1 Basic Statistical Concepts

Consider a poll with two answers, A and B, regarding political parties. Let:

- N: total number of voters,
- M: number of voters supporting A,
- n: size of the poll,
- $X_1, X_2, \ldots, X_n$: responses,
- Each $X_i \in \{0, 1\}$ if $X_i = 1$ supports A.

Additionally, assume:

- We select n individuals from N at random and record their truthful reply,
- Every person asked replies (no selection bias),
- People can be asked repeatedly.

The aim of the poll is to estimate the fraction of party A supporters, say θ.

Definition 1 (Estimator)**.** *An intuitive estimator is:*

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

This estimator will be analyzed in the following sections to determine whether it is unbiased, consistent, and optimal.

2 Statistical Models

Let $(X, \mathcal{F})$ be a measurable space, i.e., a set X with a sigma-algebra $\mathcal{F}$, in which our statistical observations take values.

Definition 2 (Statistical Model)**.** *Let $(X, \mathcal{F})$ be some sample space. We call the parameter space Θ. A statistical model is a family of probability measures $\{P_\theta\}_{\theta \in \Theta}$.*

Remark 1. *Often $(X, \mathcal{F})$ is a product space. For example, if $X_i \in \{0, 1\}$, each P_θ is a product distribution, i.e., $X_1, X_2, \ldots, X_n$ are independent and identically distributed (iid). Then we say $\{P_\theta : \theta \in \Theta\}$ is an iid statistical model.*

Remark 2. *If every person could only be asked once, we would have P_θ as a hypergeometric distribution, which converges to the Bernoulli model as $N, M \to \infty$.*

3 Parameter Estimation

Assume $(\Omega, \mathcal{F}, P_\theta)$ is the setting of parametric statistics. Assume Θ is measurable.

Definition 3 (Estimator)**.** *An estimator for θ is any measurable function $\hat{\theta} : X \to \Theta$, i.e., any function that, based on some data X, outputs a guess $\hat{\theta}(X)$ for θ.*

4 Unbiased and Consistent Estimators

4.1 Unbiased Estimator

Definition 4 (Unbiased Estimator)**.** *Let $(\Omega, \mathcal{F}, P_\theta)$ be a measurable space. An estimator $\hat{\theta}$ is called unbiased if:*

$$\mathbb{E}[\hat{\theta}] = \theta \quad \forall \theta \in \Theta$$

where $\mathbb{E}_{P_\theta}$ denotes expectation under the law P_θ. In more explicit terms, unbiasedness means no systematic error.

Proof. For the Bernoulli model, we compute:

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

Thus, $\hat{\theta}_n$ is an unbiased estimator of θ. □

4.2 Consistent Estimator

Definition 5 (Consistent Estimator)**.** *Let $\{P_{\theta,n} : n \geq 1\}$ be a sequence of statistical models on the same parameter space. Let $\hat{\theta}_n$ be a sequence of estimators. The sequence $\hat{\theta}_n$ is called consistent if for every $\theta \in \Theta$:*

$$\hat{\theta}_n \to \theta \quad \textit{in probability as } n \to \infty$$

or equivalently:

$$P_\theta\left(\lim_{n\to\infty} \hat{\theta}_n = \theta\right) = 1$$

Proof. For the Bernoulli model:

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

We know $\mathbb{E}[\hat{\theta}_n] = \theta$ and $\mathrm{Var}(\hat{\theta}_n) = \frac{\theta(1-\theta)}{n}$. Using Chebyshev's inequality, for any $\epsilon > 0$:

$$P\left(|\hat{\theta}_n - \theta| > \epsilon\right) \leq \frac{\mathrm{Var}(\hat{\theta}_n)}{\epsilon^2} = \frac{\theta(1-\theta)}{n\epsilon^2}$$

As $n \to \infty$, this probability tends to 0, proving that $\hat{\theta}_n$ is consistent. □

5 Maximum Likelihood Estimation (MLE)

Definition 6 (Maximum Likelihood Estimator)**.** *The maximum likelihood estimator (MLE) is the parameter that maximizes the likelihood function:*

$$L(\theta) = \prod_{i=1}^{n} P_\theta(X_i)$$

5.1 Proof: MLE for Bernoulli Model

Proof. For the Bernoulli model, $P_\theta(X_i) = \theta^{X_i}(1-\theta)^{1-X_i}$, so the likelihood function is:

$$L(\theta) = \prod_{i=1}^{n} \theta^{X_i}(1-\theta)^{1-X_i} = \theta^{\sum X_i}(1-\theta)^{n-\sum X_i}$$

Taking the logarithm:

$$\log L(\theta) = \sum X_i \log\theta + (n - \sum X_i)\log(1-\theta)$$

Setting the derivative with respect to θ equal to 0 gives:

$$\frac{d}{d\theta}\log L(\theta) = \frac{\sum X_i}{\theta} - \frac{n - \sum X_i}{1-\theta} = 0$$

Solving for θ, we get:

