned} L(\mu, \sigma^2 \mid x) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2} \left(\frac{X_i - \mu}{\sigma} \right)^2 \right) \\ &= (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2} \|X - \mu \cdot 1_n\|^2 \right), \end{aligned}$$

where by 1_n we denote the vector of ones of dimension n.

Here, the MLE is given as:

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} X_i \quad \text{[Sample mean]}, \qquad \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \hat{\mu})^2 \quad \text{[Sample variance]}.$$

$$\mathbb{E}_\theta[\hat{\mu}] = \mu, \quad \mathbb{E}_\theta[\hat{\sigma}^2] = \frac{n-1}{n} \sigma^2, \quad \text{so } \hat{\sigma}^2 \text{ is biased.}$$

$\Rightarrow$ MLE is not always unbiased, and not always a "good" method.

3.2 Bayesian method

Motivation In Bayesian statistics, a key element is the prior distribution, which we denote by π, reflecting our "beliefs" about the parameter $\theta \in \Theta$ before observing data (π is a probability measure on Θ).

A prior π, together with a model $(P_\theta : \theta \in \Theta)$, gives rise to a joint probability distribution for the pair $(\theta, x) \in \Theta \times \mathcal{X}$.

The Bayesian approach bases statistical inference on the posterior distribution of θ conditioned on x.

Joint probability:

$$(\theta, x) \mapsto \pi(\theta) P_\theta(x)$$

$$\text{conditional distribution of } \; x \mid \theta$$

Posterior:

$$\pi(\theta \mid x) = \frac{\pi(\theta)P_\theta(x)}{\int_\Theta \pi(\theta)P_\theta(x)\,d\theta}$$

Remark 7. Bayesian methods automatically generate ”error bars” because the posterior is not an estimator but a whole probability distribution.

Lecture 3

Definition 7 (Prior, Posterior, Bayes' Rule). Let $\mathcal{F}_\Theta$ be a σ-algebra on Θ, and suppose

$$\{P_\theta : \theta \in \Theta\}$$

is a dominated statistical model with densities $p_\theta(x)$, and assume that

$$(\theta, x) \mapsto p_\theta(x)$$

is ”jointly measurable” (i.e., w.r.t. $\sigma(\mathcal{F}_\Theta \times \mathcal{F})$).

Let Π be a prior distribution on Θ, with density $\pi(\theta)$ w.r.t. measure $\nu(\cdot)$. Then, define the posterior density

$$\pi(\theta \mid x) := \frac{p_\theta(x)\pi(\theta)}{\int_\Theta p_{\tilde{\theta}}(x)\,d\Pi(\tilde{\theta})}.$$

The corresponding probability measure $\Pi(\cdot \mid x)$ is called the posterior distribution:

$$\Pi(B \mid x) = \int_B \pi(\theta \mid x)\,d\nu(\theta), \quad B \in \mathcal{F}_\Theta.$$

$$= \frac{\int_B p_\theta(x)\,\pi(\theta)\,d\nu(\theta)}{\int_\Theta p_{\tilde{\theta}}(x)\,d\Pi(\tilde{\theta})},$$

$$= \frac{\int_B p_\theta(x)\,d\Pi(\theta)}{\int_\Theta p_{\tilde{\theta}}(x)\,d\Pi(\tilde{\theta})},$$

Remark 8. Think of $\Theta \subseteq \mathbb{R}^p$, $\nu(\cdot)$ as a Lebesgue measure, $\pi(\cdot)$ as a Lebesgue density.

Exception: $\Theta = \{0, 1\}$ in hypothesis testing. Then, we'd take $\nu(\cdot)$ to be the counting measure.

From the posterior, we can derive several estimators:

- **Maximum-a-posterior (MAP) estimator:**

$$\hat{\theta}_{\text{MAP}}(x) = \operatorname*{argmax}_{\theta \in \Theta} \pi(\theta \mid x).$$

- **Posterior mean:** Say $\Theta \subseteq \mathbb{R}^p$ convex

$$\hat{\theta}(x) = \int_\Theta \theta \, \pi(\theta \mid x) \, d\nu(\theta) \in \mathbb{R}^p.$$

Back to Example 1.1: *Binomial model:* $\mathcal{X} = \{0, 1, \ldots, n\}$, $p_\theta = \mathrm{Bin}(n, \theta)$, $\theta \in \Theta = [0, 1]$.

Prior (uniform): $\Pi = \mathrm{Unif}(0, 1)$.

We know:

$$\hat{\theta}_{\mathrm{MAP}} = \hat{\theta}_{\mathrm{MLE}} \quad \text{(for the uniform prior)},$$

$$\hat{\theta}_{\mathrm{MAP}} = \frac{X}{n}.$$

- **Posterior mean**:

$$\pi(\theta|x) = \frac{p_\theta(x)}{\int p_{\tilde{\theta}}(x) d\tilde{\theta}} \propto \binom{n}{k} \theta^x (1-\theta)^{n-x}.$$

- **Binomial distribution**:

$$\mathrm{Bin}(n, p)(k) = \binom{n}{k} p^k (1-p)^{n-k},$$

  where $k \in \{0, \ldots, n\}$ is the number of successes, and p is the probability of success, and n would be interpreted as the ”number of coin flips”.

$$\pi(\theta|x) \propto \theta^x (1-\theta)^{n-x}.$$

and

$$\int_0^1 \pi(\theta \mid x) d\nu = 1$$

We conclude that $\pi(\theta|x)$ is a **Beta-distribution** on $[0, 1]$,

$$\mathrm{Beta}(x+1, n-x+1).$$

The mean is given by:

$$\hat{\theta} = \frac{x+1}{n+2}.$$

Remark 9 (Beta distribution). The Beta distribution is defined as:

$$\mathrm{Beta}(a, b), \quad a, b \geq 0.$$

The probability density function of the Beta distribution is given by:

$$P_{\mathrm{Beta}(a,b)}(x) = x^a(1-x)^b.$$

Definition 8 (Conjugate Bayesian models). Let $(P_\theta : \theta \in \Theta)$ be a statistical model. Then, some family $\mathcal{D}$ of p.m.s on Θ is called *conjugate* if

$$\Pi \in \mathcal{D} \implies \Pi(\cdot|x) \in \mathcal{D} \quad \text{for all } x \in \mathcal{X}.$$

Examples:

- $(\mathrm{Bin}(n,\theta)) : \theta \in [0,1], \quad \mathcal{D} = \mathrm{Beta}(a,b),\ a, b \geq 0.$
- $(\mathcal{N}(\mu,\sigma^2)) : \mu \in \mathbb{R}, \quad \mathcal{D} = \left\{\mathcal{N}\left(\mu, n^2\right), \mu \in \mathbb{R}, n^2 > 0\right\},\ \sigma^2$ known.

4 Decision Theory

Here suppose that $\Theta \subseteq \mathbb{R}^p$.

Definition 9 (Loss function). A function $\ell : \Theta \times \mathbb{R}^p \to [0, \infty)$ is a *loss function* if for every $\theta \in \Theta$, $\ell(\theta, \cdot)$ is measurable. Given some estimator $\hat{\theta}$, the expected loss is

$$\mathcal{R}(\theta, \hat{\theta}) = \mathbb{E}_\theta \left[\ell(\theta, \hat{\theta}) \right].$$

Example: Take $\ell(\theta, \hat{\theta}) = \|\theta - \hat{\theta}\|^2_{\mathbb{R}^p}$. Then,

$$\mathcal{R}(\theta, \hat{\theta}) = \mathbb{E}_\theta \left[\|\theta - \hat{\theta}\|^2_{\mathbb{R}^p} \right]$$

is the mean squared error (MSE).

Proposition 1 (Bias-Variance Decomposition). *Let $\hat{\theta} \in L^2(\mathbb{P}_\theta)$. Then it holds that:*

$$R(\hat{\theta}, \theta) = (\mathbb{E}_\theta[\hat{\theta}] - \theta)^2 + \mathrm{Var}_\theta(\hat{\theta}).$$

Proof. We have

$$R(\hat{\theta}, \theta) = \mathbb{E}_\theta[(\hat{\theta} - \mathbb{E}_\theta[\hat{\theta}] + \mathbb{E}_\theta[\hat{\theta}] - \theta)^2]$$

Expanding the squared term:

$$= \mathbb{E}_\theta[(\hat{\theta} - \mathbb{E}_\theta[\hat{\theta}])^2] + (\mathbb{E}_\theta[\hat{\theta}] - \theta)^2 + 2\mathbb{E}_\theta[(\hat{\theta} - \mathbb{E}_\theta[\hat{\theta}])(\mathbb{E}_\theta[\hat{\theta}] - \theta)].$$

Since $\mathbb{E}_\theta[\hat{\theta} - \mathbb{E}_\theta[\hat{\theta}]] = 0$, the last term vanishes, leaving us with:

$$R(\hat{\theta}, \theta) = \mathrm{Var}_\theta(\hat{\theta}) + (\mathbb{E}_\theta[\hat{\theta}] - \theta)^2.$$

□

Definition 10 (Minimax Risk). Given an estimator $\hat{\theta}$ in a model $(\mathbb{P}_\theta : \theta \in \Theta)$, the maximal risk of $\hat{\theta}$ is

$$\sup_{\theta \in \Theta} R(\theta, \hat{\theta}).$$

The minimax risk of a model $(\mathbb{P}_\theta : \theta \in \Theta)$ is given as

$$\inf_{\hat{\theta}} \sup_{\theta \in \Theta} R(\theta, \hat{\theta}),$$

where the infimum is taken over all estimators. An estimator is called minimax if

$$\sup_{\theta\in\Theta} R(\theta,\hat{\theta}) = \inf_{\hat{\theta}} \sup_{\theta\in\Theta} R(\theta,\hat{\theta}).$$

Definition 11 (Bayes Risk). Given a prior π on Θ, the π-Bayes risk of a decision rule δ for the loss function L is defined as

$$R_\Pi(\delta) = \mathbb{E}_\Pi[R(\delta,\theta)] = \int_\Theta R(\delta,\theta)\pi(\theta)d\theta = \int_\Theta \int_{\mathcal{X}} L(\delta(x),\theta)\pi(\theta)p_\theta(x)dxd\theta.$$

A Π-Bayes decision rule $\hat{\theta}_\Pi$ is any decision rule that minimizes $R_\Pi(\hat{\theta})$.

SW: ℓ instead of L below, $p_\theta(x)$ instaed of $f(x,\theta)$ Note: Has been corrected.

Definition 12 (Posterior Risk). For a Bayesian model, the posterior risk R_Π is defined as the average loss under the posterior distribution for some observation $x\in\mathcal{X}$:

$$R_{\Pi(\cdot|x)}(\delta) = \mathbb{E}_\Pi[\ell(\delta(x),\theta)|x].$$

Here, the notation $\mathbb{E}_\Pi[\cdot|x]$ stands for the expectation under the posterior distribution.

Proposition 2 (Bayes Risk and Posterior Risk). *An estimator δ that minimizes the Π-posterior risk R_Π also minimizes the π-Bayes risk R_π.*

Proof. The π-Bayes risk can be rewritten as

$$\begin{aligned}
R_\pi(\delta) &= \int_\Theta \mathbb{E}_\theta[\ell(\delta(X),\theta)]\pi(\theta)d\theta \\
&= \int_\Theta\int_{\mathcal{X}} \ell(\delta(x),\theta)\pi(\theta)p_\theta(x)dxd\theta \\
&= \int_{\mathcal{X}}\int_\Theta \ell(\delta(x),\theta)\frac{p_\theta(x)\pi(\theta)}{\int_\Theta p_{\theta'}(x)\pi(\theta')d\theta'} \times \underbrace{\int_\Theta p_{\theta'}(x)\pi(\theta')d\theta'}_{=:n(x)\geq 0} dxd\theta \\
&= \int_{\mathcal{X}} \mathbb{E}_\Pi[\ell(\delta(x),\theta)|x]n(x)dx.
\end{aligned}$$

[Notation $n(x)$ motivated by the word ‘normalising constant’].

Let δ_Π be a decision rule that minimizes the posterior risk, i.e., such that for all $x\in\mathcal{X}$,

$$\mathbb{E}_\Pi[\ell(\delta_\Pi(x),\theta)|x] \leq \mathbb{E}_\Pi[\ell(\delta(x),\theta)|x].$$

Multiplying by $m(x)\geq 0$ and integrating on both sides over $\mathcal{X}$ yields the desired result.□

Example 3. For the quadratic risk with the squared-loss, the posterior risk is minimized by taking $\delta(X) = \mathbb{E}_\Pi[\theta|X]$, by minimizing the quadratic function in δ. Other losses will give other ways to minimize the posterior risk, and other Bayes decision rules.

Proposition 3. *Let $(\mathcal{X}, \mathcal{F}, (\mathbb{P}_\theta)_{\theta\in\Theta})$ be a statistical model and let $\hat{\theta}$ be an estimator. Then we have*

$$\sup_{\theta\in\Theta} R(\hat{\theta}, \theta) = \sup_{\Pi} \int_\Theta R(\hat{\theta}, \theta)\, \Pi(d\theta),$$

where the supremum is taken over all prior distributions Π*.*

Proof. Obviously, we have

$$\int_\Theta R(\hat{\theta}, \theta)\, \Pi(d\theta) \le \sup_{\theta\in\Theta} R(\hat{\theta}, \theta).$$

On the other hand, by using the prior distributions δ_θ (Dirac measure on $\theta \in \Theta$), we obtain

$$\sup_{\Pi} \int_\Theta R(\hat{\theta}, \theta)\, \Pi(d\theta) \ge \int_\Theta R(\hat{\theta}, \theta)\, \delta_\theta(d\theta) = R(\hat{\theta}, \theta).$$ □

[Note: In the following we use the notation δ for decision rules while on the blackboard we used $\hat{\theta}$ or $\tilde{\theta}$. If you want to adjust this please contact me so that I can give you access.]

Proposition 4. *Let π be a prior on Θ such that*

$$R_\pi(\delta_\pi) = \sup_{\theta\in\Theta} R(\delta_\pi, \theta),$$

where δ_π is a π-Bayes rule. Then it holds that

1. *The rule δ_π is minimax.*
2. *If δ_π is unique Bayes, then it is unique minimax.*

Proof. Let δ be any decision rule. Then

$$\sup_{\theta\in\Theta} R(\delta, \theta) \ge \mathbb{E}_\pi[R(\delta, \theta)],$$

$$\int_\Theta R(\delta, \theta)\pi(\theta)\, d\theta \ge \mathbb{E}_\pi[R(\delta, \theta)],$$

$$\int_\Theta R(\delta, \theta)\pi(\theta)\, d\theta = R_\pi(\delta_\pi) = \sup_{\theta\in\Theta} R(\delta_\pi, \theta).$$

Taking the infimum over δ gives the result.

2. If δ_π is unique Bayes, the second inequality is strict for any $\delta' \neq \delta_\pi$. □

Corollary 1. *If a (unique) Bayes rule δ_π has constant risk in θ, then it is (unique) minimax.*

Proof. If a Bayes rule δ_π has constant risk, then

$$R_\pi(\delta_\pi) = \mathbb{E}_\pi[R(\delta_\pi, \theta)] = \sup_{\theta \in \Theta} R(\delta_\pi, \theta),$$

where $R(\delta_\pi, \theta)$ is constant in θ. Uniqueness of the Bayes rule implies uniqueness of the minimax rule, as in part 2 of the former proposition. □

Example 4. Hence, if the maximal risk of a Bayes rule δ_π equals the Bayes risk, then π is least favorable, and the corresponding Bayes rule is minimax.

- In a $\text{Bin}(n, \theta)$ model, let $\pi_{a,b}$ be a $\text{Beta}(a, b)$ prior on $\theta \in [0, 1]$. Then the unique Bayes rule for $\pi_{a,b}$ over the quadratic risk is the posterior mean $\delta_{a,b} = \bar{\theta}_{a,b}$. Trying to solve the equation

  $$R(\delta_{a,b}, \theta) = \text{const.} \quad \forall \theta \in [0, 1]$$

  we can find a prior π_{a^*,b^*} and a corresponding Bayes rule $\delta_{\pi_{a^*,b^*}}$ of constant risk. It is therefore unique minimax, and different from the MLE (see Examples sheet).
- In a $\mathcal{N}(\theta, 1)$ model, $\bar{X}_n$ is minimax, as proved later.

4.1 Another optimality concept: Admissibility

Definition 13. A decision rule δ is *inadmissible* if there exists δ' such that

$$R(\delta', \theta) \leq R(\delta, \theta) \quad \forall \theta \in \Theta \quad \text{and} \quad R(\delta', \theta) < R(\delta, \theta) \quad \text{for some } \theta \in \Theta.$$

Remark 10.

- The intuition is that there is no reason to choose an inadmissible estimator or decision rule: it would always be better to choose another estimator that dominates it.
- Admissibility is not the only criterion to evaluate an estimator: In most cases, a constant estimator will be admissible for the quadratic risk, but it is often not reasonable.

Proposition 5.

1. *A unique Bayes rule is admissible.*
2. *If δ is admissible and has constant risk, then it is minimax.*

Proof may be done in the Examples sheet.

Definition 14. For a vector $X \in \mathbb{R}^p$, the *James–Stein estimator* is defined as

$$\delta^{JS}(X) = \left(1 - \frac{p-2}{\|X\|^2}\right) X.$$

In a Gaussian model $X \sim \mathcal{N}(\theta, I_p)$ for $\theta \in \mathbb{R}^p$ (with a single observation, to simplify notation), the risk of the MLE is given by

$$R(\hat{\theta}_{\mathrm{MLE}}, \theta) = \mathbb{E}_\theta[\|X - \theta\|^2] = \sum_{j=1}^{p} \mathbb{E}_\theta[(X_j - \theta_j)^2] = p.$$

For $X \sim \mathcal{N}(\theta, I_p)$ with $p \geq 3$, the risk of the James–Stein estimator satisfies for all $\theta \in \mathbb{R}^p$

$$R(\delta^{JS}, \theta) < p.$$

5 Confidence Sets

Definition 15 (Confidence Set). Let $(\mathcal{X}, \mathcal{F}, (\mathbb{P}_\theta)_{\theta\in\Theta})$ be a statistical model. For a given $\alpha \in [0,1]$, consider sets $C_{1-\alpha}(x) \subseteq \Theta$ for each $x \in \mathcal{X}$. Then $C_{1-\alpha}(x)$ is called a random confidence set at level $1-\alpha$ (or with coverage probability $1-\alpha$) if

$$\forall \theta \in \Theta : \mathbb{P}_\theta(\theta \in C_{1-\alpha}) = \mathbb{P}_\theta(\{x \in \mathcal{X} : \theta \in C_{1-\alpha}(x)\}) \geq 1-\alpha.$$

Note: The following example was only started in Lecture 4 and may be fully covered in Lecture 5.