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

which is the MLE. □

6 Bayesian Methods

Definition 7 (Posterior Distribution in Bayesian Inference)**.** *In Bayesian statistics, a key element is the prior distribution, denoted by $\pi(\theta)$, which reflects our beliefs about the parameter θ before observing data. The posterior distribution is given by:*

$$\pi(\theta|X) \propto P_\theta(X)\pi(\theta)$$

6.1 Example: Posterior for Bernoulli Model

Example 1. *Suppose we have a Beta prior for θ, $\pi(\theta) \sim Beta(\alpha, \beta)$, and observe $X_1, \ldots, X_n$ as Bernoulli trials. The likelihood is:*

$$P(X|\theta) = \theta^{\sum X_i}(1-\theta)^{n-\sum X_i}$$

The posterior is proportional to the product of the prior and likelihood:

$$\pi(\theta|X) \propto \theta^{\sum X_i + \alpha - 1}(1-\theta)^{n-\sum X_i+\beta-1}$$

Thus, $\pi(\theta|X) \sim Beta(\sum X_i + \alpha, n - \sum X_i + \beta)$.

Notes on Bayes and Posterior

Posterior = prior × likelihood
Normalizing Constant

$$\int \text{Posterior}\,\text{dx} = 1$$

So,

$$\int \text{Posterior}\,\text{dx} = 1$$

Prior → Posterior via Bayes.

Let $\mathcal{F}_0$ be a σ-algebra on Ω and suppose $(\Omega, \mathcal{F}_0, P_\theta)$ is a dominated statistical model with densities $p(x|\theta)$. Assume

$$x, \theta \in \Omega \quad \Rightarrow \quad p(x|\theta)$$

is jointly measurable with respect to $\mathcal{F}_0 \times \mathcal{F}_1$.

Let π be a prior distribution on Ω with density $\pi(\theta)$ with respect to measure ν. Define posterior density

$$\pi(\theta|x) = \frac{p(x|\theta)\pi(\theta)}{\int p(x|\theta)\pi(\theta)\,d\theta}$$

The corresponding probability measure is called the **posterior distribution**.

Think of $p(x|\theta)$ as a Lebesgue measure. Let ν be a Lebesgue density.

Exception: If $\Omega = \{0, 1\}$, then we take ν to be the counting measure.

From the posterior, we can derive several estimators. For example, $E[\theta|X = x]$ is convex:

$$\int \theta p(x|\theta)\,d\theta = E[\theta|X = x]$$

Example: Binomial model $X|\theta \sim \text{Binomial}(n, \theta)$ with prior $\theta \sim \text{Unif}(0, 1)$.

For a uniform prior, we know the MAP and MLE.

Posterior mean:

$$\theta_{\text{MAP}} = \frac{k+1}{n+2}$$

In the case of coin flips, $X \sim \text{Binomial}(n, \theta)$, where k is the number of heads, we conclude $\theta|X \sim \text{Beta}(k+1, n-k+1)$.

$$\theta|X \sim \text{Beta}(k+1, n-k+1)$$

Conjugate Bayes Models: Let $P_\theta \in \mathcal{P}$ be a statistical model. Then some family of priors is called **conjugate** if

$$P_\theta \in \mathcal{P} \Rightarrow \theta|X \in \mathcal{P}$$

for all $X \in \mathcal{X}$, where $\mathcal{X}$ is the sample space.

$$\theta|X \sim \text{Beta}(a,b), \quad X \sim \text{Bernoulli}(p)$$

Loss Functions and Risk

Loss Function: A function $L : \Theta \times \mathcal{X} \to [0, \infty)$ is a basis function if for every $\theta \in \Theta$, $L(\theta, \cdot)$ is measurable.

Given an estimator δ, the expected loss is

$$R(\theta, \delta) = E_\theta[L(\theta, \delta)]$$

Mean Squared Error (MSE):

$$L(x, y) = (x-y)^2 \Rightarrow R(\theta, \delta) = E_\theta[(\delta - \theta)^2]$$

Bias-Variance Decomposition:

$$L(x, y) = (x-y)^2$$

Proof: Let $\delta(x) = E[\theta|X = x]$.

$$R(\theta, \delta) = E_\theta[(\delta(X) - \theta)^2]$$

Bias-variance decomposition:

$$E[(\delta(X) - \theta)^2] = \text{Var}(\delta(X)) + (\text{Bias})^2$$

Minimax and Bayes Risk

Minimax Risk: Given an estimator δ in a model $P_\theta \in \mathcal{P}$, the maximal risk of it is

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

The minimax of a model P_θ is given as $\inf_\delta \sup_\theta R(\theta, \delta)$, where the inf is over all estimators.

An estimator is called minimax if

$$\sup_\theta R(\theta, \delta) = \inf_\delta \sup_\theta R(\theta, \delta)$$

Bayes Risk: Given an estimator δ and prior π on Θ, the Bayes risk of δ is defined as

$$R_\pi(\delta) = \int R(\theta, \delta) \, d\pi(\theta)$$

The posterior risk of an estimator $\delta(X)$ is defined by

$$R(\delta|X = x) = E[L(\theta, \delta(X))|X = x]$$

Suppose δ^* is an estimator that minimizes the posterior risk, $\delta^*(x) = E[\theta|X = x]$. Then it also minimizes the Bayes risk. ec If $L(x, y) = (x-y)^2$, the Bayes optimal estimator $\delta(x)$ is the posterior mean.

We want to construct $C(x)$ s.t. $P_\theta(\theta \in C(x)) \geq 1 - \alpha$, $\forall \theta \in [0, 1]$

$$x^{(1)} \qquad (\quad) \quad C(x^{(1)})$$

$$x^{(k)} \qquad (\quad) \quad C(x^{(k)})$$
$$\theta \to \quad \to \quad \to \quad \text{contains true param } 3/4 \text{ times}$$

Example cont.:

Best guess: $C(x) = \left[\frac{\bar{X}_n - a}{n}, \frac{\bar{X}_n + b}{n}\right]$

$$P_\theta^n(\theta \in C(x)) = P_\theta^n\left(\frac{\bar{X}_n}{n} - \theta \in [-b, a]\right)$$
$$= F_\theta^n(a) - F_\theta^n(-b) + \rho_n$$

where $F_\theta^n : \mathbb{R} \to [0,1], F_\theta^n(t) = P_\theta^n\left(\frac{\bar{X}_n - \theta}{n} \le t\right)$ is the CDF of $\frac{\bar{X}_n - \theta}{n}$ under P_θ and $\rho_n = P_\theta^n\left(\frac{\bar{X}_n}{n} - \theta = -b\right)$.

How to choose a and b:

$$\text{CDF} \quad \text{CDF} \quad \leftarrow \quad -b \quad a \to t$$

We'd like to choose $a = (F_\theta^n)^{-1}\left(1 - \frac{\alpha}{2}\right)$ and $b = (F_\theta^n)^{-1}\left(\frac{\alpha}{2}\right)$, where

$$(F_\theta^n)^{-1}(p) := \inf\{t \in \mathbb{R} : F_\theta^n(t) \ge x\} \quad \text{(Quantile Function)}$$

Let's use a normal approximation, for $\sigma^2 = \theta(1-\theta)$:

$$\sqrt{n}\left(\frac{\bar{X}_n}{n} - \theta\right) = \frac{1}{\sqrt{n}} \sum_{k=1}^n \frac{X_k - \theta}{\sigma} \xrightarrow{d} \mathcal{N}(0,1) \quad \text{[CLT]}$$

$X_k \sim \text{Ber}(\theta)$

Then it follows that

$$F_\theta^n(a_n) = P_\theta^n\left(\frac{\bar{X}_n}{n} - \theta \le a_n\right)$$
$$= P_\theta^n\left(\frac{\sqrt{n}}{\sigma}\left(\frac{\bar{X}_n - \theta}{n}\right) \le \sqrt{n} a_n\right)$$
$$= \Phi\left(\frac{\sqrt{n}}{\sigma} a_n\right),$$

where the convergence is valid if $a_n := \text{const.} \frac{1}{\sqrt{n}}$.

Now, let us choose

$$a := \frac{\sigma}{\sqrt{n}} z_{1-\frac{\alpha}{2}}$$

where $z_{1-\frac{\alpha}{2}} = \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)$ is the $1 - \frac{\alpha}{2}$ quantile of $\mathcal{N}(0,1)$ and $b = a$. Then

$$C(x) = \left[\frac{\bar{X}_n}{n} - \frac{\sigma}{\sqrt{n}} z_{1-\frac{\alpha}{2}}, \frac{\bar{X}_n}{n} + \frac{\sigma}{\sqrt{n}} z_{1-\frac{\alpha}{2}}\right]$$

It follows

$$P_\theta^n(\theta \in C(x)) = F_\theta^n(a_n) - F_\theta^n(b) + \rho_n = 1 - \frac{\alpha}{2} + o(1) + o(1)$$
$$= 1 - \alpha + o(1) \text{ as } n \to \infty$$

$$\Rightarrow \text{ Asymptotically valid confidence set}$$

One more problem: σ depends on θ

- Upper bound: $\sup_{\theta \in [0,1]} \theta(1-\theta) = \frac{1}{4}$ (maximized at $\theta = \frac{1}{2}$)
- Empirical Variance: $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \frac{1}{n}\sum_{i=1}^n X_i)^2$

$$\frac{\hat{\sigma}^2}{\sigma^2} \xrightarrow{P_\theta} 1$$

Slutsky's Theorem:

$$X_n \xrightarrow{d} X, \quad Y_n \xrightarrow{d} \text{const.} \Rightarrow X_n Y_n \xrightarrow{d} CX$$

Exercise: Use this to deduce that $a_n = \frac{\hat{\sigma}}{\sqrt{n}} z_{1-\frac{\alpha}{2}}$ is also valid

Remark:

Hypothesis Testing

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

- A statistical test is a measurable function of the data $\varphi : (\mathcal{X}, \mathcal{F}) \to [0,1]$
- If $\forall x \in \mathcal{X}, \varphi(x) \in \{0,1\}$, then φ is a non-randomized test
- Else φ is randomized

Definitions:

- $H_0 : \theta \in \Theta_0$ is called the null hypothesis
- $H_1 : \theta \in \Theta_1$ is called the alternative hypothesis
- The map $\theta \to \beta_\varphi(\theta) = P_\theta[\varphi = 1]$ is called the power function of a test φ