Example 5. Consider the statistical model $(\{0,1\}^n, \mathcal{P}(\{0,1\}^n), (\mathbb{P}_p)_{p\in[0,1]})$ with $\mathbb{P}_p = \mathrm{Ber}(p)^{\otimes n}$ and independent observations $X_k \sim \mathrm{Ber}(p)$, $k \in \{0, \dots, n\}$. We are looking for a confidence interval $C_{1-\alpha}$ around $\hat{p} = \overline{X}_n = \frac{1}{n}\sum_{k=1}^n X_k$, i.e.,

$$C_{1-\alpha} = [\overline{X}_n - a, \overline{X}_n + b] \quad \Big(C_{1-\alpha}(x) = [\overline{X}_n(x) - a(x), \overline{X}_n(x) + b(x)]\Big)$$

should satisfy (where a and b might be random) that

$$1-\alpha \leq \mathbb{P}_p\left(p \in C_{1-\alpha}\right) = \mathbb{P}_p\Big(\overline{X}_n - a \leq p \leq \overline{X}_n + b\Big) = \mathbb{P}_p\Big(-b \leq \overline{X}_n - p \leq a\Big).$$

Let $t \mapsto F_p^n(t) := \mathbb{P}_p\Big(\overline{X}_n - p \leq t\Big)$ be the distribution function. Then,

$$\begin{aligned}\mathbb{P}_p\Big(-b \leq \overline{X}_n - p \leq a\Big) &= \mathbb{P}_p\Big(\overline{X}_n - p \leq a\Big) - \mathbb{P}_p\Big(\overline{X}_n - p < -b\Big)\\ &= F_p^n(a) - F_p^n(-b) + R_n,\end{aligned}$$

where $R_n = \mathbb{P}_p\Big(\overline{X}_n - p = -b\Big)$. Choose a, b as quantiles of $\mathbb{P}_p^n$, i.e., $a = \Big(F_p^n\Big)^{-1}(1-\alpha/2)$ and $-b = \Big(F_p^n\Big)^{-1}(\alpha/2)$ (with quantile function $t \mapsto \Big(F_p^n\Big)^{-1}(t) := \inf\{t \in \mathbb{R} : F_p^n(t) \geq t\}$).

However, F_p^n and thus a, b are unknown. Consider two possibilities:

Normal Approximation. It holds that $\mathbb{E}_p^n[X_k] = p$, $\sigma := \mathrm{Var}_p^n(X_k) = p(1-p)$. By the central limit theorem, we have

$$\frac{\sqrt{n}}{\sigma}\left(\overline{X}_n - p\right) = \frac{1}{\sqrt{n}}\sum_{k=1}^{n}\frac{X_k - p}{\sigma} \xrightarrow{d} N(0,1), \quad n \to \infty.$$

For $Z \sim N(0,1)$, it holds that

$$F_p^n(a) = \mathbb{P}_p^n\left(\overline{X}_n - p \le a\right) = \mathbb{P}_p^n\left(\frac{\sqrt{n}}{\sigma}\left(\overline{X}_n - p\right) \le \frac{\sqrt{n}}{\sigma}a\right) \approx \mathbb{P}(|Z|\le \frac{\sqrt{n}}{\sigma}a)$$

$$= \Phi\left(\frac{\sqrt{n}}{\sigma}a\right) = \Phi(z_\beta)$$

for $a := \frac{\sigma}{\sqrt{n}}z_\beta$ (where z_β is the β-quantile of the $N(0,1)$-distribution, i.e., $\Phi(z_\beta) = \beta$). In particular, $R_n = o(1)$ (i.e., $R_n \to 0$ as $n \to \infty$). For $a = b$ (since the $N(0,1)$-distribution is symmetric) and because $\Phi(-x) = 1 - \Phi(x)$, it follows that

$$\mathbb{P}_p^n(p \in C_{1-\alpha}) = F_p^n(a) - F_p^n(-a) + R_n \approx \Phi(z_\beta) - (1 - \Phi(z_\beta)) + o(1) = 2\Phi(z_\beta) - 1 + o(1).$$

For $\beta = 1 - \alpha/2$, $C_{1-\alpha}$ is an asymptotically correct confidence interval. However, p and therefore σ and a are unknown. Solutions:

- Estimate $\sigma = p(1-p) \le 1/4$ to widen the confidence interval.
- For the empirical variance $\hat{\sigma}^2 = \frac{1}{n}\sum_{k=1}^{n}(X_k - \overline{X}_n)^2$, we have $\hat{\sigma}^2 \to \sigma^2$ almost surely (by the law of large numbers). Using Slutsky's lemma (Lemma: For random variables $(X_n, Y_n)_{n\ge 1}$ with $X_n \xrightarrow{d} X$, $Y_n \xrightarrow{p} c \in \mathbb{R}$ (where c is deterministic), it holds that $X_n + Y_n \xrightarrow{d} X + c$ and $X_n \cdot Y_n \xrightarrow{d} c \cdot X$), it follows that

$$\frac{\sqrt{n}}{\hat{\sigma}}\left(\overline{X}_n - p\right) = \frac{\sigma}{\hat{\sigma}} \cdot \frac{\sqrt{n}}{\sigma}\left(\overline{X}_n - p\right) \xrightarrow{d} N(0,1), \quad n \to \infty.$$

From this, we derive $a = \frac{\hat{\sigma}}{\sqrt{n}}z_{1-\alpha/2}$ (randomly chosen).

$$\mathbb{P}_p^n(p \in C_{1-\alpha}) = \mathbb{P}_p^n\left(|\overline{X}_n - p|\le a\right) = \mathbb{P}_p^n\left(\left|\frac{\sqrt{n}}{\hat{\sigma}}(\overline{X}_n - p)\right| \le z_{1-\alpha/2}\right)$$

$$\approx \mathbb{P}(|Z|\le z_{1-\alpha/2}) = 2\Phi(z_{1-\alpha/2}) - 1 = 1 - \alpha.$$

5.1 Hypothesis Testing

5.1.1 Basic Definitions

Let $(P_\theta : \theta \in \Theta)$ be a statistical model, and let $\Theta = \Theta_0 \cup \Theta_1$ be a partition:

- A **statistical test** is a measurable function of the data $\varphi : (\mathcal{X}, \mathcal{F}) \to [0, 1]$.
- If $\varphi(x) \in \{0, 1\}$ for all $x \in \mathcal{X}$, then φ is a **non-randomized test**; otherwise, it is **randomized**.
- $H_0 : \theta \in \Theta_0$ is the **null hypothesis**.
- $H_1 : \theta \in \Theta_1$ is the **alternative hypothesis**.
- The map $\theta \to \beta_\varphi(\theta) = P_\theta(\varphi = 1)$ is called the **power function** of a test φ.

5.1.2 Type I and Type II Errors

- For $\theta \in \Theta_0$, $\beta_\varphi(\theta)$ represents the **Type I error** (wrongly rejecting the null).
- For $\theta \in \Theta_1$, $1 - \beta_\varphi(\theta)$ represents the **Type II error** (failing to reject the alternative when it is true).

$$1 \qquad \beta_\varphi(\theta) \quad 0 \qquad \Theta_0 \qquad \Theta_1 \qquad \Theta$$

Note:

$$1 - P_\theta(\varphi = 1) = P_\theta(\varphi = 0) = P_\theta \text{ (wrongly accepting the null)}$$

5.1.3 Level and Uniformly Most Powerful Tests

Definition 16 (Level). A test $\varphi : \mathcal{X} \to [0, 1]$ has **level** $\alpha \in [0, 1]$ if

$$\sup_{\theta \in \Theta_0} \beta_\varphi(\theta) \leq \alpha.$$

Definition 17 (Uniformly Most Powerful Test). Given a level $\alpha \in (0, 1)$, $\varphi : \mathcal{X} \to [0, 1]$ is called **uniformly most powerful (UMP)** if, for every other test φ' of level α and all $\theta \in \Theta_1$,

$$\beta_\varphi(\theta) \geq \beta_{\varphi'}(\theta).$$

5.1.4 The Neyman-Pearson Lemma

The Neyman-Pearson Lemma provides a basis for constructing the most powerful tests for simple hypotheses:

Theorem 1 (Neyman-Pearson Lemma). *Let $\Theta_0 = \{\theta_0\}$ and $\Theta_1 = \{\theta_1\}$ be simple hypotheses:*

1. ***Existence:*** *There exists a test φ and a constant $k \in [0, \infty)$ such that $P_{\theta_0}(\varphi = 1) = \alpha$, with*
$$\varphi(x) = \begin{cases} 1, & \textit{if } \frac{p_{\theta_1}(x)}{p_{\theta_0}(x)} > k \\ 0, & \textit{if } \frac{p_{\theta_1}(x)}{p_{\theta_0}(x)} < k \end{cases}$$
Here, p_{θ_1} and p_{θ_0} are densities with respect to some dominated measure μ.
2. ***Sufficiency:*** *If φ satisfies $P_{\theta_0}(\varphi = 1) = \alpha$ and the above form, then φ is a UMP level α test.*
3. ***Necessity:*** *If φ_k is UMP for level α, then it must be of the form shown above.*

5.1.5 Proof of the Neyman-Pearson Lemma

1. Define the likelihood ratio $r(x) = \frac{p_{\theta_1}(x)}{p_{\theta_0}(x)} \in [0, \infty)$. Let F_0 be the CDF of $r(x)$ under P_{θ_0}.
$$F_0(t) = P_{\theta_0}(r(x) \leq t).$$
Define $\alpha(t) = 1 - F_0(t) = P_{\theta_0}(r(x) > t)$ and note:
   - α is right-continuous:
$$\lim_{\epsilon \to 0} \alpha(t + \epsilon) = P_{\theta_0}(r(x) > t).$$
   - α is non-increasing.
   - α has left limits.

   α **is cadlag:** It is continuous from the right and has a left limit.
$$\text{There exists } k \in [0, \infty) \text{ such that } \alpha \leq \alpha(k^-) \quad \text{and} \quad \alpha \geq \alpha(k).$$
We define the test
$$\varphi(x) = \begin{cases} 1 & \text{if } r(x) > k, \\ \gamma & \text{if } r(x) = k, \\ 0 & \text{if } r(x) < k. \end{cases}$$

Set $\gamma = \frac{\alpha - \alpha(k)}{\alpha(k^-) - \alpha(k)}$.

The level of φ is

$$E_{\theta_0}[\varphi(x)] = P_{\theta_0}(\varphi(x) = 1).$$

$$= P_{\theta_0}(r(x) > k) + P_{\theta_0}(r(x) = k) \cdot \gamma = \alpha.$$

6 Lecture 6: Neyman-Pearson Lemma and Likelihood Ratio Tests

Neyman-Pearson Lemma

Power of a Test

The **power** of a test is defined as:

$$E_{\theta_1}[\varphi] = P_{\theta_1}(\varphi = 1)$$

Likelihood Ratio Test

The **likelihood ratio** is given by:

$$\Lambda(x) = \frac{p_{\theta_1}(x)}{p_{\theta_0}(x)} = r(x)$$

Likelihood Ratio (LR) Test

The LR test is defined as:

$$\varphi(x) = \begin{cases} 1 & \text{if } r(x) > k, \\ \gamma & \text{if } r(x) = k, \\ 0 & \text{if } r(x) < k, \end{cases}$$

where $k \in [0, \infty)$ and $\gamma \in [0, 1]$.

Note: LR tests are Uniformly Most Powerful (UMP) for simple hypothesis testing:

- Given a significance level α, if the LR test satisfies $E_{\theta_0}[\varphi] = \alpha$, it controls the Type I error.
- The LR test minimizes the Type II error:

$$E_{\theta_1}[\varphi] \geq E_{\theta_1}[\varphi'] \quad \forall \varphi'$$

Continuation of Proof (Part of UMP)

Let φ' be another level α test such that $E_{\theta_0}[\varphi'] \leq \alpha$.

Goal: Show that $E_{\theta_1}[\varphi] \geq E_{\theta_1}[\varphi']$.

Let μ be the dominating measure. Consider:

$$\int (\varphi(x) - \varphi'(x))\,(p_{\theta_1}(x) - kp_{\theta_0}(x))\,d\mu(x) = 0$$

Claim: $p \geq 0$.

Observation:

- If $p_{\theta_1}(x) - kp_{\theta_0}(x) > 0$, then $\frac{p_{\theta_1}(x)}{p_{\theta_0}(x)} > k$, implying $\varphi(x) = 1$.
- If $p_{\theta_1}(x) - kp_{\theta_0}(x) < 0$, then $\varphi(x) = 0$.
- If $p_{\theta_1}(x) - kp_{\theta_0}(x) = 0$, then the integrand is 0.

Thus, $p = 0$, leading to:

$$\int (\varphi - \varphi')p_{\theta_1}\,d\mu = \int (\varphi - \varphi')p_{\theta_0}\,d\mu = k\left[E_{\theta_0}[\varphi] - E_{\theta_0}[\varphi']\right] \geq 0$$

Therefore,

$$E_{\theta_1}[\varphi] \geq E_{\theta_1}[\varphi']$$

Part (3) UMP $\Rightarrow$ LR

Assume φ^* is a UMP test with $E_{\theta_0}[\varphi^*] = \alpha$. Let φ be the LR test satisfying $E_{\theta_0}[\varphi] = \alpha$.

Goal: Show that $\varphi = \varphi^*$ almost everywhere except on $\{r(x) = k\}$.

Define the sets:

$$x^+ = \{x : \varphi(x) > \varphi^*(x)\}, \quad x^- = \{x : \varphi(x) < \varphi^*(x)\}, \quad x^0 = \{x : \varphi(x) = \varphi^*(x)\}$$

$$\tilde{x} = (x^+ \cup x^-) \cap \{x : p_{\theta_1}(x) \neq kp_{\theta_0}(x)\}$$

It suffices to show $\mu(\tilde{x}) = 0$.

On $\tilde{x}$:

$$(\varphi - \varphi^*)(p_{\theta_1} - kp_{\theta_0}) > 0$$

If $\mu(\tilde{x}) > 0$, then:

$$\int_{\mathcal{X}} (\varphi - \varphi^*)(p_{\theta_1} - kp_{\theta_0})\,d\mu \geq 0$$

$$\int_{\tilde{x}} (\varphi - \varphi^*)(p_{\theta_1} - kp_{\theta_0})\,d\mu \geq 0$$

However,

$$E_{\theta_1}[\varphi] - E_{\theta_1}[\varphi^*] > k\left[E_{\theta_0}[\varphi] - E_{\theta_0}[\varphi^*]\right] \geq 0$$

This leads to a contradiction, implying $\mu(\tilde{x}) = 0$ and thus $\varphi = \varphi^*$ almost everywhere.

Example: Gaussian Location Model

Consider:

$$X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$$

Testing:

$$H_0 : \mu = \mu_0 \quad \text{vs.} \quad H_1 : \mu = \mu_1, \quad \mu_0 < \mu_1$$

The likelihood ratio is:

$$\frac{p_1(X_1, \ldots, X_n)}{p_0(X_1, \ldots, X_n)} = \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - \mu_1)^2 + \frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - \mu_0)^2\right)$$

Simplifying:

$$= \exp\left(-\frac{n}{2\sigma^2}(\mu_1^2 - \mu_0^2) - \frac{2(\mu_1 - \mu_0)}{\sigma^2}\sum_{i=1}^{n} X_i\right) \geq K_\alpha$$

This implies:

$$\frac{1}{n}\sum_{i=1}^{n} X_i \geq K_\alpha, \quad \text{for some } K_\alpha \in \mathbb{R}$$

To determine K_α:

$$\bar{X}_n := \frac{1}{n}\sum X_i \overset{H_0}{\sim} \mathcal{N}\left(\mu_0, \frac{\sigma^2}{n}\right)$$

Thus:

$$P_{H_0}\left(\bar{X}_n \geq K_\alpha\right) = 1 - \Phi\left(\frac{\sqrt{n}}{\sigma}(K_\alpha - \mu_0)\right)$$

Solving for K_α:

$$K_\alpha = \mu_0 + \frac{\sigma}{\sqrt{n}}\Phi^{-1}(1 - \alpha)$$

Therefore, the LR test is:

$$\varphi(X_1, \ldots, X_n) = \begin{cases} 1 & \text{if } \bar{X}_n \geq \mu_0 + \frac{\sigma}{\sqrt{n}}\Phi^{-1}(1 - \alpha), \\ 0 & \text{otherwise.} \end{cases}$$

Corollary

Consider simple hypothesis testing. Let φ be a UMP test at level α. Then:

$$\alpha = E_{H_0}[\varphi] \leq E_{\theta_1}[\varphi]$$

Suppose $E_{\theta_1}[\varphi] = E_{\theta_1}[\varphi_0]$. Then φ_0 is also UMP, implying φ_0 is an LR test:

$$\varphi_0 = \begin{cases} 1 & \text{if } \frac{p_{\theta_1}}{p_{\theta_0}} \geq K \quad \text{a.s., for some } K, \\ 0 & \text{otherwise.} \end{cases}$$

Since $\varphi_0 \in \{\varphi, \beta\}$, it follows that $p_{\theta_1} = K p_{\theta_0}$ almost surely.

Moreover:

$$\int p_{\theta_0}\, d\mu = K \int p_{\theta_0}\, d\mu = 1 \quad \Rightarrow \quad K = 1$$

Correspondence Theorem

Statement: There is a correspondence between tests and confidence regions.

$$\text{Tests} \quad \longleftrightarrow \quad \text{Confidence regions } C(x)$$

with

$$\Pr_\theta(\theta \in C(x)) \geq 1 - \alpha$$

and

$$\Pr_\theta(\phi_\theta(x) = 1) = \alpha$$

Theorem: Let $\{P_\theta : \theta \in \Theta\}$ be a statistical model and $\alpha \in (0, 1)$.

(i) If $C = C(X)$ is a level-α confidence set, then

$$\phi_{\theta_0}(x) = \mathbb{I}\{\theta_0 \notin C(x)\}$$

is a level-α test for $H_0 : \theta = \theta_0$ vs. $H_1 : \theta \neq \theta_0$.

(ii) If $\{\phi_\theta : \theta \in \Theta\}$ is a family of level-α tests, then

$$C(X) = \{\theta \in \Theta : \phi_\theta(X) = 0\}$$

is a $1 - \alpha$ confidence set.

Proof:

(i)

$$\Pr_{\theta_0}(\phi_{\theta_0} = 1) = \Pr_{\theta_0}(\theta_0 \notin C(X)) \leq \alpha$$

(ii)

$$\Pr_{\theta}(\theta \notin C(X)) = \Pr_{\theta}(\phi_\theta(X) = 1) \leq \alpha$$

UMPT Tests in Models with Monotone Likelihoods

Proposition: Let $\Theta \subseteq \mathbb{R}$. Consider testing:

$$H_0 : \theta \leq \theta_0 \quad \text{vs.} \quad H_1 : \theta > \theta_0,$$

for some $\theta_0 \in \mathbb{R}$.

Assume there exists a test statistic $T : X \to \mathbb{R}$ and a function $h : \mathbb{R} \times \Theta \times \Theta \to \mathbb{R}$ such that:

$$\frac{P_\theta(X)}{P_{\tilde{\theta}}(X)} = h(T(X), \theta, \tilde{\theta})$$

and for all $\theta \geq \tilde{\theta}$, the function $t \mapsto h(t, \theta, \tilde{\theta})$ is monotone increasing.

Conclusion: LR tests are also UMP for level α. Specifically, the LR test of $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_1$ for any $\theta_1 > \theta_0$ will be UMP.

7 Linear Models

7.1 Introduction to Linear Regression Models

Linear regression models are fundamental tools in statistical analysis, enabling us to understand and quantify the relationship between a dependent variable and one or more independent variables. In this lecture, we will explore the simplest form of linear regression, discuss the statistical framework underpinning it, and delve into key estimation techniques and their optimality properties.

7.2 Simple Linear Regression Model

Consider the simplest scenario where we model the relationship between a dependent variable Y_i and an independent variable X_i using a linear relationship:

$$Y_i = aX_i + b + \varepsilon_i \tag{7.1}$$

for $i = 1, \dots, n$, where:

- a and b are unknown parameters representing the slope and intercept, respectively.
- ε_i is the error term, assumed to be centered, i.e., $E(\varepsilon_i) = 0$, and having constant variance $\mathrm{Var}(\varepsilon_i) = \sigma^2$.