1 $\beta_\varphi(\theta)$ 0 Θ_0 Θ_1 Θ

- For $\theta \in \Theta_0$, $\beta_\varphi(\theta)$ is the type-I-error under θ [Wrongly rejecting the null]
- For $\theta \in \Theta_1$, $1 - \beta_\varphi(\theta)$ is the type-II-error

Note:

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

Definition: [Level]

$\varphi : \mathcal{X} \to [0,1]$ has level $\alpha \in [0,1]$ if

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

Definition: [Uniformly most powerful test]

Given a level $\alpha \in (0,1)$, $\varphi : \mathcal{X} \to [0,1]$ is called UMP if for every other test φ' of level α and all $\theta \in \Theta_1$,

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

1 α 0 $\beta_\varphi(\theta)$ $\beta_{\varphi'}(\theta)$ Θ_0 Θ_1

Remark:

In general, it is very hard to find UMP tests. But: for simple hypotheses, i.e. $\Theta_0 = \{\theta_0\}, \Theta_1 = \{\theta_1\}$, it is possible. Here, likelihood ratio tests are UMP.

Theorem: [Neyman-Pearson Lemma]

Let $\Theta_0 = \{\theta_0\}, \Theta_1 = \{\theta_1\}$ be simple:

1. **Existence:** There exists a test φ and a constant $k \in [0, \infty)$, s.t. $P_{\theta_0}(\varphi = 1) = \alpha$, of the form

$$\varphi(x) = \begin{cases} 1, & \text{if } \frac{p_{\theta_1}(x)}{p_{\theta_0}(x)} > k \\ 0, & \text{if } \frac{p_{\theta_1}(x)}{p_{\theta_0}(x)} < k \end{cases} \qquad (*)$$

   Here $p_{\theta_1}, p_{\theta_0}$ are densities w.r.t. some dominated measure μ, e.g. $\mu = p_{\theta_0} + p_{\theta_1}$. Finite Θ implies measure is always dominated (likelihood always exists).

2. **Sufficiency:** If φ satisfies $P_{\theta_0}(\varphi = 1) = \alpha$ and $(*)$ 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 $(*)$, and it also satisfies $P_{\theta_0}(\varphi_k = 1) = \alpha$, or else it must satisfy $P_{\theta_1}(\varphi_k = 1) = 1$.

Proof:

1. Define $r(x) = \frac{p_{\theta_1}(x)}{p_{\theta_0}(x)} \in [0, \infty) \cup \{\pm\infty\}$. Let F_0 be the CDF of $r(x)$ under P_{θ_0}.

$$F_0(t) = P_{\theta_0}(r(x) \leq t)$$

   Then define also $\alpha(t) = 1 - F_0(t) = P_{\theta_0}(r(x) > t)$

   - α is right-continuous:

$$\lim_{\epsilon \to 0} \alpha(t + \epsilon) = \lim_{\epsilon \to 0} P_{\theta_0}(r(x) > t + \epsilon) = P_{\theta_0}(r(x) > t) = \alpha(t)$$

   - α is non-increasing
   - α has left limits

$$\lim_{\epsilon \to 0} \alpha(t - \epsilon) = P_{\theta_0}(r(x) > t - \epsilon) = \alpha(t^-)$$

   α **is cadlag:**

   - Continuous from the right
   - Limit from the left

$$\text{There exists some } k \in [0, \infty) \text{ s.t. } \alpha \leq \alpha(k^-) \quad \text{and} \quad \alpha \geq \alpha(k)$$

   We define our test

$$\varphi(x) = \begin{cases} 1 & \text{if } r(x) > k \\ \gamma & \text{if } r(x) = k \quad \text{[reject null w.p. } \gamma] \\ 0 & \text{if } r(x) < k \end{cases}$$

   We set

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

   The level of φ is

$$\begin{aligned} E_{\theta_0}[\varphi(x)] &= P_{\theta_0}(\varphi(x) = 1) \\ &= P_{\theta_0}(r(x) > k) + P_{\theta_0}(r(x) = k) \cdot \gamma \\ &= \alpha(k) + \left[\alpha(k^-) - \alpha(k)\right] \cdot \frac{\alpha - \alpha(k)}{\alpha(k^-) - \alpha(k)} = \alpha \end{aligned}$$

   (randomizing the test)

Lecture 6

Neyman-Pearson

Power of a test:

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

Likelihood ratio test:

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

LR test

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

for some $k \in [0, \infty)$, $\gamma \in [0, 1]$.

Note: LR tests are UMP for simple hypothesis testing:

- Given some α, if LR satisfies $E_{\theta_0}[\varphi] = \alpha$, it represents a Type I error.
- φ minimizes the Type II error
$$E_{\theta_1}[\varphi] \geq E_{\theta_1}[\varphi'] \quad \forall \varphi'$$

Cont. of proof (part of UMP)

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

Goal: $E_{\theta_1}[\varphi] \geq E_{\theta_1}[\varphi']$. 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: $p \geq 0$.