We further assume that the error terms are normally distributed, $\varepsilon_i \sim N(0, \sigma^2)$, with σ known.

7.2.1 Statistical Model

The statistical model can be formally defined as:

$$(\mathbb{R}, \mathcal{B}(\mathbb{R}), \left(\bigotimes_{i=1}^{n} N(aX_i + b, \sigma^2)\right)_{(a,b)\in\mathbb{R}^2})$$

Here:

- $\mathbb{R}$ denotes the real line.
- $\mathcal{B}(\mathbb{R})$ is the Borel sigma-algebra on $\mathbb{R}$.
- $\bigotimes_{i=1}^{n} N(aX_i + b, \sigma^2)$ represents the product measure of the normal distributions for each observation.

7.2.2 Likelihood Function

Within this statistical model, the likelihood function is given by:

$$L((a,b) \mid y) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - aX_i - b)^2}{2\sigma^2}\right)$$

Simplifying, we obtain:

$$L((a,b) \mid y) = (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - aX_i - b)^2\right)$$

7.2.3 Maximum Likelihood Estimation (MLE)

The maximum likelihood estimators (MLE) for the parameters a and b are obtained by maximizing the likelihood function $L((a,b) \mid y)$. Equivalently, since the logarithm is a monotonically increasing function, we can maximize the log-likelihood:

$$\log L((a,b) \mid y) = -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - aX_i - b)^2$$

Maximizing the log-likelihood is equivalent to minimizing the sum of squared residuals:

$$(\hat{a}, \hat{b}) = \arg \min_{(a,b)\in\mathbb{R}^2} \sum_{i=1}^{n} (y_i - aX_i - b)^2$$

Provided that $X_i \neq X_j$ for $i \neq j$, the least squares problem has a unique solution, historically attributed to Gauss (1801), given by:

$$(\hat{a}, \hat{b}) = \left(\frac{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}, \bar{Y} - \hat{a}\bar{X} \right)$$

where $\bar{X}$ and $\bar{Y}$ denote the sample means of X and Y, respectively.

7.3 General Linear Models

To generalize the simple linear regression model, we introduce the framework of linear models in multiple dimensions.

Definition 18 (Linear Model). A random vector $Y = (Y_1, \dots, Y_n)^T \in \mathbb{R}^n$ stems from a linear model if there exists a parameter vector $\beta \in \mathbb{R}^p$, a matrix $X \in \mathbb{R}^{n \times p}$, and a random vector $\varepsilon \in \mathbb{R}^n$ such that

$$Y = X\beta + \varepsilon$$

1. **Regular Linear Model**: A linear model is called *regular* if
   a) The number of parameters does not exceed the sample size, i.e., $p \leq n$.
   b) The design matrix X has full rank, $\text{rank}(X) = p \leq n$, ensuring a unique solution.
   c) The error vector satisfies $E(\varepsilon) = 0$, meaning the noise is centered.
   d) The covariance matrix of the errors is positive definite, $\Sigma = \text{Cov}(\varepsilon_i, \varepsilon_j)_{i,j \in [n]}$.
2. **Ordinary Linear Model**: A linear model is called *ordinary* if $\Sigma = \sigma^2 I_n$, where I_n is the $n \times n$ identity matrix. Typically, the noise is assumed to be Gaussian in this case.

Remark 11.

1. **Terminology:** Various synonyms are used in the literature:
   - Y: Dependent variable, response, regressand.
   - X: Independent variable, predictor, design matrix, regressor.
   - ε: Error, perturbation, regression function.
2. **Covariance Matrix Σ:** The matrix Σ is symmetric and diagonalizable, i.e., $\Sigma = UDU^T$ for some orthogonal matrix U and diagonal matrix $D = \text{diag}(\lambda_1, \dots, \lambda_n) \in \mathbb{R}^{n \times n}$.
3. **Positive Semi-definiteness:** The covariance matrix Σ is positive semi-definite, meaning $\lambda_i \geq 0$ for all i, which can be shown as:

$$\langle \Sigma u, u \rangle = \langle E[(\varepsilon - E[\varepsilon])(\varepsilon - E[\varepsilon])^T]u, u \rangle = E[(\varepsilon - E[\varepsilon])^2] \geq 0, \quad \forall u \in \mathbb{R}^n$$

4. **Positive Definiteness and Inverses:** If Σ is positive definite (i.e., $\lambda_i > 0$ for all i), then its inverse Σ^{-1} and square root $\Sigma^{-1/2}$ exist and can be expressed as:

$$\Sigma^{-1} = UD^{-1}U^T \quad \text{and} \quad \Sigma^{-1/2} = UD^{-1/2}U^T$$

5. **Random Design:** If the matrix X is not deterministic but random, the model is referred to as having a *random design*.

7.4 Least Squares Estimation

In the context of a regular linear model, the least squares estimator (LSE) seeks to minimize the weighted sum of squared residuals. Specifically, the LSE $\hat{\beta}$ satisfies:

$$\|\sigma^{-1/2}(Y - X\hat{\beta})\|^2 = \inf_{\beta \in \mathbb{R}^p} \|\sigma^{-1/2}(Y - X\beta)\|^2 = \inf_{\beta \in \mathbb{R}^p} \|\Sigma^{-1/2}Y - X_\Sigma\beta\|^2$$

where $X_\Sigma = \Sigma^{-1/2}X$.

7.4.1 Geometric Interpretation

The estimator $X_\Sigma\hat{\beta}$ represents the point within the subspace:

$$U = \{X_\Sigma\beta \mid \beta \in \mathbb{R}^p\} \subseteq \mathbb{R}^n$$

that is closest to the vector $\Sigma^{-1/2}Y$ in terms of Euclidean distance. Formally, this can be expressed using the orthogonal projection Π_U onto U:

$$X_\Sigma\hat{\beta} = \Pi_U(\Sigma^{-1/2}Y)$$

The orthogonal projection satisfies:

- $\Pi_U u = u$ for all $u \in U$.
- $\langle \Pi_U v - v, u \rangle = 0$ for all $u \in U$ and $v \in \mathbb{R}^n$.

Provided that $(X_\Sigma^T X_\Sigma)^{-1}$ exists, we can confirm by direct computation that the projection operator Π_U is given by:

$$\Pi_U = X_\Sigma(X_\Sigma^T X_\Sigma)^{-1}X_\Sigma^T$$

For any $u = X_\Sigma\beta$, we have:

$$X_\Sigma(X_\Sigma^T X_\Sigma)^{-1}X_\Sigma^T X_\Sigma\beta = X_\Sigma\beta = u$$

By symmetry, for any $u \in U$ and $v \in \mathbb{R}^n$:

$$\langle \Pi_U v - v, u \rangle = \langle v, \Pi_U u \rangle - \langle v, u \rangle = \langle v, u \rangle - \langle v, u \rangle = 0$$

7.5 Representation for the Least Squares Estimator

Lemma 1 (Representation for the LSE). *Consider a regular linear model. Then the least squares estimator (LSE) exists uniquely and is given by:*

$$\hat{\beta} = (X_\Sigma^T X_\Sigma)^{-1} X_\Sigma^T \Sigma^{-1/2} Y = X_\Sigma^+ \Sigma^{-1/2} Y$$

where X_Σ^+ denotes the Moore-Penrose pseudoinverse of X_Σ.

Proof. Since X has full rank and $p \leq n$, the matrix $X_\Sigma^T X_\Sigma$ is invertible. Suppose $v \in \ker(X_\Sigma^T X_\Sigma)$, then:

$$0 = v^T X_\Sigma^T X_\Sigma v = (X_\Sigma v)^T (X_\Sigma v) = \|X_\Sigma v\|^2 = \|\Sigma^{-1/2} X v\|^2$$

This implies $\|Xv\|^2 = 0$, hence $Xv = 0$. Given that X has full rank, the only solution is $v = 0$. Therefore, $X_\Sigma^T X_\Sigma$ is invertible.

The projection of $\Sigma^{-1/2} Y$ onto U is:

$$X_\Sigma \hat{\beta} = \Pi_U(\Sigma^{-1/2} Y) = X_\Sigma (X_\Sigma^T X_\Sigma)^{-1} X_\Sigma^T \Sigma^{-1/2} Y$$

Multiplying both sides by X_Σ^T:

$$X_\Sigma^T X_\Sigma \hat{\beta} = X_\Sigma^T X_\Sigma (X_\Sigma^T X_\Sigma)^{-1} X_\Sigma^T \Sigma^{-1/2} Y = X_\Sigma^T \Sigma^{-1/2} Y$$

Hence, solving for $\hat{\beta}$:

$$\hat{\beta} = (X_\Sigma^T X_\Sigma)^{-1} X_\Sigma^T \Sigma^{-1/2} Y$$

□

Remark 12. 1. If $p > n$, the matrix $X_\Sigma^T X_\Sigma$ is not invertible, and the LSE is not unique. In this case, the set of solutions is a $(p - n)$-dimensional subspace, and each solution interpolates the data perfectly, i.e.,

$$\{\beta \in \mathbb{R}^p \mid \|\Sigma^{-1/2} Y - X_\Sigma \beta\|^2 = 0\}$$

7.6 Optimality of the Least Squares Estimator

The least squares estimator possesses several optimality properties under the ordinary linear model. This is formalized in the Gauss-Markov Theorem.

Theorem 2 (Gauss-Markov Theorem). *Consider an ordinary linear model with $\sigma > 0$. Then:*

1. *The least squares estimator $\hat{\beta} = (X^T X)^{-1} X^T Y$ is a linear and unbiased estimator for the parameter β.*
2. *For any desired linear combination of parameters $\alpha = \langle \beta, v \rangle$ where $v \in \mathbb{R}^p$, the estimator $\hat{\alpha} = \langle \hat{\beta}, v \rangle$ is the best linear unbiased estimator (BLUE) of α. This means that $\hat{\alpha}$ has the smallest variance among all linear unbiased estimators of α.*
3. *The estimator $\hat{\sigma}^2 = \frac{\|Y - X\hat{\beta}\|^2}{n-p}$ is an unbiased estimator of σ^2.*

Proof.

1. **Linearity and Unbiasedness of $\hat{\beta}$:**

   The estimator $\hat{\beta}$ is linear because it can be expressed as a linear transformation of Y:
   $$\hat{\beta} = (X^T X)^{-1} X^T Y$$
   For any two vectors y and $\tilde{y}$ in $\mathbb{R}^n$,
   $$\hat{\beta}(y + \tilde{y}) = (X^T X)^{-1} X^T (y + \tilde{y}) = \hat{\beta}(y) + \hat{\beta}(\tilde{y})$$
   demonstrating linearity.

   To show unbiasedness, compute the expectation of $\hat{\beta}$:
   $$\begin{aligned} E[\hat{\beta}] &= (X^T X)^{-1} X^T E[Y] \\ &= (X^T X)^{-1} X^T E[X\beta + \varepsilon] \\ &= (X^T X)^{-1} X^T (X\beta + E[\varepsilon]) \\ &= (X^T X)^{-1} X^T X\beta \quad \text{(since } E[\varepsilon] = 0) \\ &= \beta \end{aligned}$$
   Hence, $\hat{\beta}$ is unbiased.

2. **Optimality of $\hat{\alpha}$ as BLUE:**

   Consider a linear unbiased estimator $\tilde{\alpha}$ for $\alpha = \langle \beta, v \rangle$. Since $\tilde{\alpha}$ is linear, there exists a vector $w \in \mathbb{R}^n$ such that:
   $$\tilde{\alpha} = \langle Y, w \rangle$$
   For $\tilde{\alpha}$ to be unbiased, we require:
   $$E[\tilde{\alpha}] = \langle E[Y], w \rangle = \langle X\beta, w \rangle = \langle \beta, X^T w \rangle = \alpha = \langle \beta, v \rangle$$

This implies that:

$$v = X^T w$$

The variance of $\tilde{\alpha}$ is:

$$\text{Var}(\tilde{\alpha}) = \text{Var}(\langle Y, w\rangle) = \text{Var}(\langle X\beta + \varepsilon, w\rangle) = \text{Var}(\langle \varepsilon, w\rangle) = \sigma^2 \|w\|^2$$

Now, consider the variance of $\hat{\alpha} = \langle \hat{\beta}, v\rangle$:

$$\begin{aligned}
\text{Var}(\hat{\alpha}) &= \text{Var}(\langle \hat{\beta} - \beta, v\rangle) \\
&= \text{Var}\left(\langle (X^T X)^{-1} X^T \varepsilon, v\rangle\right) \\
&= \text{Var}\left(\langle \varepsilon, X(X^T X)^{-1} v\rangle\right) \\
&= \sigma^2 \|X(X^T X)^{-1} v\|^2 \\
&= \sigma^2 \|X(X^T X)^{-1} X^T w\|^2 \quad \text{(since } v = X^T w\text{)} \\
&= \sigma^2 \|\Pi_U w\|^2 \\
&\leq \sigma^2 \|w\|^2 \quad \text{(since projection does not increase the norm)}
\end{aligned}$$

Therefore, $\text{Var}(\hat{\alpha}) \leq \text{Var}(\tilde{\alpha})$, showing that $\hat{\alpha}$ has the smallest variance among all linear unbiased estimators of α.

3. **Unbiasedness of $\hat{\sigma}^2$:**

The estimator $\hat{\sigma}^2$ is given by:

$$\hat{\sigma}^2 = \frac{\|Y - X\hat{\beta}\|^2}{n - p}$$

To show that $\hat{\sigma}^2$ is unbiased, compute its expectation:

$$\begin{aligned}
E[\hat{\sigma}^2] &= \frac{E[\|Y - X\hat{\beta}\|^2]}{n - p} \\
&= \frac{E[\|\varepsilon\|^2 - \|X\hat{\beta}\|^2 + 2\langle \varepsilon, X\hat{\beta}\rangle]}{n - p} \\
&= \frac{E[\|\varepsilon\|^2]}{n - p} \quad \text{(since } E[\langle \varepsilon, X\hat{\beta}\rangle] = 0\text{)} \\
&= \frac{n\sigma^2}{n - p} \quad \text{(since } \|\varepsilon\|^2 \text{ is chi-squared with } n \text{ degrees of freedom)} \\
&= \sigma^2 \quad \text{(since } E[\|\varepsilon\|^2] = n\sigma^2\text{)}
\end{aligned}$$

Hence, $\hat{\sigma}^2$ is an unbiased estimator of σ^2. □

Remark 13. 1. The Gauss-Markov theorem establishes that under the ordinary linear model assumptions, the least squares estimator is the best (in the sense of having the smallest variance) among all linear unbiased estimators. This optimality holds without requiring the error terms to be normally distributed.

7.7 Conclusion

In this lecture, we have introduced the fundamental concepts of linear regression models, both in their simplest form and in a more general framework. We discussed the estimation of model parameters using the method of least squares, explored the geometric interpretation of the estimator, and established its optimality through the Gauss-Markov theorem. Understanding these foundational elements is crucial for further studies in statistical modeling and inference.

8 Lecture 9: Ridge Regression and Its Properties

In this lecture, we delve into Ridge Regression, a technique used to address multicollinearity in linear models. We will explore its derivation, properties, and how it compares to Ordinary Least Squares (OLS). Additionally, we will discuss the bias-variance trade-off inherent in Ridge Regression and present propositions regarding its Mean Squared Error (MSE).

Proof of Ridge Regression Estimator

We begin by deriving the Ridge Regression estimator through optimization of the penalized loss function.

Objective: Minimize the cost function

$$\mathcal{J}(\boldsymbol{\beta}) = \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 + \frac{\sigma^2}{\tau^2}\|\boldsymbol{\beta}\|^2$$

with respect to the coefficient vector $\boldsymbol{\beta}$.

Steps:

1. **Compute the Gradient:**

To find the minimum, we take the gradient of $\mathcal{J}(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$:

$$\nabla_{\boldsymbol{\beta}}\mathcal{J}(\boldsymbol{\beta}) = 2\boldsymbol{X}^\top(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) + \frac{2\sigma^2}{\tau^2}\boldsymbol{\beta}$$

2. **Set Gradient to Zero:**

Setting the gradient equal to zero to find the critical point:

$$2\boldsymbol{X}^\top(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) + \frac{2\sigma^2}{\tau^2}\boldsymbol{\beta} = 0$$

Simplifying:

$$(\boldsymbol{X}^\top\boldsymbol{X} + \frac{\sigma^2}{\tau^2}\boldsymbol{I})\boldsymbol{\beta} = \boldsymbol{X}^\top\boldsymbol{Y}$$

3. **Solve for $\boldsymbol{\beta}$:**

Finally, we solve for $\boldsymbol{\beta}$:

$$\boldsymbol{\beta} = (\boldsymbol{X}^\top \boldsymbol{X} + \frac{\sigma^2}{\tau^2}\boldsymbol{I})^{-1}\boldsymbol{X}^\top \boldsymbol{Y}$$

This is the Ridge Regression estimator, which introduces a penalty term $\lambda = \frac{\sigma^2}{\tau^2}$ to shrink the coefficients.

Posterior Mean

In a Bayesian framework, the posterior mean of $\boldsymbol{\beta}$ given the data can be expressed as:

$$\mu_{\text{post}} = \boldsymbol{\Sigma}_{\text{post}}^{-1}(\boldsymbol{X}^\top \boldsymbol{Y} + \boldsymbol{M}_0^{-1}\boldsymbol{\mu}_0)$$

Upon substituting $\boldsymbol{M}_0^{-1} = \frac{\sigma^2}{\tau^2}\boldsymbol{I}_p$, we obtain:

$$\mu_{\text{post}} = (\sigma^{-2}\boldsymbol{X}^\top \boldsymbol{X} + \tau^{-2}\boldsymbol{I}_p)^{-1}\sigma^{-2}\boldsymbol{X}^\top \boldsymbol{Y}$$

Simplifying further:

$$\mu_{\text{post}} = (\boldsymbol{X}^\top \boldsymbol{X} + \frac{\sigma^2}{\tau^2}\boldsymbol{I})^{-1}\boldsymbol{X}^\top \boldsymbol{Y}$$

This aligns with the Ridge Regression estimator derived earlier.

Remark on Ridge Regression's Existence

An important property of Ridge Regression is that the estimator $\boldsymbol{\beta}$ is always defined, even when the design matrix $\boldsymbol{X}$ does not have full rank. Specifically, Ridge Regression provides a unique solution even in cases where $n < p$ (i.e., when there are more predictors than observations), addressing issues of multicollinearity and overfitting inherent in OLS.

Definition of Ridge Regression Estimator

$$\hat{\boldsymbol{\beta}}_{\text{ridge}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmin}} \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|^2,$$

is termed the **Ridge Regression** estimator. Here, $\lambda > 0$ serves as the regularization parameter that controls the strength of the penalty on the size of the coefficients. A key feature of Ridge Regression is that $\hat{\boldsymbol{\beta}}_{\text{ridge}}$ is always uniquely defined, making it a robust alternative to OLS, especially in high-dimensional settings.

When the model is specified as $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where $\boldsymbol{\varepsilon} \sim N(0, \sigma^2 \boldsymbol{I}_n)$, the Ridge estimator takes the form:

$$\hat{\boldsymbol{\beta}}_{\text{ridge}} = (\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I}_p)^{-1} \boldsymbol{X}^\top \boldsymbol{Y}.$$

Notably, this estimator is independent of σ^2, highlighting its focus on minimizing the penalized residual sum of squares without direct dependence on the noise variance.