Observe:

- If $p_{\theta_1}(x) - kp_{\theta_0}(x) > 0 \Rightarrow \frac{p_{\theta_1}(x)}{p_{\theta_0}(x)} > k \Rightarrow \varphi(x) = 1$.
- If $p_{\theta_1}(x) - kp_{\theta_0}(x) < 0 \Rightarrow \varphi(x) = 0$.
- If $p_{\theta_1}(x) - kp_{\theta_0}(x) = 0 \Rightarrow$ integrand $= 0$.

$$\Rightarrow p = 0$$

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

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

Part (3) UMP $\Rightarrow$ (LR): Take φ^* a UMP test, $E_{\theta_0}[\varphi^*] = \alpha$, and let φ be the LR test with $E_{\theta_0}[\varphi] = \alpha$ with (*).

Goal: $\varphi = \varphi^*$ a.e. except on $\{r(x) = k\}$.

Define

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

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

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

Like before, we have

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

Thus if $\mu(\tilde{x}) > 0$,

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

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

But also

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

$$\Rightarrow \text{Cannot be } \varphi^* \text{ is UMP.}$$

Example (Gaussian Location Model)

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

$$H_0 : \mu = \mu_0, \quad H_1 : \mu = \mu_1, \quad \mu_0 < \mu_1$$

Then:

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

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

To determine K_α:

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

$$\begin{aligned}
\Rightarrow \mathbb{L} &= P_{H_0}\left(\bar{X}_n \geq K_\alpha\right) = 1 - P_{H_0}\left(\bar{X}_n < K_\alpha\right) \\
&= 1 - \Phi\left(\frac{\sqrt{n}}{\sigma}(K_\alpha - \mu_0)\right) \quad (\text{CDF for } \mathcal{N}(0,1))
\end{aligned}$$

$$\Rightarrow \text{solving for } K_\alpha \text{ gives } K_\alpha = \mu_0 + \frac{\sigma}{\sqrt{n}}\Phi^{-1}(1-\alpha),$$

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

Corollary

Consider simple hypothesis testing. Let φ be UMP, for level α. Then,

$$\alpha = E_{H_0}[\varphi_0] = E_{\theta_0}[\varphi_0] \leq E_{\theta_1}[\varphi]$$

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

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

Also since $\varphi_0 \in \{\varphi, \beta\}$ we conclude that $p_{\theta_1} = Kp_{\theta_0}$ a.s.

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

Correspondence theorem

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

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

$$\text{If } \Pr_\theta(\phi_\theta = 1) = \alpha$$

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

(i) Let $C = C(X)$ be a level-α confidence set, then
$$\phi_{\theta_0}(x) = 1\{\theta_0 \notin C(x)\}$$
is a level-α test of $\theta = \theta_0$ vs. $\theta \neq \theta_0$.

(ii) Suppose $\{\phi_{\theta_0} : \theta_0 \in \Theta\}$ 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) $\Pr_{\theta_0}(\phi_{\theta_0} = 1) = \Pr_{\theta_0}(\theta_0 \notin C(X)) = \alpha$

(ii) $\Pr_\theta(\theta \notin C(X)) = \Pr_\theta(\theta \notin \{\tilde{\theta} \in \Theta : \phi_{\tilde{\theta}}(X) = 0\}) = \Pr_\theta(\phi_\theta(X) = 1) \leq \alpha$

UMPT Tests in Models with Monotone Likelihoods
Proposition: Let $\Theta \subseteq \mathbb{R}$. Consider testing $H_0 : \theta \leq \theta_0$ vs. $H_1 : \theta > \theta_0$, for some $\theta_0 \in \mathbb{R}$. Assume there exists some test statistic $T : X \to \mathbb{R}$ and a function $h : \mathbb{R} \times \Theta \times \Theta$ such that
$$\frac{P_\theta(X)}{P_{\tilde{\theta}}(X)} = h(T(X), \theta, \tilde{\theta})$$
and for all $\theta \geq \tilde{\theta}$, $t \mapsto h(t, \theta, \tilde{\theta})$ is monotone increasing.

The simplest model for the relationship between Y_i and X_i assumes a linear relationship:

$$Y_i = aX_i + b + \varepsilon_i$$

for $i = 1, \ldots, n$, where ε_i is centered, i.e., $E(\varepsilon) = 0$ and $\mathrm{Var}(\varepsilon) = \sigma^2$. Suppose $\varepsilon \sim N(0, \sigma^2)$ with σ known.

The statistical model is given by

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

The likelihood within the statistical model is

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

The MLE satisfies the optimization problem

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

Provided that $x_i \neq x_j$ for $i \neq j$, the least squares problem has a solution with minimum given by (Gauss, 1801):

$$(\hat{a}, \hat{b}) = \left(\frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}, \bar{y} - \hat{a}\bar{x}\right)$$

Definition 8 (Linear Model). *A random vector $Y = (Y_1, \ldots, Y_n)^T \in \mathbb{R}^n$ 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. *A linear model is called regular if*
   (a) *$p \leq n$ (parameter size is smaller than sample size),*
   (b) *X has full rank. $rank(X) = p \leq n$ (design with full rank)*
   (c) *$E(\varepsilon) = 0$ (noise is controlled)*
   (d) *The covariance matrix is postive definate, $\Sigma = (Cov(\varepsilon_i, \varepsilon_j))_{i,j\in[n]}$*
2. *A linear model is called ordinary if $\Sigma = \sigma^2 E_n$ (and is usually the noise is Gaussian)*