Proposition: Mean Squared Error (MSE) of $\hat{\beta}_{\text{ridge}}$

We now analyze the MSE of the Ridge estimator under specific conditions.

Assumptions:

- The linear model is given by $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$.
- The error term $\boldsymbol{\varepsilon} \sim N(0, \sigma^2 \boldsymbol{I}_n)$, with $\sigma^2 > 0$ known.
- The design matrix satisfies $\boldsymbol{X}^\top \boldsymbol{X} = n\boldsymbol{I}_p$, indicating an orthonormal design.

Definition: Let $\mathcal{J} := \langle \boldsymbol{\beta}, \boldsymbol{v} \rangle$ for some $\boldsymbol{v} \in \mathbb{R}^p$, and define:

$$\delta_{\text{ridge}} = \langle \hat{\boldsymbol{\beta}}_{\text{ridge}}, \boldsymbol{v} \rangle.$$

Statements:

1. The expected squared error of δ_{ridge} is:

$$\mathbb{E}_{\boldsymbol{\beta}}[(\delta_{\text{ridge}} - \mathcal{J})^2] = (1+\lambda)^{-2} \langle \boldsymbol{\beta}_v, \boldsymbol{v} \rangle^2 + \frac{\sigma^2}{n} \|\boldsymbol{v}\|^2 (1+\lambda)^{-2}.$$

2. The expected squared error of the Ridge estimator vector is:

$$\mathbb{E}_{\boldsymbol{\beta}}[\|\hat{\boldsymbol{\beta}}_{\text{ridge}} - \boldsymbol{\beta}\|^2] = (1+\lambda)^{-2} \|\boldsymbol{\beta}\|^2 + \frac{p\sigma^2}{n} \frac{1}{(1+\lambda)^2}.$$

Derivation:

Starting from the Ridge estimator:

$$\hat{\boldsymbol{\beta}}_{\text{ridge}} = (\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I}_p)^{-1} \boldsymbol{X}^\top \boldsymbol{Y}.$$

Given the orthonormal design $\boldsymbol{X}^\top \boldsymbol{X} = n\boldsymbol{I}_p$, this simplifies to:

$$\hat{\boldsymbol{\beta}}_{\text{ridge}} = (n\boldsymbol{I}_p + \lambda \boldsymbol{I}_p)^{-1} \boldsymbol{X}^\top \boldsymbol{Y} = \frac{1}{1+\frac{\lambda}{n}} (\boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{X}^\top \boldsymbol{\varepsilon}).$$

Simplifying further:

$$\hat{\boldsymbol{\beta}}_{\text{ridge}} = \frac{1}{1+\frac{\lambda}{n}}\boldsymbol{\beta} + \frac{1}{1+\frac{\lambda}{n}}\boldsymbol{X}^\top\boldsymbol{\varepsilon}.$$

This expression separates the estimator into a bias term and a variance term, which will be instrumental in understanding the bias-variance decomposition.

Bias-Variance Decomposition

The Mean Squared Error (MSE) of an estimator can be decomposed into the sum of the squared bias and the variance of the estimator:

$$\mathbb{E}\left[(\hat{\boldsymbol{\beta}}_{\text{ridge}} - \mathcal{J})^2\right] = (\mathbb{E}[\hat{\boldsymbol{\beta}}_{\text{ridge}}] - \mathcal{J})^2 + \text{Var}(\hat{\boldsymbol{\beta}}_{\text{ridge}}).$$

Substituting from our Ridge estimator:

$$= \left((1+\frac{\lambda}{n})^{-1}\langle\boldsymbol{\beta}, \boldsymbol{v}\rangle\right)^2 + \frac{\lambda^2}{(1+\lambda)^2}\text{Var}(\boldsymbol{X}^\top\boldsymbol{\varepsilon}, \nu).$$

Observations:

$$(1+\frac{\lambda}{n})^{-1} = \frac{1}{(1+\frac{\lambda}{n})}.$$

Furthermore, the variance term can be expressed as:

$$\text{Var}(\boldsymbol{X}^\top\boldsymbol{\varepsilon}, \boldsymbol{\nu}) = \boldsymbol{\nu}^\top\boldsymbol{X}\text{Cov}(\boldsymbol{\varepsilon})\boldsymbol{X}^\top\boldsymbol{\nu} = \sigma^2\|\boldsymbol{\nu}\|^2.$$

Thus, the bias-variance decomposition clearly illustrates the trade-off controlled by the regularization parameter λ.

Corollary: MSE Under Orthonormal Design

Under the same assumptions, specifically the orthonormal design $\boldsymbol{X}^\top\boldsymbol{X} = n\boldsymbol{I}_p$, the MSE of the Ridge estimator vector is:

$$\mathbb{E}[\|\hat{\boldsymbol{\beta}}_{\text{ridge}} - \boldsymbol{\beta}\|^2] = \frac{1}{(1+\frac{\lambda}{n})^2}\|\boldsymbol{\beta}\|^2 + \frac{p\sigma^2}{n(1+\frac{\lambda}{n})^2}.$$

This expression highlights how both the bias and variance components of the estimator's error are influenced by λ.

Remark on Ridge Regression Behavior

An insightful observation about Ridge Regression is its behavior in relation to the norm of $\boldsymbol{\beta}$:

$$\text{For small } \|\boldsymbol{\beta}\|, \text{ Ridge Regression converges to OLS.}$$

This implies that the optimal choice of λ depends on the true underlying parameter $\|\boldsymbol{\beta}\|$. When $\|\boldsymbol{\beta}\|$ is small, a smaller λ is preferred to minimize bias, whereas larger λ values are beneficial when $\|\boldsymbol{\beta}\|$ is large to control variance.

Confidence Sets & Tests in Linear Model

While Ridge and OLS estimators provide point estimates for $\boldsymbol{\beta}$, it's crucial to quantify the uncertainty associated with these estimates. This involves constructing confidence sets and performing hypothesis tests, which inherently depend on the variance σ^2.

Assumption: Throughout this discussion, we assume $\boldsymbol{\varepsilon} \sim N(0, \sigma^2 \boldsymbol{I}_n)$, ensuring the errors are normally distributed with known variance.

Easy Case: Known σ^2

When the variance σ^2 is known, the distribution of the OLS estimator is straightforward:

$$\hat{\boldsymbol{\beta}}_{\text{OLS}} \sim N(\boldsymbol{\beta}, \sigma^2 (\boldsymbol{X}^\top \boldsymbol{X})^{-1}).$$

Given an orthonormal design ($\boldsymbol{X}^\top \boldsymbol{X} = n\boldsymbol{I}_p$), this simplifies to:

$$\hat{\boldsymbol{\beta}}_{\text{OLS}} \sim N\left(\boldsymbol{\beta}, \frac{\sigma^2}{n} \boldsymbol{I}_p\right).$$

For any linear combination $\mathcal{J} = \langle \boldsymbol{\beta}, \boldsymbol{\nu} \rangle$, the estimator is:

$$\hat{\mathcal{J}} = \langle \hat{\boldsymbol{\beta}}_{\text{OLS}}, \boldsymbol{\nu} \rangle \sim N\left(\mathcal{J}, \sigma^2 \boldsymbol{\nu}^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{\nu}\right).$$

This allows us to construct a 95% confidence interval for $\mathcal{J}$:

$$I_{95\%}(\mathcal{J}) = \left[\hat{\mathcal{J}} \pm 1.96 \sqrt{\boldsymbol{\nu}^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{\nu}}\right].$$

Notes on *t*- and *F*-distributions

In practice, the variance σ^2 is often unknown and must be estimated from the data. This necessitates the use of the t-distribution and the F-distribution for constructing confidence intervals and conducting hypothesis tests.

Definitions

t-distribution: The t-distribution with $n \geq 1$ degrees of freedom on $\mathbb{R}$ has the probability density function:

$$f_n(x) = C_n \left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}},$$

where C_n is the normalizing constant ensuring that the total probability integrates to 1.

Special Case: For $n = 1$, the t-distribution simplifies to:

$$f_1(x) = C_1 \frac{1}{1 + x^2},$$

which corresponds to the **Cauchy distribution**, a distribution with heavy tails.

F-distribution: The F-distribution with $(m, n) \in \mathbb{N}^2$ degrees of freedom has the density:

$$f_{m,n}(x) = C_{m,n} \frac{x^{\frac{m}{2}-1}}{(mx+n)^{\frac{m+n}{2}}}, \quad x \in (0, \infty),$$

where $C_{m,n}$ is the normalizing constant. The F-distribution is pivotal in variance ratio tests and regression analysis.

Utility of *t*- and *F*-distributions

These distributions are instrumental in constructing confidence intervals and performing hypothesis tests when dealing with normally distributed errors, especially when the variance is unknown and must be estimated from the data.

Lemma 2. *Let $X_1, \ldots, X_n, Y_1, \ldots, Y_n$ be independent and identically distributed (i.i.d.) $N(0, \Delta)$ random variables. Then:*

1.

$$T_n := \frac{X_n}{\sqrt{\frac{1}{n}\sum_{i=1}^{n} Y_i^2}} \sim t_n.$$

2.

$$F_{m,n} := \frac{\frac{1}{m}\sum_{i=1}^{m} X_i^2}{\frac{1}{n}\sum_{j=1}^{n} Y_j^2} \sim F_{m,n}.$$

Remarks on the Lemma

1. The t-distribution emerges when evaluating the ratio of a normally distributed variable to the square root of a scaled chi-squared variable, reflecting the relationship between the sample mean and sample variance.
2. As the degrees of freedom n increase, the t-distribution T_n converges in distribution to the standard normal distribution $N(0, 1)$, illustrating the asymptotic behavior of the estimator.

Proof of the Lemma

Proof. 1. To show that T_n follows a t-distribution with n degrees of freedom, observe that:

$$T_n^2 = F_{1,n}.$$

Considering the symmetry of the t-distribution around 0, we utilize a change of variables:

$$f_{F_{m,n}}(x) = f_{F_{m,n}}(x^2) \cdot 2x, \quad x > 0.$$

Given that T_n is symmetric, we derive:

$$f_{T_n}(x) = f_{F_{m,n}}(x^2)|x| = C_n \left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}}.$$

This aligns with the known density of the t-distribution.

2. To establish the distribution of $F_{m,n}$, define:

$$X = \sum_{i=1}^{m} X_i^2, \quad Y = \sum_{j=1}^{n} Y_j^2.$$

Since each X_i and Y_j is normally distributed, it follows that:

$$X \sim \chi_m^2, \quad Y \sim \chi_n^2,$$

where χ_m^2 and χ_n^2 denote chi-squared distributions with m and n degrees of freedom, respectively. The density of χ_m^2 is:

$$f(x) \propto x^{m/2-1} e^{-x/2}, \quad x > 0.$$ □

Derivation of the F-distribution

To derive the distribution of $F_{m,n}$, consider the ratio:

$$W = \frac{X}{Y}.$$

The cumulative distribution function (CDF) of W is:

$$\mathbb{P}\left(\frac{X}{Y} < z\right) = \int_0^\infty \int_0^{zy} 1 \cdot f_X(x) f_Y(y)\, dx\, dy.$$

By substituting $x = wy$, we obtain:

$$= \int_0^\infty \int_0^{z} f_X(wy) f_Y(y) \cdot y\, dw\, dy.$$

This simplifies to:

$$= \int_0^\infty f_X(zy) f_Y(y) \cdot y\, dy.$$

Substituting the chi-squared densities:

$$\propto \int_0^\infty (zy)^{\frac{m}{2}-1} y^{\frac{n}{2}-1} e^{-(z+y)/2}\, dy.$$

Introducing the change of variable $a = \frac{z}{z+1} y$, the integral becomes:

$$\propto z^{\frac{m}{2}-1} (z+1)^{-\frac{m+n}{2}} \int_0^\infty a^{\frac{m}{2}-1} e^{-a}\, da.$$

Recognizing that the integral is the gamma function $\Gamma\left(\frac{m}{2}\right)$, we conclude:

$$f_F(z) \propto z^{\frac{m}{2}-1} (z+1)^{-\frac{m+n}{2}}.$$

Finally, by appropriate scaling, we arrive at the density function of the F-distribution:

$$f_{m,n}(z) = \frac{m^{m/2} n^{n/2}}{B\left(\frac{m}{2}, \frac{n}{2}\right)} \frac{z^{m/2-1}}{(mz+n)^{(m+n)/2}},$$

where $B(\cdot, \cdot)$ is the beta function.

9 Lecture 10: Confidence Sets and Hypothesis Testing in Linear Models

Building upon the foundations laid in the previous lecture, we now focus on constructing confidence sets and performing hypothesis tests within the context of linear models. We will explore both t-tests and F-tests, essential tools for statistical inference in regression analysis.

Ridge and OLS Estimators

Recall that in the linear model $Y = X\beta + \varepsilon$, with $\varepsilon \sim N(0, \sigma^2 I_n)$, the OLS estimator is given by:

$$\hat{\beta} = (X^T X)^{-1} X^T Y \sim N(\beta, \sigma^2 (X^T X)^{-1}).$$

When σ^2 is unknown, it is typically estimated by:

$$\hat{\sigma}^2 = \frac{||Y - X\hat{\beta}_{OLS}||^2}{n-p} \sim \sigma^2 \frac{\chi^2(n-p)}{n-p}.$$

This estimator is unbiased and follows a scaled chi-squared distribution.

Distributions of Estimators

$$t_n = \frac{N(0,1)}{\sqrt{\chi^2(n)/n}} \quad \text{and} \quad F_{m,n} = \frac{\chi^2(m)/m}{\chi^2(n)/n}$$

These ratios form the basis of the t- and F-distributions, respectively.

Lemma 3. *Let $\xi \sim N(0, I_n)$, a random variable in $\mathbb{R}^n$, and let $R \in \mathbb{R}^{n \times n}$ be an orthogonal projection matrix ($R = R^2$, $R = R^T$), with rank$(R) = r \leq n$. Then:*

1. *$\xi^T R \xi = ||R\xi||^2 \sim \chi^2(r)$.*
2. *If $B \in \mathbb{R}^{p \times n}$ is such that $BR = 0$, then $B\xi$ is independent of $R\xi$.*
3. *If $S \in \mathbb{R}^{n \times n}$ is another orthogonal projection with rank$(S) = s \leq n$ and $RS = 0$,*

then

$$\frac{s}{r}\frac{\xi^T R\xi}{\xi^T S\xi} \sim F(r,s).$$

Proof. 1. Since R is an orthogonal projection, there exists an orthogonal matrix T such that:

$$R = T\begin{pmatrix} I_r & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} T^T = TD_rT^T.$$

Applying the orthogonal transformation T^T to ξ, we maintain the distribution:

$$T^T\xi \sim N(0, I_n).$$

Therefore,

$$\xi^T R\xi = \xi^T(TD_rT^T)\xi = (T^T\xi)^T D_r(T^T\xi) = \sum_{i=1}^{r}(T^T\xi)_i^2 \sim \chi^2(r).$$

2. Let $A_1 = B\xi$ and $A_2 = R\xi$. The covariance between A_1 and A_2 is:

$$\mathrm{Cov}(A_1, A_2) = \mathrm{Cov}(B\xi, R\xi) = B\mathrm{Cov}(\xi,\xi)R^T = BR^T = BR = 0,$$

since $BR = 0$ by assumption. Given that both A_1 and A_2 are linear combinations of jointly normal variables, zero covariance implies independence.

3. Given that $RS = 0$, by part (2), $S\xi$ and $R\xi$ are independent. From part (1), we have $S\xi \sim \chi^2(s)$ and $R\xi \sim \chi^2(r)$. Thus, the ratio:

$$\frac{s}{r}\frac{\xi^T R\xi}{\xi^T S\xi} \sim F(r,s)$$

follows directly from the definition of the F-distribution as the ratio of two scaled chi-squared variables. □

Theorem: Confidence Sets in Linear Models with Unknown σ^2

Theorem 3 (Linear Model Confidence Sets - Unknown σ^2). *Assume the regular linear model $Y = X\beta + \varepsilon$, with rank$(X) = p \le n$ and $\varepsilon \sim N(0, \sigma^2 I_n)$. Let $\alpha \in (0,1)$. Then:*

1. *Let $q_{F_{p,n-p},1-\alpha}$ be the $1-\alpha$ quantile of the $F_{p,n-p}$ distribution. Then the set*

$$C(Y,X) = \left\{\beta \in \mathbb{R}^p \mid \frac{||X(\beta - \hat{\beta}_{OLS})||^2}{p\hat{\sigma}^2} \le q_{F_{p,n-p},1-\alpha}\right\}$$

is a $1-\alpha$ confidence set for β.

2. *For a specific linear combination $\alpha = \langle \beta, v\rangle$ for some $v \in \mathbb{R}^p$, a $1-\alpha$ confidence interval is given by:*

$$C = C(Y,X) = \left\{ \alpha \in \mathbb{R} \mid \left| \frac{\alpha - \hat{\alpha}}{\hat{\sigma}\sqrt{v^T(X^TX)^{-1}v}} \right| < q \right\},$$

where $\hat{\alpha} = \langle \hat{\beta}_{OLS}, v\rangle$ and q is the $1-\frac{\alpha}{2}$ quantile of the t_{n-p} distribution.

Proof. 1. Consider the residual sum of squares under the null hypothesis $H_0 : \beta = \hat{\beta}_{OLS}$:

$$RSS = ||Y - X\hat{\beta}_{OLS}||^2.$$

The residual sum of squares under the alternative hypothesis is:

$$RSS_{H_0} = ||Y - X\beta||^2.$$

Given the orthonormal design and the properties of the Ridge estimator, we have:

$$RSS_{H_0} - RSS = (K\hat{\beta}_{OLS} - d)^2(K(X^TX)^{-1}K^T)^{-1},$$

where K is the contrast matrix. Under H_0, this difference follows a chi-squared distribution with r degrees of freedom:

$$\frac{RSS_{H_0} - RSS}{\sigma^2} \sim \chi^2(r).$$

Consequently, the ratio:

$$\frac{RSS_{H_0} - RSS}{r\hat{\sigma}^2} \sim F(r, n-p).$$

This establishes that:

$$\frac{||X(\beta - \hat{\beta}_{OLS})||^2}{p\hat{\sigma}^2} \sim F(p, n-p),$$

leading to the construction of the confidence set $C(Y,X)$ as stated.

2. For the specific linear combination $\alpha = \langle \beta, v\rangle$, the OLS estimator satisfies:

$$\hat{\alpha} = \langle \hat{\beta}_{OLS}, v\rangle \sim N\left(\alpha, \sigma^2 v^T(X^TX)^{-1}v\right).$$

The standardized statistic is:

$$\frac{\alpha - \hat{\alpha}}{\sigma\sqrt{v^T(X^TX)^{-1}v}} \sim t_{n-p}.$$

However, since σ^2 is unknown and estimated by $\hat{\sigma}^2$, the statistic becomes:

$$\frac{\alpha - \hat{\alpha}}{\hat{\sigma}\sqrt{v^T(X^TX)^{-1}v}} \sim t_{n-p}.$$

Thus, the confidence interval C is constructed by:

$$\left|\frac{\alpha - \hat{\alpha}}{\hat{\sigma}\sqrt{v^T(X^TX)^{-1}v}}\right| < q,$$

where q is the critical value from the t_{n-p} distribution corresponding to the desired confidence level. □

9.0.1 The t- and F-Tests

Confidence sets facilitate hypothesis testing. Here, we outline the procedures for conducting t-tests and F-tests in the linear model context.