Remark 3. 1. *There are several synonyms*
   (a) *Y a dependent variable, responce, regressand*
   (b) *X, a independent variable, predictor, design matrix, regressor*
   (c) *ε Error, perturbation, reression function*

2. *The matrix Σ is symmetric and diagonalizable, i.e. $\Sigma = UDU^T$ for some diagonal matrix, $D = diag(\lambda_1, \ldots, \lambda_n) \in \mathbb{R}^{n\times n}$*
3. *Positive semi-definate, i.e. $\lambda_i \geq 0$*

$$\langle \Sigma u, u\rangle = \langle E[(\varepsilon - E[\varepsilon])(\varepsilon - E[\varepsilon])^T]u, u\rangle$$

$$= E[(\varepsilon - E[\varepsilon])^2] \geq 0, u \in \mathbb{R}^n$$

   item If Σ is positve definate ($\lambda_i > 0$) for $i = 1, \ldots, n$, then there exists the inverse $\Sigma^{-1} = UD^{-1}U^T$ and $\Sigma^{-1/2} = UD^{-1/2}U^T$.

4. *If X is not deterministic, we speak of random design.*

In the regular linear model, $\hat{\beta}$ is called weighted least squares estimate, (LSE). if

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

where $X_\Sigma = \Sigma^{-1/2}X$. $X_\Sigma\hat{\beta}$ is the point within the subspace,

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

with the smallest distance to the vector $\Sigma^{-1/2}Y$. Thus, $X_\Sigma\hat{\beta} = \Pi_U(\Sigma^{-1/2}Y)$ where Π_U is the orthogonal projection onto U. $\Pi_U u = u$ for all $u \in U$ $\langle \Pi_U v - v, u\rangle = 0$ for all $u \in U$ and $r \in \mathbb{R}^n$. Provided that $(X_\Sigma^T X_\Sigma)^{-1}$ exists, we can confirm by direct computation that the projection satisfies

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

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

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

By symmetry,

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

for all $u \in U$.

Lemma 1. *Representation for the LSE Consider a regular linear model, then the LSE exists uniquely, and is given by*

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

Proof. $\ker(X_\Sigma^T X_\Sigma)$ is invertible. Suppose that $X_\Sigma^T X_\Sigma v = 0$ ($v \in \ker(X_\Sigma^T X_\Sigma)$)

$$0 = v^T X_\Sigma^T X_\Sigma v = (X_\Sigma^T v)^T X_\Sigma v = \langle X_\Sigma v, X_\Sigma v\rangle = ||X_\Sigma v||^2 = ||\Sigma^{-1/2} X v||^2 \implies ||Xv||^2 = 0 \implies v = 0$$

So then

$$\begin{aligned} X_\Sigma \hat{\beta} &= \Pi_u \Sigma^{-1/2} Y = X_\Sigma (X_\Sigma^T X_\Sigma)^{-1} X_\Sigma^T \Sigma^{-1/2} Y \\ X_\Sigma^T X_\Sigma \hat{\beta} &= X_\Sigma^T X_\Sigma (X_\Sigma^T X_\Sigma)^{-1} X_\Sigma^T \Sigma^{-1/2} Y \\ \implies \hat{\beta} &= (X_\Sigma^T X_\Sigma)^{-1} X_\Sigma^T \Sigma^{-1/2} Y \end{aligned}$$

□

Remark 4. *1. If $p > n$, then $(X_\Sigma^T X_\Sigma)^{-1}$ does not exist and the LSE is not unique.*

$$\left\{\beta \cdot ||\Sigma^{-1/2} Y - X_\Sigma \beta||^2 = 0\right\}$$

is a $p - n$ dim subspace and each solution interpolates the data

Theorem 1. *Optimality of the LSE, Gauss-Markov Theorem Consider an ordinary linear model for $\sigma > 0$, then*

1. *The least squares estimator $\hat{\beta} = (X^T X)^{-1} X^T Y$ is linear and the unbiased parameter for the parameter β.*
2. *For the desired parameter $\alpha = \langle \beta, v\rangle$ for $v \in \mathbb{R}$, the estimator $\hat{\alpha} = \langle \hat{\beta}, v\rangle$ is the best linear unbiased estimator (BLUE), meaning that $\hat{\alpha}$ has the optimal value within the class of linear unbiased estimators for α*
3. *$\hat{\sigma}^2 = \frac{||Y - X\hat{\beta}||^2}{n-p}$ is an unbiased estimator of σ^2*

Proof.

$$\hat{\beta}(y + \tilde{y}) = \hat{\beta}(y) + \hat{\beta}(\tilde{y}) \text{for } y, \tilde{y} \in \mathbb{R}^n$$

$$\begin{aligned} E[\hat{\beta}] &= (X^T X)^{-1} X^T E[Y] && (1) \\ &= (X^T X)^{-1} X^T E[X\beta + \varepsilon] && (2) \\ &= (X^T X)^{-1} (X^T X) \beta && (3) \\ &= \beta && (4) \end{aligned}$$

Suppose that $\tilde{\alpha}$ is some other linear unbiased estimator of α. Since the estimator is linear, there exists some element w such that $\tilde{\alpha} = \langle y, w \rangle$

$$\langle \beta, v \rangle = \alpha = E[\tilde{\alpha}] = E[\langle y, w \rangle] = \langle X\beta, w \rangle = \langle \beta, X^T w \rangle$$

This implies that $v = X^T w$, therefore we have,

$$\begin{aligned} \text{Var} &= \text{Var}(\langle x\beta, w \rangle + \langle \varepsilon, w \rangle) && (5)\\ &= \text{Var}(\langle \varepsilon, w \rangle) + E\left[\left(\sum_{i=1}^{n} \varepsilon w\right)^2\right] && (6)\\ &= \sigma^2 \sum_{i=1}^{p} w_i^2 = \sigma^2 ||w||^2 && (7) \end{aligned}$$

$$\begin{aligned} \text{Var}(\hat{\alpha}) &= E[\langle \hat{\beta} - \beta, v \rangle^2] && (8)\\ &= E[\langle (X^T X)^{-1} X^T \beta + (X^T X)^{-1} X^T \varepsilon - \beta, v \rangle^2] && (9)\\ &= E[\langle (X^T X)^{-1} X^T \varepsilon, v \rangle^2] && (10)\\ &= \sigma^2 ||X(X^T X)^{-1} v||^2 = \sigma^2 ||X(X^T X)^{-1} X^T w||^2 && (11)\\ &= \sigma^2 ||\Pi_u w||^2 && (12) \end{aligned}$$

Thus, $\text{Var}(\hat{\alpha}) \leq \text{Var}\tilde{\alpha}$ □

7 Lecture 8

Recall linear model

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

where $\text{cov}(\varepsilon) = \Sigma$.

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

$X\hat{\beta} =$ Projection of $\Sigma^{-1/2} Y$ onto span $\{X_{\varepsilon,1}, \ldots, X_{\varepsilon,p}\}$

Theorem 2 (Gauss-Markov)**.** *1. $\hat{\beta}_{OLS}$ is the best linear unbiased est (BLUE)*

2. $\alpha_i = \langle \beta, v \rangle$ is BLUE.

3. $\hat{\sigma}^2 = \frac{||Y - X\hat{\beta}||^2}{n-p}$ is unbiased est for $\sigma^2 > 0$

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}^T + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$ Where our data is $(Y_i, X_i)_{i=1}^n \in (\mathbb{R} \times \mathbb{R}^p)^{\otimes p}$

Remark 5. *Is this an iid model? Depends!*

1. *Typically ε_i are iid.*
2. *If X_i are random then "random design".*
3. *If X_i are iid, then linear model is iid model.*
4. *If X_i are deterministic, then not iid model.*

$\beta \mapsto ||Y - X\hat{\beta}||$.

Proof. This is a continuation of point 3 in our theorem above.

We already introdiced $\Pi_U = X(X^T X)^{-1} X^T$ projection onto col space U of X. Thus $I_n - \Pi_U$ is another projection operator, onto $U^\perp$ (othrogonal complement),

$$U^\perp = \{z \in \mathbb{R}^n \mid \langle z, X_k \rangle \forall k = 1, \ldots, p\}.$$

Choose a basis $e_1, \dots e_{n-p}$, orthonormal, of $U^\perp$, then

$$(I_n - \Pi_U)z = \Pi_{U^\perp} z = \sum_{n=1}^{n-p} \langle z, e_k \rangle e_k.$$

$$\begin{aligned}
||Y - X\hat{\beta}|| &= ||Y - \underbrace{X(X^TX)^{-1}X^T}_{\Pi_U} Y||^2 && (13)\\
&= ||(I_n - \Pi_n)Y||^2 && (14)\\
&= ||(I_n - \Pi_n)(X\beta + \varepsilon)||^2 && (15)\\
&= ||(I_n - \Pi_n)\varepsilon||^2 && (16)\\
&= \sum_{i=1}^{n-p} \langle \varepsilon, e_i \rangle^2 && (17)\\
& && (18)
\end{aligned}$$

Hence,

$$E[||Y - X\hat{\beta}||^2] = \sum_{i=1}^{n-p} E[\langle \varepsilon, e_i \rangle^2] = n - p \implies E[\hat{\sigma}] = n - p$$

□

Remark 6. *Recall the $N(\mu, \sigma^2)$ model, where the MLE is*

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

The unbiased estimator for σ^2 was $\frac{1}{n-1}\sum_{i=1}^n (X - i - \hat{\mu})^2$. This is related to the $n - p$ factor in point 3.