Remark: Method (t-test)

In a regular linear model with $\varepsilon \sim N(0, \sigma^2 I_n)$, consider the null hypothesis:

$$H_0 : \gamma = \gamma_0 \quad \text{vs} \quad H_1 : \gamma \neq \gamma_0,$$

where $\gamma = \langle \beta, v \rangle$ for some $v \in \mathbb{R}^p$.

The two-sided t-test statistic is defined as:

$$T_{\gamma_0, n-p} = \frac{\gamma_0 - \hat{\gamma}}{\hat{\sigma}\sqrt{v^T(X^TX)^{-1}v}},$$

where $\hat{\gamma} = \langle \hat{\beta}_{OLS}, v \rangle$.

The corresponding test function is:

$$\varphi_{\alpha_0}(Y, X) = \mathbf{1}\left(|T_{\gamma_0, n-p}(Y, X)| > q\right),$$

where q is the critical value from the t_{n-p} distribution corresponding to the significance level α_0.

Remark: Method (F-test)

For the same linear model setting, consider the null hypothesis:

$$H_0 : \beta = \beta_0 \quad \text{vs} \quad H_1 : \beta \neq \beta_0,$$

where $\beta_0 \in \mathbb{R}^p$.

The F-test statistic is defined as:

$$F_{\beta_0,n-p}(Y, X) = \frac{||X(\beta - \hat{\beta}_{OLS})||^2}{p\hat{\sigma}^2},$$

which follows an $F_{p,n-p}$ distribution under H_0.

The corresponding test function is:

$$\varphi_{\beta_0}(Y, X) = \mathbb{1}\left(|F_{\beta_0,n-p}(Y, X)| > q\right),$$

where q is the critical value from the $F_{p,n-p}$ distribution corresponding to the desired significance level.

9.0.2 General Linear Hypothesis Testing Problems

Beyond testing individual parameters, we often encounter more complex hypotheses involving multiple linear constraints on β. This section addresses such scenarios.

Definition 19. A **linear hypothesis testing problem** is of the form:

$$H_0 : K\beta = d \quad \text{vs} \quad H_1 : K\beta \neq d,$$

where $K \in \mathbb{R}^{r\times p}$ is a matrix with $\text{rank}(K) = r \leq p$ and $d \in \mathbb{R}^r$. Essentially, this represents r linear constraints on the parameter vector β. The matrix K is referred to as the **contrast matrix**.

Theorem 4. *Assume the regular linear model $Y = X\beta + \varepsilon$, with rank$(X) = p \leq n$ and $\varepsilon \sim N(0, \sigma^2 I_n)$. Consider the hypothesis:*

$$H_0 : K\beta = d \quad \textit{vs} \quad H_1 : K\beta \neq d.$$

Define the residual sum of squares (RSS) as:

$$RSS = ||Y - X\hat{\beta}_{OLS}||^2 \quad \textit{and} \quad RSS_{H_0} = ||Y - X\hat{\beta}_{H_0}||^2,$$

where $\hat{\beta}_{H_0}$ is the estimator of β under the constraint $K\beta = d$.

Then:

1. *The constrained estimator is:*

$$\hat{\beta}_{H_0} = \hat{\beta}_{OLS} - (X^TX)^{-1}K^T(K(X^TX)^{-1}K^T)^{-1}(K\hat{\beta}_{OLS} - d).$$

2. *The difference in RSS satisfies:*

$$RSS_{H_0} - RSS = (K\hat{\beta}_{OLS} - d)^T(K(X^TX)^{-1}K^T)^{-1}(K\hat{\beta}_{OLS} - d) \sim \sigma^2\chi^2(r).$$

3. *The test statistic:*

$$F = \frac{(RSS_{H_0} - RSS)/r}{RSS/(n-p)} \sim F_{r,n-p}$$

under the null hypothesis H_0.

Proof. 1. The constrained estimator $\hat{\beta}_{H_0}$ minimizes the RSS subject to $K\beta = d$. By applying the method of Lagrange multipliers or directly solving the constrained optimization problem, we obtain:

$$\hat{\beta}_{H_0} = \hat{\beta}_{OLS} - (X^TX)^{-1}K^T(K(X^TX)^{-1}K^T)^{-1}(K\hat{\beta}_{OLS} - d).$$

2. The difference in RSS between the unconstrained and constrained models is:

$$RSS_{H_0} - RSS = (K\hat{\beta}_{OLS} - d)^T(K(X^TX)^{-1}K^T)^{-1}(K\hat{\beta}_{OLS} - d).$$

Under H_0, $K\hat{\beta}_{OLS} \sim N(d, \sigma^2K(X^TX)^{-1}K^T)$. Thus, the quadratic form above follows a chi-squared distribution with r degrees of freedom:

$$\frac{RSS_{H_0} - RSS}{\sigma^2} \sim \chi^2(r).$$

3. Defining the F-statistic as:

$$F = \frac{(RSS_{H_0} - RSS)/r}{RSS/(n-p)} = \frac{RSS_{H_0} - RSS}{r\hat{\sigma}^2} \sim F_{r,n-p},$$

where $\hat{\sigma}^2 = \frac{RSS}{n-p}$ is the unbiased estimator of σ^2. Under H_0, this statistic follows the F-distribution with r and $n-p$ degrees of freedom. □

9.0.3 The t- and F-Tests

Remark: t-Test Method

Consider a regular linear model where $\varepsilon \sim N(0, \sigma^2 I_n)$. To test the null hypothesis $H_0 : \gamma = \gamma_0$ against the alternative $H_1 : \gamma \neq \gamma_0$, where $\gamma = \langle \beta, v \rangle$ for some $v \in \mathbb{R}^p$, we employ the two-sided t-test.

The t-test statistic is defined as:

$$T_{\gamma_0, n-p} = \frac{\gamma_0 - \hat{\gamma}}{\hat{\sigma}\sqrt{v^T (X^T X)^{-1} v}},$$

where $\hat{\gamma} = \langle \hat{\beta}_{OLS}, v \rangle$.

The corresponding test function is:

$$\varphi_{\alpha_0}(Y, X) = \mathbf{1}\left(|T_{\gamma_0, n-p}(Y, X)| > q\right),$$

where q is the critical value from the t_{n-p} distribution corresponding to the significance level α_0.

Remark: F-Test Method

In the same linear model setting, to test the null hypothesis $H_0 : \beta = \beta_0$ against $H_1 : \beta \neq \beta_0$, we use the F-test.

The F-test statistic is defined as:

$$F_{\beta_0, n-p}(Y, X) = \frac{||X(\beta - \hat{\beta}_{OLS})||^2}{p\hat{\sigma}^2},$$

which follows an $F_{p,n-p}$ distribution under H_0.

The corresponding test function is:

$$\varphi_{\beta_0}(Y, X) = \nVdash\left(|F_{\beta_0, n-p}(Y, X)| > q\right),$$

where q is the critical value from the $F_{p,n-p}$ distribution at the desired significance level.

9.0.4 General Linear Hypothesis Testing Problems

Beyond testing individual coefficients or simple contrasts, we often need to test more general linear hypotheses involving multiple parameters. This section formalizes such testing problems and provides the necessary statistical framework.

Definition 20. A **linear hypothesis testing problem** involves testing $H_0 : K\beta = d$ against $H_1 : K\beta \neq d$, where:

- $K \in \mathbb{R}^{r \times p}$ is the contrast matrix with $\text{rank}(K) = r \leq p$.
- $d \in \mathbb{R}^r$ is the vector of constants.

Theorem 5. *Assume the regular linear model $Y = X\beta + \varepsilon$, with rank$(X) = p \leq n$ and $\varepsilon \sim N(0, \sigma^2 I_n)$. Consider the linear hypothesis:*

$$H_0 : K\beta = d \quad vs \quad H_1 : K\beta \neq d.$$

Define the residual sum of squares (RSS) under H_0 as:

$$RSS_{H_0} = ||Y - X\hat{\beta}_{H_0}||^2,$$

where $\hat{\beta}_{H_0}$ is the constrained estimator minimizing RSS subject to $K\beta = d$.

Then:

1. *The constrained estimator is:*

$$\hat{\beta}_{H_0} = \hat{\beta}_{OLS} - (X^T X)^{-1} K^T (K(X^T X)^{-1} K^T)^{-1} (K\hat{\beta}_{OLS} - d).$$

2. *The difference in RSS is given by:*

$$RSS_{H_0} - RSS = (K\hat{\beta}_{OLS} - d)^T (K(X^T X)^{-1} K^T)^{-1} (K\hat{\beta}_{OLS} - d) \sim \sigma^2 \chi^2(r).$$

3. *The F-statistic is defined as:*

$$F = \frac{RSS_{H_0} - RSS}{r\hat{\sigma}^2} \sim F_{r,n-p},$$

under the null hypothesis H_0.

Proof.

1. To derive the constrained estimator $\hat{\beta}_{H_0}$, we minimize the RSS subject to $K\beta = d$. Utilizing the method of Lagrange multipliers or projecting the OLS estimator onto the constraint set, we obtain:

$$\hat{\beta}_{H_0} = \hat{\beta}_{OLS} - (X^T X)^{-1} K^T (K(X^T X)^{-1} K^T)^{-1} (K\hat{\beta}_{OLS} - d).$$

2. The difference in RSS between the constrained and unconstrained models is a quadratic form:

$$RSS_{H_0} - RSS = (K\hat{\beta}_{OLS} - d)^T (K(X^T X)^{-1} K^T)^{-1} (K\hat{\beta}_{OLS} - d).$$

Under H_0, $K\hat{\beta}_{OLS}$ follows a normal distribution centered at d, and the quadratic form follows a scaled chi-squared distribution:

$$\frac{RSS_{H_0} - RSS}{\sigma^2} \sim \chi^2(r).$$

3. Defining the F-statistic as:

$$F = \frac{RSS_{H_0} - RSS}{r\hat{\sigma}^2},$$

and noting that $\hat{\sigma}^2 = \frac{RSS}{n-p}$ is an unbiased estimator of σ^2, we conclude:

$$F \sim F_{r,n-p}.$$

This completes the proof. □

10 Lecture 11

10.1 Introduction

In this lecture, we delve into hypothesis testing within the framework of regular linear models (LM) and explore the Analysis of Variance (ANOVA). We will cover the derivation and proof of key theorems related to the F-test in linear models and the decomposition of the Residual Sum of Squares (RSS) in ANOVA.

10.2 Hypothesis Testing in Linear Models

Consider the standard linear model:

$$Y = X\beta + \varepsilon,$$

where:

- $Y \in \mathbb{R}^n$ is the response vector,
- $X \in \mathbb{R}^{n \times p}$ is the design matrix,
- $\beta \in \mathbb{R}^p$ is the parameter vector,
- $\varepsilon \sim N(0, \sigma^2 I_n)$ is the error term.

We are interested in testing linear hypotheses of the form:

$$H_0 : K\beta = d \quad \text{vs.} \quad H_1 : K\beta \neq d,$$

where $K \in \mathbb{R}^{m \times p}$ is a known matrix and $d \in \mathbb{R}^m$ is a known vector.

10.2.1 Residual Sum of Squares (RSS)

Define the Residual Sum of Squares as:

$$RSS = \|Y - X\beta\|^2,$$

and the RSS under the null hypothesis H_0 as:

$$RSS_{H_0} = \|Y - X\beta_{H_0}\|^2,$$

where β_{H_0} is the Ordinary Least Squares (OLS) estimator constrained by $K\beta = d$. The constrained estimator β_{H_0} is given by:

$$\beta_{H_0} = \hat{\beta} - (X^\top X)^{-1}K^\top \left(K(X^\top X)^{-1}K^\top\right)^{-1} (K\hat{\beta} - d),$$

where $\hat{\beta} = (X^\top X)^{-1}X^\top Y$ is the unconstrained OLS estimator.

10.2.2 The F-Test Theorem

Theorem: Assume a regular linear model with $\varepsilon \sim N(0, \sigma^2 I_n)$. Consider the hypotheses:

$$H_0 : K\beta = d \quad \text{vs.} \quad H_1 : K\beta \neq d.$$

Define:

$$F = \frac{1}{p}\frac{RSS_{H_0} - RSS}{RSS/n}.$$

Under H_0, F follows an F-distribution with m and $n - p$ degrees of freedom:

$$F \sim F_{m,n-p}.$$

Proof:

1. **Relation between RSS_{H_0} and RSS**
We first show that:

$$RSS_{H_0} - RSS = \|X(\beta_{H_0} - \hat{\beta})\|^2 = (K\hat{\beta} - d)^\top \left(K(X^\top X)^{-1}K^\top\right)^{-1} (K\hat{\beta} - d).$$

Since β_{H_0} minimizes RSS subject to $K\beta = d$, it follows that:

$$K\beta_{H_0} = d.$$

Substituting the expression for β_{H_0}:

$$K\beta_{H_0} = K\hat{\beta} - K(X^\top X)^{-1}K^\top \left(K(X^\top X)^{-1}K^\top\right)^{-1} (K\hat{\beta} - d) = d.$$

2. **Distribution under H_0**
Under H_0, we have:

$$RSS_{H_0} \sim \sigma^2\chi^2(n - m).$$

Additionally, the difference $RSS_{H_0} - RSS$ is:

$$RSS_{H_0} - RSS = \|X(\beta_{H_0} - \hat{\beta})\|^2.$$

Let $Z = K\hat{\beta} - d$. Then:

$$\mathbb{E}[Z] = K\mathbb{E}[\hat{\beta}] - d = K\beta - d = 0 \quad \text{under } H_0.$$

The variance of Z is:

$$\text{Var}(Z) = \sigma^2 K(X^\top X)^{-1}K^\top.$$

Therefore:

$$Z \sim \mathcal{N}(0, \sigma^2 K(X^\top X)^{-1}K^\top).$$

Consequently:

$$\frac{RSS_{H_0} - RSS}{\sigma^2} = \frac{Z^\top \left(K(X^\top X)^{-1}K^\top\right)^{-1} Z}{\sigma^2} \sim \chi^2(m).$$

Also, under H_0:

$$RSS \sim \sigma^2\chi^2(n-p).$$

3. **Construction of the F-statistic**
   Combining the results, we define the F-statistic as:

$$F = \frac{\frac{RSS_{H_0} - RSS}{m}}{\frac{RSS}{n-p}}.$$

Under H_0, since $RSS_{H_0} - RSS$ and RSS are independent,

$$F \sim F_{m,n-p}.$$

This concludes the proof of the F-test theorem.

10.3 Analysis of Variance (ANOVA)

ANOVA is a statistical method used to compare means across multiple groups to determine if at least one group mean is different from the others. It decomposes the total variability in the data into components attributable to different sources.

10.3.1 Motivation

Suppose we have data from k different groups. We aim to test whether the means of these groups are equal.

10.3.2 Model Specification

We are given data:

$$Y_{ij} = \mu_i + \varepsilon_{ij}, \quad i = 1, \dots, k, \quad j = 1, \dots, n_i,$$

where:

- Y_{ij} is the observation in the i-th group and j-th replicate,
- μ_i is the mean of the i-th group,
- $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$ are independent error terms.

10.3.3 Factor Model

ANOVA can be viewed as a special case of a linear model known as the *factor model* with one categorical predictor (factor). The index $i = 1, \dots, k$ represents different levels (groups) of the factor.

- $n = \sum_{i=1}^{k} n_i$ is the total sample size.
- The design is **balanced** if all groups have the same number of observations, i.e., $n_1 = n_2 = \ldots = n_k$.

Remark: ANOVA can be expressed in matrix form as:

$$\begin{pmatrix} Y_{1,1} \\ Y_{1,2} \\ \vdots \\ Y_{k,n_k} \end{pmatrix} = \begin{pmatrix} \mathbf{1}_{n_1} & 0 & \cdots & 0 \\ 0 & \mathbf{1}_{n_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \mathbf{1}_{n_k} \end{pmatrix} \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_k \end{pmatrix} + \varepsilon,$$

where $\mathbf{1}_{n_i}$ is a column vector of ones of length n_i.

10.3.4 Hypothesis Testing in ANOVA

We aim to test:

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_k \quad \text{vs.} \quad H_a : \exists i, j \text{ with } \mu_i \neq \mu_j.$$

10.3.5 Decomposition of RSS

To perform ANOVA, we decompose the Total Sum of Squares (SST) into the Sum of Squares Between Groups (SSB) and the Sum of Squares Within Groups (SSW).

Definitions:

$$\bar{Y}_{i.} = \frac{1}{n_i}\sum_{j=1}^{n_i} Y_{ij}, \quad i = 1, \ldots, k \quad \text{(Group Means)},$$
$$\bar{Y}_{..} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i} Y_{ij} \quad \text{(Overall Mean)}.$$

Sum of Squares:

$$SSB = \sum_{i=1}^{k} n_i(\bar{Y}_{i.} - \bar{Y}_{..})^2 \quad \text{(Sum of Squares Between Groups)},$$
$$SSW = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_{i.})^2 \quad \text{(Sum of Squares Within Groups)}.$$

Total Sum of Squares:

$$SST = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_{..})^2.$$

Theorem (Decomposition of RSS):

$$SST = SSB + SSW.$$

Proof:

$$
\begin{aligned}
SST &= \sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij}-\bar{Y}_{..})^2 \\
&= \sum_{i=1}^{k}\sum_{j=1}^{n_i}\Big[(Y_{ij}-\bar{Y}_{i.})+(\bar{Y}_{i.}-\bar{Y}_{..})\Big]^2 \\
&= \sum_{i=1}^{k}\sum_{j=1}^{n_i}\Big[(Y_{ij}-\bar{Y}_{i.})^2+(\bar{Y}_{i.}-\bar{Y}_{..})^2+2(Y_{ij}-\bar{Y}_{i.})(\bar{Y}_{i.}-\bar{Y}_{..})\Big] \\
&= \sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij}-\bar{Y}_{i.})^2+\sum_{i=1}^{k}\sum_{j=1}^{n_i}(\bar{Y}_{i.}-\bar{Y}_{..})^2+2\sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij}-\bar{Y}_{i.})(\bar{Y}_{i.}-\bar{Y}_{..}) \\
&= SSW+SSB+C,
\end{aligned}
$$

where:

$$C = 2\sum_{i=1}^{k}(\bar{Y}_{i.}-\bar{Y}_{..})\sum_{j=1}^{n_i}(Y_{ij}-\bar{Y}_{i.}).$$

Since:

$$\sum_{j=1}^{n_i}(Y_{ij}-\bar{Y}_{i.}) = 0 \quad \text{for each } i,$$

it follows that:

$$C = 0.$$

Thus:

$$SST = SSB + SSW.$$

10.3.6 ANOVA F-Test

Theorem:

1. The least squares estimator for $\mu = (\mu_1, \dots, \mu_k)^\top \in \mathbb{R}^k$ is:

$$\hat{\mu} = (\bar{Y}_{1.}, \dots, \bar{Y}_{k.})^\top.$$

2. Under H_0:

$$\frac{SSW}{\sigma^2} \sim \chi^2(n-k).$$

3. Under H_0:

$$\frac{SSB}{\sigma^2} \sim \chi^2(k-1).$$

4. SSW and SSB are independent under H_0, and the F-statistic defined by:

$$F = \frac{\frac{SSB}{k-1}}{\frac{SSW}{n-k}} \sim F(k-1, n-k) \quad \text{under } H_0.$$

Proof:

1. **Estimator for μ**
   The least squares estimator minimizes SSW. Taking derivatives with respect to μ_i and setting them to zero yields:

   $$\hat{\mu}_i = \bar{Y}_{i.} \quad \text{for each } i = 1, \dots, k.$$

2. **Distribution of SSW under H_0**
   Under H_0, all group means are equal to the overall mean $\bar{Y}_{..}$. Therefore, SSW represents the variability within groups:

   $$SSW = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i.})^2.$$

   Each term $(Y_{ij} - \bar{Y}_{i.})^2$ is chi-squared distributed with 1 degree of freedom. Summing over all groups and observations, we have:

   $$\frac{SSW}{\sigma^2} \sim \chi^2(n-k).$$

3. **Distribution of SSB under H_0**
   Under H_0, the between-group variability SSB is based on deviations of group means from the overall mean:

   $$SSB = \sum_{i=1}^{k} n_i (\bar{Y}_{i.} - \bar{Y}_{..})^2.$$

   Each term $n_i(\bar{Y}_{i.} - \bar{Y}_{..})^2/\sigma^2$ follows a chi-squared distribution with 1 degree of freedom. Summing over k groups, we obtain:

   $$\frac{SSB}{\sigma^2} \sim \chi^2(k-1).$$

4. **Independence of SSB and SSW and F-Statistic Distribution**
   Under H_0, SSB and SSW are independent because SSB depends only on the group means, while SSW depends on the deviations within groups, which are orthogonal to the group means.

The F-statistic is defined as:

$$F = \frac{\frac{SSB}{k-1}}{\frac{SSW}{n-k}}.$$

Since $\frac{SSB}{\sigma^2} \sim \chi^2(k-1)$ and $\frac{SSW}{\sigma^2} \sim \chi^2(n-k)$, and they are independent, it follows that:

$$F \sim F(k-1, n-k) \quad \text{under } H_0.$$

10.4 Conclusion

In this lecture, we established the foundational theorems for hypothesis testing in linear models and ANOVA. We derived the distribution of the F-statistic under the null hypothesis, which allows us to test the equality of multiple group means. Understanding these concepts is crucial for analyzing variance and making informed statistical inferences in various applied contexts.

11 Lecture 12

11.1 Analysis of Variance (ANOVA)

ANOVA is a statistical method used to compare the means of three or more groups to determine if at least one group mean is different from the others. It is particularly useful when dealing with experimental data where multiple treatments or conditions are being tested.

11.1.1 Linear Model for ANOVA

Consider a linear model where we have k groups (or factors/categories), each with n_i observations. The model can be expressed as:

$$Y_{i,j} = \mu_i + \varepsilon_{ij} \quad \text{for } i = 1, \ldots, k \text{ and } j = 1, \ldots, n_i$$

where:

- $Y_{i,j}$ is the j-th observation in the i-th group.
- μ_i is the mean of the i-th group.
- ε_{ij} are the random errors, assumed to be normally distributed with mean 0 and variance σ^2, i.e., $\varepsilon_{ij} \sim N(0, \sigma^2)$.

11.1.2 Ordinary Least Squares (OLS) Estimation

In the context of ANOVA, the OLS estimates of the group means μ_i can be derived. Let X be the design matrix for the linear model. A key note is that:

$$X^T X = \|X\|_{\mathbb{R}^n}^2$$

This relationship is fundamental in deriving the OLS estimates.

11.1.3 Theorem: Properties of the ANOVA Model

Theorem 6. *In the ANOVA model with $\varepsilon_{ij} \sim N(0, \sigma^2)$, the following properties hold:*

1. *The OLS estimate of the group means is given by:*

$$\hat{\mu} = (\bar{y}_{1,\cdot}, \dots, \bar{y}_{k,\cdot}) \quad \textit{where} \quad \bar{y}_{i,\cdot} = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij}$$

2. *The Within-Group Sum of Squares (SSW) scaled by σ^2 follows a chi-squared distribution:*

$$\frac{SSW}{\sigma^2} = \frac{1}{\sigma^2} \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i,\cdot})^2 \sim \chi^2(n-k)$$

where $n = \sum_{i=1}^{k} n_i$ is the total number of observations.

3. *Under the null hypothesis $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$, the Between-Group Sum of Squares (SSB) scaled by σ^2 follows a chi-squared distribution:*

$$\frac{SSB}{\sigma^2} = \frac{1}{\sigma^2} \sum_{i=1}^{k} n_i \left(\bar{y}_{i,\cdot} - \bar{y}\right)^2 \sim \chi^2(k-1)$$

where $\bar{y} = \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{n_i} y_{ij}$ is the overall mean.

4. *SSW and SSB are independent. Under H_0, the ratio*

$$\frac{(n-k)SSB}{(k-1)SSW} \sim F(k-1, n-k)$$

follows an F-distribution with $(k-1, n-k)$ degrees of freedom.

Proof. We will prove each part of the theorem sequentially.

1. **OLS Estimate of μ**: The OLS estimator is given by

$$\hat{\mu} = (X^T X)^{-1} X^T Y$$

Given the structure of $X^T X$, we have:

$$(X^T X)^{-1} = \begin{pmatrix} \frac{1}{n_1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \frac{1}{n_k} \end{pmatrix}$$

Multiplying this with $X^T Y$, we get:

$$\hat{\mu} = \begin{pmatrix} \frac{1}{n_1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \frac{1}{n_k} \end{pmatrix} \begin{pmatrix} \sum_{j=1}^{n_1} Y_{1,j} \\ \vdots \\ \sum_{j=1}^{n_k} Y_{k,j} \end{pmatrix} = \begin{pmatrix} \bar{Y}_{1,\cdot} \\ \vdots \\ \bar{Y}_{k,\cdot} \end{pmatrix}$$

where $\bar{Y}_{i,\cdot} = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{i,j}$.

2. **Distribution of SSW**: The Within-Group Sum of Squares is defined as

$$SSW = \|Y - X\hat{\mu}\|_{\mathbb{R}^n}^2 = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(Y_{ij} - \bar{Y}_{i,\cdot} \right)^2$$

Since $\varepsilon_{ij} \sim N(0, \sigma^2)$, each squared deviation $\left(Y_{ij} - \bar{Y}_{i,\cdot} \right)^2$ follows a scaled chi-squared distribution. Aggregating over all groups and observations, we have:

$$\frac{SSW}{\sigma^2} \sim \chi^2(n-k)$$

where $n-k$ degrees of freedom account for the estimation of k group means.

3. **Distribution of SSB under H_0**: Under the null hypothesis $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$, the Between-Group Sum of Squares is given by

$$SSB = \sum_{i=1}^{k} n_i \left(\bar{Y}_{i,\cdot} - \bar{Y} \right)^2$$

Under H_0, all group means are equal to the overall mean $\bar{Y}$. Therefore, SSB measures the deviation of each group mean from the overall mean. The distribution of SSB scaled by σ^2 is:

$$\frac{SSB}{\sigma^2} \sim \chi^2(k-1)$$

This follows from the properties of chi-squared distributions and the fact that there are $k-1$ independent comparisons among the k group means.

4. **F-Statistic and Independence of SSW and SSB**: The total sum of squares (SST) can be decomposed as:

$$SST = SSW + SSB$$

Under H_0, SSW and SSB are independent because the residuals within groups do not provide information about the between-group variability. The ratio of the scaled sums of squares follows an F-distribution:

$$\frac{(n-k)SSB}{(k-1)SSW} \sim F(k-1, n-k)$$

This result is derived from the general linear hypothesis testing framework, specifically Theorem 2.2.30 in *Methoden der Statistik*. □

11.2 Exponential Families

Exponential families encompass a wide range of probability distributions and include many of the common distributions used in statistics. They also generalize linear models, making them a fundamental concept in statistical theory.

11.2.1 General Model and Hierarchy

$$\text{General Model} \supseteq \text{Exponential Families} \supseteq \text{Linear Models}$$

This hierarchy indicates that exponential families are a subset of general statistical models and that linear models are a subset of exponential families.

11.2.2 Regularity Assumptions

For the theory to hold, certain regularity conditions must be satisfied by the statistical model $(P_\theta : \theta \in \Theta)$:

1. **Dominated Model**: There exists a measure μ such that every P_θ is absolutely continuous with respect to μ, denoted as $P_\theta \ll \mu$ for all $\theta \in \Theta$.
2. **Parameter Space**: The parameter space $\Theta \subseteq \mathbb{R}^p$ is an open set with $p \geq 1$.
3. **Likelihood Function**: The likelihood function $p_\theta(x) > 0$ for all $\theta \in \Theta$ and $x \in X$, ensuring that $\log p_\theta(x)$ is well-defined.

11.2.3 Score Function

Definition 21 (Score). The score vector is defined as the gradient of the log-likelihood with respect to the parameter θ:

$$U_\theta(x) = \nabla_\theta \log p_\theta(x) = \begin{pmatrix} \frac{\partial}{\partial \theta_1} \log p_\theta(x) \\ \vdots \\ \frac{\partial}{\partial \theta_p} \log p_\theta(x) \end{pmatrix}$$

whenever it exists.

11.2.4 Fisher Information

Definition 22 (Fisher Information). For $\theta \in \Theta$, the Fisher Information (FI) matrix is defined as:

$$I(\theta) = \mathbb{E}\left[U_\theta(x)U_\theta(x)^T\right] \in \mathbb{R}^{p\times p}$$

whenever the expectation exists.

11.2.5 Additional Regularity Assumptions

To further develop the theory, we impose additional regularity conditions:

1. **Twice Differentiable Likelihood**: The likelihood function $p_\theta(x)$ is twice differentiable with respect to θ, ensuring that the score function $U_\theta(x)$ is well-defined.
2. **Finite Fisher Information**: For all $\theta \in \Theta$,

   $$\mathbb{E}_\theta\left[\|U_\theta(x)\|^2_{\mathbb{R}^p}\right] < \infty$$

   ensuring that the Fisher Information matrix $I(\theta)$ is well-defined.
3. **Interchange of Integration and Differentiation**: For relevant functions $h(x)$,

   $$\int h(x)\nabla_\theta p_\theta(x)\,\mu(dx) = \nabla_\theta \int h(x)p_\theta(x)\,\mu(dx)$$

   allowing the differentiation under the integral sign.

11.2.6 Properties of the Score and Fisher Information

Lemma 4. *Let $(P_\theta : \theta \in \Theta)$ be a regular statistical model as defined above. Then:*

1. *The expected value of the score function is zero:*

   $$\mathbb{E}_\theta\left[U_\theta(x)\right] = 0$$

2. *The Fisher Information matrix is equal to the covariance matrix of the score function:*

   $$I(\theta) = Cov(U_\theta(x))$$

Proof. 1. **Expected Score is Zero**:

$$\mathbb{E}_\theta\left[U_\theta(x)\right] = \int_X \nabla_\theta \log p_\theta(x)\, p_\theta(x)\,\mu(dx) = \int_X \frac{\nabla_\theta p_\theta(x)}{p_\theta(x)} p_\theta(x)\,\mu(dx) = \int_X \nabla_\theta p_\theta(x)\,\mu(dx)$$

Since $\int_X p_\theta(x)\,\mu(dx) = 1$ for all θ, differentiating both sides with respect to θ gives:

$$\nabla_\theta \int_X p_\theta(x)\,\mu(dx) = \int_X \nabla_\theta p_\theta(x)\,\mu(dx) = 0$$

Therefore, $\mathbb{E}_\theta\left[U_\theta(x)\right] = 0$.

2. **Fisher Information as Covariance**: By definition,

$$I(\theta) = \mathbb{E}\left[U_\theta(x)U_\theta(x)^T\right]$$

Since $\mathbb{E}_\theta\left[U_\theta(x)\right] = 0$, the covariance matrix is:

$$\operatorname{Cov}\left(U_\theta(x)\right) = \mathbb{E}\left[U_\theta(x)U_\theta(x)^T\right] - \mathbb{E}\left[U_\theta(x)\right]\mathbb{E}\left[U_\theta(x)^T\right] = I(\theta)$$ □

11.3 Uniformly Minimum Variance Unbiased Estimators (UMVUE)

In statistical estimation, an unbiased estimator is one whose expected value equals the parameter it estimates. Among all unbiased estimators, the UMVUE is the one with the smallest variance, making it the most efficient unbiased estimator.

11.3.1 Definition

Definition 23 (UMVUE). Let $p(\theta) \in \mathbb{R}$ be a parameter of interest. A statistic $T : X \to \mathbb{R}$ is called a Uniformly Minimum Variance Unbiased Estimator (UMVUE) of $p(\theta)$ if:

- T is unbiased for $p(\theta)$, i.e., $\mathbb{E}_\theta[T(x)] = p(\theta)$ for all $\theta \in \Theta$.
- For any other unbiased estimator $S : X \to \mathbb{R}$ of $p(\theta)$, the variance of T is less than or equal to the variance of S for all $\theta \in \Theta$, i.e.,

$$\operatorname{Var}_\theta(T) \leq \operatorname{Var}_\theta(S) \quad \forall \theta \in \Theta$$

11.3.2 Remarks

1. UMVUEs are the best possible unbiased estimators in terms of variance.
2. Analogous to the Gauss-Markov theorem where the OLS estimator $\hat{\beta}_{OLS}$ is the best linear unbiased estimator (BLUE), the UMVUE is the best unbiased estimator without the linearity constraint.

3. The Mean Squared Error (MSE) of an estimator T can be decomposed as:

$$\mathbb{E}_\theta\left[\|T(x)-\rho(\theta)\|^2\right] = \text{Bias}^2(T) + \text{Var}(T)$$

For unbiased estimators, $\text{Bias}(T) = 0$, hence:

$$\mathbb{E}_\theta\left[\|T(x)-\rho(\theta)\|^2\right] = \text{Var}(T) \leq \mathbb{E}_\theta\left[\|S(x)-\rho(\theta)\|^2\right] = \text{Var}(S)$$

11.3.3 Cramér-Rao Lower Bound

The Cramér-Rao Lower Bound provides a lower bound on the variance of unbiased estimators. It is a fundamental result in estimation theory that helps identify the efficiency of estimators.

Theorem 7 (Cramér-Rao Lower Bound). *Let $(P_\theta : \theta \in \Theta)$ be a regular statistical model. Let $\rho : \Theta \to \mathbb{R}$ be a continuously differentiable function of the parameter θ. For any unbiased estimator T of $\rho(\theta)$, i.e., $\mathbb{E}_\theta[T] = \rho(\theta)$, the variance of T satisfies:*

$$Var_\theta(T) \geq \nabla_\theta \rho(\theta)^T I(\theta)^{-1} \nabla_\theta \rho(\theta)$$

Remark 14. If the parameter space is one-dimensional ($\Theta \subseteq \mathbb{R}$) and $\rho(\theta) = \theta$, then the Cramér-Rao bound simplifies to:

$$\text{Var}_\theta(T) \geq \frac{1}{I(\theta)}$$

where $I(\theta)$ is the Fisher Information.

Proof. Assume that $\Theta \subseteq \mathbb{R}$. Let T be an unbiased estimator of $\rho(\theta)$, so $\mathbb{E}_\theta[T] = \rho(\theta)$.

Consider the covariance between the score function $U_\theta(x)$ and the estimator $T(x)$:

$$\text{Cov}_\theta(U_\theta, T) = \mathbb{E}_\theta[U_\theta T] - \mathbb{E}_\theta[U_\theta]\mathbb{E}_\theta[T] = \mathbb{E}_\theta[U_\theta T]$$

since $\mathbb{E}_\theta[U_\theta] = 0$.

By the Cauchy-Schwarz inequality:

$$|\text{Cov}_\theta(U_\theta, T)|^2 \leq \text{Var}_\theta(U_\theta) \cdot \text{Var}_\theta(T)$$

which implies:

$$\text{Var}_\theta(T) \geq \frac{|\text{Cov}_\theta(U_\theta, T)|^2}{\text{Var}_\theta(U_\theta)}$$

Next, compute the covariance:

$$\mathrm{Cov}_\theta(U_\theta, T) = \mathbb{E}_\theta[U_\theta T] = \int_X U_\theta(x)T(x)p_\theta(x)\,\mu(dx) = \int_X \nabla_\theta \log p_\theta(x)T(x)p_\theta(x)\,\mu(dx)$$

$$= \int_X T(x)\nabla_\theta p_\theta(x)\,\mu(dx) = \nabla_\theta \int_X T(x)p_\theta(x)\,\mu(dx) = \nabla_\theta \mathbb{E}_\theta[T(x)] = \nabla_\theta \rho(\theta)$$

since T is unbiased.

Therefore:

$$\mathrm{Cov}_\theta(U_\theta, T) = \rho'(\theta)$$

and:

$$\mathrm{Var}_\theta(U_\theta) = I(\theta)$$

Substituting these into the inequality:

$$\mathrm{Var}_\theta(T) \geq \frac{(\rho'(\theta))^2}{I(\theta)}$$

In higher dimensions, this generalizes to:

$$\mathrm{Var}_\theta(T) \geq \nabla_\theta \rho(\theta)^T I(\theta)^{-1} \nabla_\theta \rho(\theta)$$

□

11.3.4 Invertibility of the Fisher Information Matrix

An important regularity condition for the Cramér-Rao Lower Bound and related results is that the Fisher Information matrix $I(\theta)$ is invertible. This ensures that the lower bound on the variance of unbiased estimators is well-defined and finite.

11.4 Conclusion

In this lecture, we delved into the foundations of ANOVA, exploring the linear model framework, OLS estimation, and the distributional properties of the sum of squares under various hypotheses. We also introduced exponential families, discussing their place within the broader context of statistical models and the regularity conditions required for their analysis. Furthermore, we examined the concept of UMVUEs and established the Cramér-Rao Lower Bound, providing essential tools for assessing the efficiency of estimators.

Understanding these concepts is crucial for conducting rigorous statistical analysis and for developing efficient estimation procedures in various applications.

12 Lecture 13

Regular Statistical Models

In this lecture, we delve into the framework of regular statistical models, explore the Cramér-Rao (CR) Inequality, and examine the structure of exponential families. We will begin by defining regular statistical models and proceed to establish foundational results, including the Fisher Information Matrix and the CR Inequality. Finally, we will explore the properties and examples of exponential families.

Definition of a Regular Statistical Model

A **regular statistical model** is characterized by a parameter space and a family of probability density functions that satisfy certain smoothness and positivity conditions. Formally, a statistical model is regular if it satisfies the following:

- $\Theta \subset \mathbb{R}^p$ is an open set, where p denotes the number of parameters.
- For every $\vartheta \in \Theta$ and $x \in \mathcal{X}$, the density $p_\vartheta(x)$ is strictly positive, i.e., $p_\vartheta(x) > 0$.
- The density $p_\vartheta(x)$ is continuously differentiable with respect to the parameter ϑ.

These conditions ensure that the model is smooth enough to allow for the interchange of differentiation and integration, which is crucial for deriving key statistical properties.

Fisher Information Matrix

Given a regular statistical model, the **Fisher Information Matrix** quantifies the amount of information that an observable random variable X carries about the unknown parameter ϑ. It is defined as:

$$I(\vartheta) = \mathbb{E}_\vartheta \left[\nabla_\vartheta \log p_\vartheta(X) \cdot \nabla_\vartheta \log p_\vartheta(X)^T \right]$$

Here, ∇_ϑ denotes the gradient with respect to the parameter vector ϑ, and the expectation is taken with respect to the distribution p_ϑ.

- The Fisher Information Matrix $I(\vartheta)$ exists for all $\vartheta \in \Theta$.

- $I(\vartheta)$ is positive definite, which implies that its inverse $I(\vartheta)^{-1}$ exists.

The positivity and invertibility of $I(\vartheta)$ are critical for deriving lower bounds on the variance of unbiased estimators, as we will see next.

Cramér-Rao / Information Inequality

The Cramér-Rao Inequality provides a lower bound on the variance of unbiased estimators of a parameter. This fundamental result connects the variance of an estimator to the Fisher Information, establishing a benchmark for the efficiency of estimators.

Theorem 8 (Cramér-Rao Inequality). *Let $(p_\vartheta, \vartheta \in \Theta)$ be a regular statistical model. Suppose $g : \Theta \to \mathbb{R}$ is a continuously differentiable function. Let $T : \mathcal{X} \to \mathbb{R}$ be an unbiased estimator of $g(\vartheta)$, meaning that $\mathbb{E}_\vartheta[T] = g(\vartheta)$ for all $\vartheta \in \Theta$.*

Then, for all $\vartheta \in \Theta$,

$$Var_\vartheta(T) \geq \left(g'(\vartheta)\right)^T I(\vartheta)^{-1} g'(\vartheta).$$

This inequality is known as the ***Cramér-Rao / Information Inequality****.*

Score Vector

The **Score Vector** plays a pivotal role in the analysis of statistical models. It is defined as the gradient of the log-likelihood function with respect to the parameter vector.

$$U_\vartheta(x) = \nabla_\vartheta \log p_\vartheta(x)$$

The Score Vector captures the sensitivity of the log-likelihood to changes in the parameter ϑ.

Remarks

- If the Fisher Information Matrix $I(\vartheta)$ is large, it suggests that more information is contained in the data about the parameter ϑ, thereby enabling better estimation.
- The Fisher Information Matrix provides another interpretation related to the curvature of the log-likelihood function, linking it to the precision of parameter estimates.

Derivation of the Fisher Information

Consider the case where the parameter space Θ is one-dimensional, i.e., $\Theta \subseteq \mathbb{R}$. Suppose that the density $p_\vartheta(x)$ is twice differentiable with respect to ϑ. We derive expressions for the first and second derivatives of the log-likelihood function.

$$(\log p_\vartheta(x))' = \frac{p'_\vartheta(x)}{p_\vartheta(x)}$$

$$(\log p_\vartheta(x))'' = \frac{p''_\vartheta(x) p_\vartheta(x) - (p'_\vartheta(x))^2}{p_\vartheta(x)^2}$$

Taking the expectation of the first derivative:

$$\mathbb{E}_\vartheta \left[(\log p_\vartheta(x))' \right] = \int_x \frac{p'_\vartheta(x)}{p_\vartheta(x)} p_\vartheta(x) dx = \int_x p'_\vartheta(x) dx = \frac{d}{d\vartheta} \int_x p_\vartheta(x) dx = 0.$$

This result follows from the fact that the integral of a probability density function over its entire support is always 1, and hence its derivative with respect to ϑ is zero.

Next, we compute the expectation of the square of the first derivative:

$$\mathbb{E}_\vartheta \left[(\log p_\vartheta(x))^2 \right] = -\mathbb{E}_\vartheta \left[(\log p_\vartheta(x))'' \right] = -\mathbb{E}_\vartheta \left[U_\vartheta(x)^2 \right] = -I(\vartheta).$$

This establishes the relationship between the second derivative of the log-likelihood and the Fisher Information.

Attaining the Cramér-Rao Bound

The Cramér-Rao Inequality provides a lower bound on the variance of unbiased estimators. However, it is often of interest to determine when this bound is attained. The following theorem characterizes the conditions under which an estimator achieves the Cramér-Rao bound.

Theorem 9. *Let $(P_\theta, \theta \in \Theta)$ be a regular model with $\Theta \subseteq \mathbb{R}$. Let $\rho : \Theta \to \mathbb{R}$ be a continuously differentiable function, and let T be an unbiased estimator of $\rho(\theta)$, i.e., $\mathbb{E}_\theta[T] = \rho(\theta)$ for all $\theta \in \Theta$.*

Then, T attains equality in the Cramér-Rao bound if and only if

$$T(x) = \rho(\theta) + \rho'(\theta) I(\theta)^{-1} U_\theta(x)$$

almost surely for all $\theta \in \Theta$.

Proof. Define $v(\theta) = \rho'(\theta)I(\theta)^{-1}$. Consider the estimator T as given. We analyze the variance:

$$0 \leq \text{Var}(T - v(\theta)U_\theta) = \text{Var}(T) + v(\theta)^2\mathbb{E}_\theta[U_\theta^2] - 2v(\theta)\text{Cov}_\theta(T, U_\theta)$$

Substituting the definitions, we have:

$$\text{Var}(T) - \rho'(\theta)^2 I(\theta)^{-1} \geq 0$$

Equality holds if and only if:

$$T - v(\theta)U_\theta = \text{Constant}$$

Since T is unbiased, we have $\mathbb{E}_\theta[T] = \rho(\theta)$, which implies that the constant must be $\rho(\theta)$. Therefore:

$$T = \rho(\theta) + v(\theta)U_\theta$$

This establishes the necessary and sufficient condition for T to attain the Cramér-Rao bound. □

Remark 15.

1. The estimator $T(x)$ in the above equation is not always a measurable function of x.
2. If T attains the Cramér-Rao bound, we say that T is a **Cramér-Rao efficient** estimator.

Corollary: Exponential Family Representation

Under certain scaling conditions, the likelihood can be expressed in the form of an exponential family. Specifically, we have the following corollary.

Corollary 2. *Assume the previous scaling conditions and that $\rho(\theta) \neq 0$ for all $\theta \in \Theta$. Then, the likelihood can be written as:*

$$p_\theta(x) = c(x) \exp\left(n(\theta)T(x) - \Psi(\theta)\right)$$

where:

- *$n : \Theta \to \mathbb{R}$ satisfies $n'(\theta) = \frac{I(\theta)}{\rho'(\theta)}$.*
- *$c(x)$ and $\Psi(\theta)$ are functions that ensure the equality holds.*

Proof. From the previous theorem, we have:

$$T(x) = \rho(\theta) + \rho'(\theta) I^{-1}(\theta) U_\theta(x)$$

Substituting $U_\theta(x) = \frac{d}{d\theta} \log p_\theta(x)$, we obtain:

$$T(x) = \rho(\theta) + \rho'(\theta) I^{-1}(\theta) \frac{d}{d\theta} \log p_\theta(x)$$

Rearranging terms, we get:

$$(T(x) - \rho(\theta)) \frac{I(\theta)}{\rho'(\theta)} = \frac{d}{d\theta} \log p_\theta(x)$$

Integrating both sides with respect to θ yields:

$$\log p_\theta(x) = n(\theta) T(x) - \Psi(\theta) + \text{constant}(x)$$

Exponentiating both sides gives:

$$p_\theta(x) = c(x) \exp\left(n(\theta) T(x) - \Psi(\theta)\right)$$

where $c(x) = \exp(\text{constant}(x))$ and $\Psi(\theta)$ absorbs any remaining terms. □

Exponential Families

Exponential families are a broad class of probability distributions that have a particular structure, allowing for elegant mathematical properties and ease of analysis. They are central to many areas of statistics, including generalized linear models and Bayesian statistics.

Definition 24 (Exponential Families). A regular statistical model $(P_\theta : \theta \in \Theta)$ is called a **k-parameter Exponential Family** ($k \geq 1$) if there exist measurable functions:

1. $n : \Theta \to \mathbb{R}^k$
2. $T : \mathcal{X} \to \mathbb{R}^k$
3. $c : \mathcal{X} \to [0, \infty)$

such that the density can be expressed as:

$$p_\theta(x) = c(x) \exp\left(\langle n(\theta), T(x) \rangle_{\mathbb{R}^k} - \Psi(\theta)\right)$$

for all $\theta \in \Theta$ and $x \in \mathcal{X}$, where:

$$\Psi(\theta) = \log \left(\int_{\mathcal{X}} c(x) \exp\left(\langle n(\theta), T(x) \rangle_{\mathbb{R}^k}\right) d\mu(x) \right)$$

Remark 16. 1. A key feature of exponential families is the factorization of the term $\langle n(\theta), T(x) \rangle_{\mathbb{R}^k}$, which allows for the derivation of sufficient statistics and facilitates the computation of maximum likelihood estimates.

2. Exponential forms are motivated by the quest to find general models where Cramér-Rao efficient procedures exist, leveraging the structure provided by the exponential family.

Examples of Exponential Families

Example 1: Binomial Distribution

Consider the Binomial distribution with parameters n (number of trials) and θ (probability of success). The probability mass function is:

$$p_\theta(k) = \binom{n}{k} \theta^k (1-\theta)^{n-k}$$

This can be rewritten in exponential family form as:

$$\begin{aligned} p_\theta(k) &= \binom{n}{k} \exp\left(k \log\theta + (n-k)\log(1-\theta)\right) \\ &= \underbrace{\binom{n}{k}}_{c(k)} \exp\left(\underbrace{k}_{T(k)} \underbrace{\log\left(\frac{\theta}{1-\theta}\right)}_{n(\theta)} + \underbrace{n\log(1-\theta)}_{\Psi(\theta)} \right) \end{aligned}$$

Here, we identify:

- $c(k) = \binom{n}{k}$
- $T(k) = k$
- $n(\theta) = \log\left(\frac{\theta}{1-\theta}\right)$
- $\Psi(\theta) = n\log(1-\theta)$

Example 2: Normal Distribution

Consider the Normal distribution with parameters μ (mean) and σ^2 (variance), denoted as $N(\mu, \sigma^2)$. The probability density function is:

$$p_{\mu,\sigma^2}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

This can be expressed in exponential family form by expanding the exponent:

$$p_{\mu,\sigma^2}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{x^2 - 2x\mu + \mu^2}{2\sigma^2}\right)$$

By identifying appropriate functions, we have:

$$p_{\mu,\sigma^2}(x) = \underbrace{\frac{1}{\sqrt{2\pi\sigma^2}}}_{c(x)} \exp\left(\langle n(\theta), T(x)\rangle - \Psi(\theta)\right)$$

where:

- $T(x) = \begin{pmatrix} x^2 \\ x \end{pmatrix}$
- $n(\theta) = \begin{pmatrix} -\frac{1}{2\sigma^2} \\ \frac{\mu}{\sigma^2} \end{pmatrix}$
- $\Psi(\theta) = \frac{\mu^2}{2\sigma^2} + \frac{1}{2}\log(2\pi\sigma^2)$

Example 3: Poisson Distribution

Consider the Poisson distribution with parameter $\theta > 0$. The probability mass function is:

$$p_\theta(k) = e^{-\theta}\frac{\theta^k}{k!}$$

This can be written in exponential family form as:

$$p_\theta(k) = \underbrace{\frac{1}{k!}}_{c(k)} \exp\left(k \log\theta - \theta\right)$$

Here, we identify:

- $c(k) = \frac{1}{k!}$
- $T(k) = k$
- $n(\theta) = \log \theta$
- $\Psi(\theta) = \theta$

Natural Exponential Families

Natural exponential families are a subclass of exponential families where the natural parameter space coincides with the interior of the set where the log-partition function is finite.

Definition 25 (Natural Exponential Family). Define:

$$\Xi = \left\{ n \in \mathbb{R}^k \;\middle|\; \int_{\mathcal{X}} c(x) \exp\left(\langle n, T(x) \rangle_{\mathbb{R}^k}\right) d\mu(x) < \infty \right\}$$

A model $(P_n \mid n \in \Xi)$ is called a **Natural Exponential Family** if:

- $\Theta = \Xi$
- The density can be written as:

  $$\frac{dP_n}{d\mu}(x) = c(x) \exp\left(\langle n, T(x) \rangle_{\mathbb{R}^k} - \Psi(n)\right)$$

  for all $x \in \mathcal{X}$ and $n \in \Xi$.

Remark 17. A natural exponential family is fully specified by the functions $c(x)$, $T(x)$, and the natural parameter space Ξ. This structure facilitates the derivation of properties such as sufficient statistics and conjugate priors in Bayesian analysis.

Properties of Natural Exponential Families

We explore some key properties of natural exponential families, particularly focusing on the relationship between the log-partition function and moments of the sufficient statistic.

Lemma 5. *Let $(P_n \mid n \in \Xi)$ be a 1-parameter natural exponential family. For all $n \in int(\Xi)$,*

1. $\Psi'(n) = \mathbb{E}_n[T]$
2. $\Psi''(n) = Var_n[T]$

Proof. Let $n \in \mathrm{int}(\Xi)$ and define:

$$\gamma(n) = e^{\Psi(n)} = \int_{\mathcal{X}} c(x) \exp(nT(x)) d\mu(x)$$

We first show that γ is infinitely differentiable at n. Observe that:

$$\frac{d}{dn} \left(c(x) \exp(nT(x)) \right) = c(x)T(x) \exp(nT(x))$$

To apply the Dominated Convergence Theorem (DCT), we require that for some $\varepsilon > 0$, the function $c(x)T(x)\exp(nT(x))$ is dominated by an integrable function. Specifically, for sufficiently small ε, we have:

$$T(x) \exp(nT(x)) \leq C \exp((n + \varepsilon)T(x))$$

for some constant $C > 0$, ensuring integrability.

Applying DCT, we differentiate $\gamma(n)$:

$$\gamma'(n) = \int_{\mathcal{X}} c(x)T(x) \exp(nT(x)) d\mu(x) = \mathbb{E}_n[T]$$

Therefore, the derivative of $\Psi(n)$ is:

$$\Psi'(n) = \frac{d}{dn} \log \gamma(n) = \frac{\gamma'(n)}{\gamma(n)} = \mathbb{E}_n[T]$$

Differentiating again, we obtain:

$$\Psi''(n) = \frac{d}{dn} \mathbb{E}_n[T] = \mathrm{Var}_n[T]$$

This follows from the properties of exponential families, where the second derivative of the log-partition function corresponds to the variance of the sufficient statistic. □

Remark 18. The lemma highlights the intimate connection between the derivatives of the log-partition function $\Psi(n)$ and the moments of the sufficient statistic T. Specifically, the first derivative gives the mean, and the second derivative gives the variance, providing a direct link between the parameters of the exponential family and the moments of the data.

13 Lecture 14: Exponential Families and Generalized Linear Models

13.1 Exponential Families

In statistical theory, exponential families play a central role due to their rich mathematical structure and wide applicability. They encompass many of the commonly used statistical distributions and offer convenient properties for estimation and inference.

13.1.1 Definition

Let $\Theta \subseteq \mathbb{R}$ be an open set. A family of probability distributions $\{P_\theta\}_{\theta\in\Theta}$ on a measurable space $(X, \mathcal{A})$ is called an **exponential family** if the probability density function (with respect to a measure μ) can be expressed in the form:

$$p_\theta(x) = c(x) \exp\{n(\theta)T(x) - \Psi(\theta)\}, \tag{13.1}$$

where:

- $c(x)$ is a non-negative function that does not depend on θ,
- $n(\theta)$ is a real-valued function of θ,
- $T(x)$ is a sufficient statistic,
- $\Psi(\theta)$ is the log-partition function ensuring that the density integrates to one.

13.1.2 Properties

1. **Natural Exponential Family (NEF)**: When the density takes the form

$$p_n(x) = c(x) \exp\{nT(x) - \Psi(n)\}, \tag{13.2}$$

the family is called a *natural* exponential family. Here, the parameter n is known as the *natural parameter*.

2. **Natural Parameter Space**: The set of all values t for which the integral converges is known as the natural parameter space Ξ:

$$\Xi = \left\{ t \in \mathbb{R} \;\middle|\; \int_X c(x) \exp\left(tT(x)\right) d\mu(x) = e^{\Psi(t)} < \infty \right\}. \tag{13.3}$$

   It is important to verify that Ξ is an open interval in $\mathbb{R}$.

3. **Regularity Assumptions**: For statistical inference, we often require that $p_\theta(x)$ satisfies certain regularity conditions, such as differentiability with respect to θ, to ensure the validity of results like the Cramér-Rao lower bound.

13.1.3 Lemma

Lemma 6. *Let $(P_n)_{n\in\Xi}$ be a natural and regular exponential family. Then, for every $n \in \Xi$:*

1. $\Psi'(n) = E_n[T]$,

2. $\Psi''(n) = \mathrm{Var}_n(T)$.

Proof. 1. This result was established in the previous lecture. Briefly, differentiating $\Psi(n)$ with respect to n, we obtain the expected value of the sufficient statistic T under P_n:

$$\Psi'(n) = \frac{d}{dn}\Psi(n) = E_n[T].$$

2. Recall that the log-partition function $\Psi(n)$ is given by

$$\Psi(n) = \log \int_X c(x) \exp\left(nT(x)\right) d\mu(x) = \log \alpha(n),$$

where

$$\alpha(n) = \int_X c(x) \exp\left(nT(x)\right) d\mu(x).$$

We have previously shown that $\alpha(n)$ is a C^∞ (infinitely differentiable) function on Ξ. Differentiating $\Psi(n)$ twice with respect to n, we get:

$$\begin{aligned}\Psi''(n) &= \frac{d^2}{dn^2}\log\alpha(n) = \frac{\alpha''(n)}{\alpha(n)} - \left(\frac{\alpha'(n)}{\alpha(n)}\right)^2 \\ &= \int_X c(x)T(x)^2 \exp\left(nT(x)\right) d\mu(x) - \left(\int_X c(x)T(x)\exp\left(nT(x)\right) d\mu(x)\right)^2 \\ &= E_n[T^2] - (E_n[T])^2 = \mathrm{Var}_n(T).\end{aligned}$$

□

13.1.4 Examples

Exponential families include many common distributions. Often, the sufficient statistic $T(x)$ is a simple function, such as $T(x) = x$. Let's consider two examples:

1. **Binomial Distribution**:

   Let $X \sim \text{Bin}(n, \theta)$, where $\theta \in (0, 1)$. The probability mass function is

   $$p_\theta(k) = \binom{n}{k} \theta^k (1 - \theta)^{n-k}.$$

   We can express this in exponential family form by setting $T(k) = k$ and $n = \log\left(\frac{\theta}{1-\theta}\right)$. Then,

   $$\begin{aligned} p_\theta(k) &= \binom{n}{k} \exp\left(k \log \frac{\theta}{1-\theta} + n \log(1-\theta)\right) \\ &= \binom{n}{k} \exp\left(kn - n \log\left(\frac{1}{1-\theta}\right)\right). \end{aligned}$$

   The log-partition function is

   $$\begin{aligned} \Psi(n) = -n \log(1-\theta) &= -n \log\left(1 - \frac{e^n}{1+e^n}\right) \\ &= -n \log\left(\frac{1}{1+e^n}\right) = n \log(1+e^n). \end{aligned}$$

   We can compute the first and second derivatives:

   $$\Psi'(n) = \frac{d}{dn}\Psi(n) = n\frac{e^n}{1+e^n} = n\theta = E_n[T],$$

   which corresponds to the mean of the binomial distribution $E_n[T] = n\theta$. Similarly,

   $$\Psi''(n) = \text{Var}_n(T) = n\theta(1-\theta),$$

   which is the variance of the binomial distribution.

2. **Poisson Distribution**:

   Let $X \sim \text{Poisson}(\lambda)$, with $\lambda > 0$. The probability mass function is

   $$p_\lambda(k) = e^{-\lambda}\frac{\lambda^k}{k!}.$$

   This can be rewritten in exponential family form by setting $T(k) = k$, $n = \log \lambda$, and $\Psi(n) = e^n$. Then,

$$p_n(k) = \frac{1}{k!} \exp\left(kn - e^n\right).$$

The derivatives of $\Psi(n)$ are

$$\Psi'(n) = e^n, \quad \Psi''(n) = e^n.$$

Therefore,

$$E_n[T] = \Psi'(n) = e^n = \lambda, \quad \mathrm{Var}_n(T) = \Psi''(n) = e^n = \lambda.$$

13.2 Maximum Likelihood Estimators and UMVUE in Exponential Families

Exponential families have convenient properties for estimation, particularly for maximum likelihood estimators (MLEs) and uniformly minimum variance unbiased estimators (UMVUE).

13.2.1 Theorem

Theorem 10 (MLEs in Natural Exponential Families and UMVUE Estimators). *Let $(P_\theta)_{\theta\in\Xi}$ be a natural one-parameter regular exponential family. Then:*

1. *If a unique maximum likelihood estimator $\hat{\theta}_{MLE}$ exists, then it is given by*

$$\hat{\theta}_{MLE} = (\Psi')^{-1}(T).$$

2. *For a function $\rho(\theta) = E_\theta[T]$, the statistic T is the uniformly minimum variance unbiased estimator (UMVUE) for $\rho(\theta)$.*

Proof. 1. The maximum likelihood estimator $\hat{\theta}_{\mathrm{MLE}}$ maximizes the likelihood function. Considering the logarithm of the likelihood for a single observation (since the family is regular and observations are independent), we have:

$$\begin{aligned}
\hat{\theta}_{\mathrm{MLE}} &= \arg\max_{\theta\in\Xi} p_\theta(x) \\
&= \arg\max_{\theta} \log p_\theta(x) \\
&= \arg\max_{\theta} \left(\theta T(x) - \Psi(\theta)\right).
\end{aligned}$$

Differentiating with respect to θ and setting the derivative to zero yields:

$$\frac{d}{d\theta}\left(\theta T(x) - \Psi(\theta)\right)\bigg|_{\theta=\hat{\theta}} = 0 \implies T(x) = \Psi'(\hat{\theta}).$$

Since $\Psi''(\theta) = \mathrm{Var}_\theta(T) > 0$, Ψ' is strictly increasing, and thus invertible. Therefore, the MLE is given by

$$\hat{\theta}_{\mathrm{MLE}} = (\Psi')^{-1}(T(x)).$$

2. To show that T is the UMVUE for $\rho(\theta) = E_\theta[T]$, we utilize the Cramér-Rao lower bound. For any unbiased estimator S of $\rho(\theta)$, the variance satisfies

$$\mathrm{Var}_\theta(S) \geq \frac{(\rho'(\theta))^2}{I(\theta)},$$

where $I(\theta)$ is the Fisher information, and $\rho'(\theta)$ is the derivative of $\rho(\theta)$ with respect to θ.

In the exponential family, we have:

$$\rho'(\theta) = \frac{d}{d\theta}E_\theta[T] = \Psi''(\theta) = \mathrm{Var}_\theta(T).$$

The score function is:

$$\ell'(\theta; x) = \frac{\partial}{\partial\theta}\left(\theta T(x) - \Psi(\theta)\right) = T(x) - \Psi'(\theta).$$

The Fisher information is:

$$I(\theta) = E_\theta\left[\left(\ell'(\theta; X)\right)^2\right] = \mathrm{Var}_\theta(T).$$

Therefore, the Cramér-Rao lower bound becomes:

$$\mathrm{Var}_\theta(S) \geq \frac{(\mathrm{Var}_\theta(T))^2}{\mathrm{Var}_\theta(T)} = \mathrm{Var}_\theta(T).$$

Since T achieves this bound (its variance equals the lower bound), it is the UMVUE for $\rho(\theta) = E_\theta[T]$. □

13.3 Generalized Linear Models (GLMs)

Generalized Linear Models extend traditional linear models to accommodate response variables that have error distribution models other than a normal distribution. They are particularly useful when dealing with non-normal data, such as binary or count data.

13.3.1 Motivation

Traditional linear models establish a relationship between a dependent variable Y and independent variables X, assuming that the residuals are normally distributed. However, many types of data (e.g., binary outcomes, counts) do not satisfy this assumption.

By combining linear models with exponential families, we obtain GLMs, which allow for:

- Establishing relations between predictors and responses ($X_i \to Y_i$),
- Modeling different types of data (continuous, discrete),
- Using link functions to relate the mean of the distribution to the linear predictors.

13.3.2 Definition

Given data (X_i, Y_i), where $X_i \in \mathbb{R}^p$ are covariates and Y_i are responses, a GLM assumes:

1. The distribution of Y_i belongs to an exponential family with density:

$$p_{\theta_i}(y_i) = c(y_i) \exp\left(\theta_i y_i - \Psi(\theta_i)\right),$$

   where θ_i is the natural parameter associated with Y_i.

2. A link function g relates the expected value $\mu_i = E[Y_i|X_i]$ to the linear predictors:

$$g(\mu_i) = X_i^\top \beta.$$

The function $g : \mathbb{R} \to \mathbb{R}$ is known as the **link function**. When $g(\mu_i) = \theta_i$, it is called the **canonical link function**.

13.3.3 Remarks

1. Under the canonical link function, the density simplifies to:

$$p_\beta(y_i) = c(y_i) \exp\left(y_i X_i^\top \beta - \Psi(X_i^\top \beta)\right).$$

2. The link function connects the linear predictors to the mean of the response variable:
$$E[Y_i|X_i, \beta] = g^{-1}(X_i^\top \beta).$$

13.3.4 Examples

Example 6 (Logistic Regression). Suppose we have binary response data $Y_i \in \{0, 1\}$ and covariates $X_i \in \mathbb{R}^p$.

1. The distribution of Y_i given X_i is Bernoulli with probability p_i:
$$P(Y_i = 1|X_i) = p_i.$$

2. The logistic regression model uses the logit link function:
$$\log\left(\frac{p_i}{1-p_i}\right) = X_i^\top \beta.$$
Equivalently,
$$p_i = \frac{\exp(X_i^\top \beta)}{1 + \exp(X_i^\top \beta)}.$$

3. The natural parameter is $\theta_i = \log\left(\frac{p_i}{1-p_i}\right)$, and the canonical link function is the logit function.

Example 7 (Poisson Regression). Suppose we have count data $Y_i \in \{0, 1, 2, \dots\}$ and covariates $X_i \in \mathbb{R}^p$.

1. The distribution of Y_i given X_i is Poisson with mean λ_i:
$$P(Y_i = k|X_i) = \frac{e^{-\lambda_i}\lambda_i^k}{k!}.$$

2. The Poisson regression model uses the log link function:
$$\log \lambda_i = X_i^\top \beta.$$
Equivalently,
$$\lambda_i = \exp(X_i^\top \beta).$$

3. The natural parameter is $\theta_i = \log \lambda_i$, and the canonical link function is the logarithm.

4. This model is commonly used for modeling count data, such as the number of events occurring in a fixed interval.

13.3.5 Link Functions

The choice of the link function g is critical in GLMs. It should be chosen based on the nature of the response variable and the desired relationship between the mean of the response and the linear predictors.

- **Canonical Link Function**: The link function that sets $g(\mu_i) = \theta_i$. It often simplifies calculations and has desirable statistical properties.
- **Common Link Functions**:
  - **Identity Link**: $g(\mu) = \mu$, used in linear regression.
  - **Logit Link**: $g(\mu) = \log\left(\frac{\mu}{1-\mu}\right)$, used in logistic regression.
  - **Log Link**: $g(\mu) = \log(\mu)$, used in Poisson regression.
  - **Probit Link**: $g(\mu) = \Phi^{-1}(\mu)$, where Φ is the cumulative distribution function of the standard normal distribution.

13.3.6 Estimation in GLMs

Estimation of the parameter vector β in GLMs is typically performed using maximum likelihood estimation. Due to the nonlinear nature of the link functions, closed-form solutions are rare, and iterative algorithms such as Newton-Raphson or Fisher scoring are employed.

13.3.7 Inference in GLMs

Statistical inference in GLMs involves:

- **Hypothesis Testing**: Testing hypotheses about the parameters β, such as testing whether a particular predictor has a significant effect.
- **Confidence Intervals**: Constructing confidence intervals for the parameters.
- **Model Diagnostics**: Assessing the goodness-of-fit of the model and checking model assumptions.

13.3.8 Advantages of GLMs

GLMs offer several advantages:

- Flexibility in modeling different types of data (binary, count, continuous).
- Ability to handle non-normal error distributions.
- Unified framework for estimation and inference.

13.4 Conclusion

Exponential families provide a powerful and flexible class of probability distributions with properties that facilitate statistical inference. Generalized Linear Models extend the linear modeling framework to accommodate response variables that are not normally distributed, leveraging the structure of exponential families. Understanding these concepts is essential for modeling a wide range of data types and for conducting effective statistical analysis in various fields.

14 Lecture 15: GLMs & Model Selection

Natural Exponential Family (Nat. EF)

We consider a distribution in the natural exponential family:

$$P_\eta(x) = c(x)\exp(\eta T(x) - \Psi(\eta)), \quad \eta \in \mathbb{R}.$$

- **What is the Fisher information?**

  The Fisher information for a single observation can be computed as:

  $$I(\eta) = \mathbb{E}_\eta\left[\left(\frac{\partial}{\partial\eta}\log P_\eta(X)\right)^2\right].$$

  Given $P_\eta(x) = c(x)\exp(\eta T(x) - \Psi(\eta))$, we have:

  $$\log P_\eta(x) = \log c(x) + \eta T(x) - \Psi(\eta).$$

  Differentiating with respect to η:

  $$\frac{\partial}{\partial\eta}\log P_\eta(x) = T(x) - \Psi'(\eta).$$

  Thus, the Fisher information is:

  $$I(\eta) = \mathbb{E}_\eta[(T(X) - \Psi'(\eta))^2] = \mathbb{E}_\eta[T(X)^2] - (\Psi'(\eta))^2 = \Psi''(\eta).$$

- **Is the Cramer-Rao lower bound attained?**

  In a one-parameter natural exponential family with a sufficient statistic that is not degenerate, the Cramer-Rao lower bound is indeed attained by the UMVUE (uniformly minimum variance unbiased estimator).

  $$\mathbb{E}_\eta[T] = \Psi'(\eta).$$

- **What happens if $T = \text{const.}$ a.s.?**

If T is almost surely a constant (does not vary with the data), then the parameter η cannot be identified from the data. In this degenerate case, the model carries no information about η, so the Fisher information would be zero, and we cannot estimate η.

Generalized Linear Models (GLM)

- Parameter: $\beta \in \mathbb{R}^p$.
- Design matrix: $X \in \mathbb{R}^{n \times p}$.
- Data: $Y_i \sim c(y) \exp\left(y X_i^\top \beta - \Psi(X_i^\top \beta)\right)$.

  We have:

$$g(\mathbb{E}[Y_i \mid X_i, \beta]) = X_i^\top \beta$$

  for some (natural) link function $g : \mathbb{R} \to \mathbb{R}$.

Example: Logistic Regression

$$Y_i \in \{0, 1\}, \quad \mathbb{E}[Y_i] = P(Y_i = 1) = p_i.$$

- Natural link function:

$$g(p_i) = \log \frac{p_i}{1 - p_i}, \quad g^{-1}(X_i^\top \beta) = \frac{\exp(X_i^\top \beta)}{1 + \exp(X_i^\top \beta)}.$$

- The probability mass function:

$$P_\beta(Y_i) = p_i^{Y_i} (1 - p_i)^{1 - Y_i}$$

$$= \exp\left(Y_i \log p_i + (1 - Y_i) \log(1 - p_i)\right)$$

$$= \exp\left(Y_i \log \frac{p_i}{1 - p_i} + \log(1 - p_i)\right)$$

$$= \exp\left(Y_i X_i^\top \beta - \Psi(X_i^\top \beta)\right), \quad \Psi(X_i^\top \beta) = \log(1 + \exp(X_i^\top \beta)).$$

MLE in Logistic Regression

Goal: Find $\hat{\beta} \in \arg\max_{\beta \in \mathbb{R}^p} P_\beta(Y)$.

$$
\begin{aligned}
\log p_\beta(Y) &= \sum_{i=1}^n \log p_i^{Y_i}(1-p_i)^{1-Y_i} \\
&= \sum_{i=1}^n Y_i \log \frac{p_i}{1-p_i} + \log(1-p_i) \\
&= \sum_{i=1}^n y_i \underbrace{\eta_i}_{X_i^T\beta} + \log \frac{1}{1+e^{\eta_i}}
\end{aligned}
$$

Differentiating:

$$\nabla_\beta \log p_\beta(Y) = \sum_{i=1}^n Y_i \nabla_\beta \eta_i - \frac{1}{1+e^{\eta_i}} e^{\eta_i} \nabla_\beta \eta_i.$$

Note that $\nabla_\beta \eta_i = X_i^T$. Thus:

$$\nabla_\beta \log p_\beta(Y) = \sum_{i=1}^n (Y_i - p_i) X_i^T = X^T(Y-p) \in \mathbb{R}^p.$$

The MLE in Logistic Regression solves:

$$X^T(Y-p) = 0,$$

where $p = (p_1, \ldots, p_n)$ and $p_i = \frac{e^{X_i^T\beta}}{1+e^{X_i^T\beta}}$.

Caution: $\hat{\beta}$ is not always defined.

We know:

$$E[Y_i] = P(Y_i = 1) = \frac{e^{X_i^T\beta}}{1+e^{X_i^T\beta}}.$$

Remark 19. $\hat{\beta}$, the MLE in GLMs, are not given in closed form, and have to be computed using optimization methods (e.g. gradient descent, Newton's Method).

Lemma 7 (Score and FI in GLMs). *Consider a GLM with canonical links. Then,*

$$\nabla_\beta \log p_\beta(Y) = \sum_{i=1}^n (Y_i - \Psi'(X_i^T\beta)) X_i^T$$

$$I(\beta) = \sum_{i=1}^n \Psi''(X_i^T\beta) X_i X_i^T \in \mathbb{R}^p.$$

Proof.

$$\log p_\beta(Y) = \sum_{i=1}^n \log c(y_i) + y_i X_i^T\beta - \Psi(X_i^T\beta)$$

$$\implies \nabla_\beta \log p_\beta(Y) = \sum_{i=1}^n Y_i X_i^T - \Psi'(X_i^T\beta)X_i^T = \sum_{i=1}^n (Y_i - \Psi'(X_i^T\beta))X_i^T.$$

One checks that

$$\frac{\partial^2}{\partial\beta_k\beta_\ell}\log p_\beta(Y) = \sum_{i=1}^n -\Psi''(X_i^T\beta)(X_i)_k(X_i)_\ell.$$

Since $I(\beta) = E_\beta[-\nabla_\beta^2(\log p_\beta(x))]$, the claim follows. □

Remark 20.

1. In natural ER and in GLMs, the curvature of $\log p_\beta$ is independent of the data!
2. If $\hat\beta$ MLE exists in $\mathbb{R}^p$ and $I(\beta)$ is positive definite, then $\hat\beta$ is unique (not clear in general).

14.1 Model Selection

Setting: Suppose we observe $Y \in \mathbb{R}^n$ of the form:

$$Y = \mu + \varepsilon$$

with unknown $\mu \in \mathbb{R}^n, \varepsilon \sim N(0, \sigma^2 I_n)$, σ^2 unknown.

For $k = 1, \dots, K \le n$, suppose we have linear models for μ:

$$X^{(k)}\beta^{(k)}, \beta^{(k)} \in \mathbb{R}^k$$

where $X^{(k)} \in \mathbb{R}^{n\times k}$ with rank k.

Example 8. Full design matrix:

$$X^{(k)} = \text{first k columns of } X \text{ where } X \text{ is a } n \times p \text{ matrix.}$$

We can ask ourselves, what's the best model?

For example,

$$\hat\beta^{(k)} = ((X^{(k)})^T X^{(k)})^{-1} X^{(k)^T} Y = \arg\min_{\beta^{(k)}\in\mathbb{R}^k} \|Y - X^{(k)}\beta^{(k)}\|.$$

This has the least MSE:

$$\|\mu - X^{(k)}\hat{\beta}^{(k)}\|^2.$$

We calculate:

$$\begin{aligned}E\|\mu - \hat{\mu}^{(k)}\|^2_{\mathbb{R}^k} &= (\mu - E[\hat{\mu}^{(k)}])^2 + E[(\hat{\mu}^{(k)} - E[\hat{\mu}^{(k)}])^2] \\ &\quad - 2E[\langle \mu - E[\hat{\mu}^{(k)}], \hat{\mu}^{(k)} - E[\hat{\mu}^{(k)}]\rangle].\end{aligned}$$

Moreover,

$$E[\hat{\mu}^{(k)}] = E[\Pi^{(k)}Y] = E[\Pi^{(k)}(\mu + \varepsilon)] = \Pi^{(k)}\mu.$$

So then,

$$\hat{\mu}^{(k)} - E[\hat{\mu}^{(k)}] = \Pi^{(k)}\varepsilon.$$

$$\implies E\|\mu - \hat{\mu}^{(k)}\|^2_{\mathbb{R}^k} = \|(I_n - \Pi^{(k)})\mu\|^2 + E[\|\Pi^{(k)}\varepsilon\|^2].$$

Since $\varepsilon \sim N(0, \sigma^2 I_n)$,

$$E[\|\Pi^{(k)}\varepsilon\|^2] = \sigma^2 \text{trace}(\Pi^{(k)}) = \sigma^2 k.$$

So we have:

$$E\|\mu - \hat{\mu}^{(k)}\|^2 = \|(I_n - \Pi^{(k)})\mu\|^2 + k\sigma^2.$$

This equals:

$$= \|(I_n - \Pi^{(k)})\mu\|^2 + k\sigma^2 = \text{ BIAS + VARIANCE with k.}$$

How well is μ approximated by $\text{col}(X^{(k)})$?

We'd like to pick:

$$\hat{k} = \arg\min_{k=1,\ldots,K} \|(I_n - \Pi^{(k)})\mu\| + k\sigma^2.$$

But $\|(I_n - \Pi^{(k)})\mu\|$ is unknown. To estimate the first term, we consider RSS (Residual Sum of Squares):

$$
\begin{aligned}
E\|Y - X^{(k)}\hat{\beta}^{(k)}\|^2 &= E\|(I_n - \Pi^{(k)})(\mu + \varepsilon)\|^2 \\
&= E\|(I_n - \Pi^{(k)})\mu\|^2 + E\|(I_n - \Pi^{(k)})\varepsilon\|^2 \\
&\quad + E[\langle (I_n - \Pi^{(k)})\mu, (I_n - \Pi^{(k)})\varepsilon\rangle] \\
&= \|(I_n - \Pi^{(k)})\mu\|^2 + \sigma^2(n-k),
\end{aligned}
$$

since the cross term vanishes (due to zero mean of ε).

This implies:

$$\|Y - X^{(k)}\hat{\beta}^{(k)}\|^2 - \sigma^2(n-k) + \sigma^2 k$$

is an unbiased risk estimator for the risk $E\|\mu - \hat{\mu}^{(k)}\|^2$.

Method (Mallow's C_p):

Pick:

$$\hat{k} = \arg\min_{k=1,\ldots,K} \|Y - X^{(k)}\hat{\beta}^{(k)}\|^2 + 2\sigma^2 k.$$

Next time, we will generalize this idea to Akaike's Information Criterion (AIC).