Remark 7. *1. If linearity is dropped, there exists better estimators than $\hat{\beta}_{OLS}$. For example a connstant estimator, $\hat{\beta} = \beta^*$*

2. The MSE of $\hat{\beta}_{OLS}$ is

$$E[||\hat{\beta}_{OLS} - \beta||^2] = E\left[\sum_{i=1}^p \langle \hat{\beta}_{OLS} - \beta, \underbrace{e_i}_{ONB of \mathbb{R}^n} \rangle^2\right] = \sum_{i=1}^n Var_\beta(\langle \hat{\beta}_{OLS}, e_i \rangle) = \sum_{i=1}^n \sigma^2 ||X(X^TX)^{-1}e_k||^2$$

We say X satisfies orthogonal design if

$$X^TX = nI_p$$

"The different covariants are uncorrelated.=" $(X^TX)_{ij} = \langle X_i, X_j \rangle = n\delta_{ij}$ For orthogonal design,

$$E_\beta[||\hat{\beta}_{OLS} - \beta||^2] = \frac{1}{n^2}\sigma^2 \sum_{i=1}^n \underbrace{||xe_i||^2}_{n} = \frac{\sigma^2 P}{n}.$$

and this is equal to noise level times the number of parameters, divided by the number of data points.

Theorem 3 (Bayes in Linear Models)**.** *Consider a linear model $Y = X\beta + \varepsilon$, and $\varepsilon \sim N(0, \sigma^2 I_n)$ with $\sigma > 0$ known and $\beta \sim N(m, M)$ where $m \in \mathbb{R}^p, M \in \mathbb{R}^{p \times p}$ positive semi definate. Then, the posterior $\Pi(\beta|Y_i X)$ is given by*

$$\Pi(\beta|Y, X) = N(\mu_{past}, \Sigma_{past}) \text{ for}$$

$$\mu_{past} = \sigma_{past}(\sigma^{-2}X^Ty + M^{-1}m) \quad \Sigma_{past} = (\sigma^{-2}X^TX + M^{-2})^{-1}$$

Remark 8. *Σ_{past} independent of Y. For "$M^{-2} \to 0$", then "$\mu_{past} \to \hat{beta}_{OLS}$"*

Proof.

$$L(X,Y,\beta)\pi(\beta) \propto \exp\left(-\frac{1}{2\sigma^2}||Y-X\beta|| - \frac{1}{2}(\beta-m)^T M^{-1}(\beta-m)\right)$$

We want this to be proportional to $\exp\left(-\frac{1}{2}(\beta-\mu_{\text{past}})^T\sigma_{\text{past}}^{-1}(\beta-\mu_{\text{past}})\right)$.

Now,

$$\exp\left(-\frac{1}{2}(\beta-\mu_{\text{past}})^T\sigma_{\text{past}}^{-1}(\beta-\mu_{\text{past}})\right) \propto \exp\left(-\frac{1}{\sigma^2}\beta^T X^T X\beta - \frac{1}{2}\beta^T M^{-1}\beta + \frac{1}{\sigma^2}\beta^T X^T Y + \beta^T M^{-1}m\right)$$

and this is equal to

$$\exp\left(-\frac{1}{2}\beta^T\left(\frac{1}{\sigma^2}X^T X + M^{-1}\right)\beta + \beta^T(\frac{1}{\sigma^2}X^T Y + M^{-1}m\right)$$

and this is

$$\propto \exp\left(-\frac{1}{2}(\beta-\mu_{\text{past}})^T\sigma_{\text{past}}^{-1}(\beta-\mu_{\text{past}})\right)$$

□

Corollary 1. *For $\ell = ||\cdot||^2$, the Bayes estimator is $\hat{\beta}_\Pi = \mu_{past}$*

Proposition 1. *Consider the previous setting (from the theorem), with $m = 0$, and $M = \tau^2 I_p$ (centered, isotropic, normal prior). The, $\mu_{past} = \hat{\beta}_\Pi$ minimizes*

$$\beta \mapsto ||Y - X\beta||^2_{\mathbb{R}^n} + \underbrace{\frac{\sigma^2}{\tau^2}||\beta||^2_{\mathbb{R}^p}}_{\text{"penalty" or "regularization"}}$$

Proof. □

8 Lecture 9

Proof:

Take gradient of $\mathcal{J}(\boldsymbol{\beta})$ w.r.t. $\boldsymbol{\beta}$:

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

Set $= 0$:

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

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

Posterior mean:

$$\begin{aligned}\mu_{\text{post}} &= \boldsymbol{\Sigma}_{\text{post}}^{-1}(\boldsymbol{X}^\top\boldsymbol{Y} + \boldsymbol{M}_0^{-1}\boldsymbol{\mu}_0)\\ &= (\sigma^{-2}\boldsymbol{X}^\top\boldsymbol{X} + \tau^{-2}\boldsymbol{I}_p)^{-1}\sigma^{-2}\boldsymbol{X}^\top\boldsymbol{Y}\\ &= (\boldsymbol{X}^\top\boldsymbol{X} + \frac{\sigma^2}{\tau^2}\boldsymbol{I})^{-1}\boldsymbol{X}^\top\boldsymbol{Y}\end{aligned}$$

Remark:

$\boldsymbol{\beta}$ is defined even if $\text{rank}(\boldsymbol{X}) < p$, in particular even for $n < p$.

Definition:

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

is called a **Ridge Regression** estimator. Here, $\lambda > 0$ is called a regularization parameter. $\hat{\boldsymbol{\beta}}_{\text{ridge}}$ is always uniquely defined.

For $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, UM:

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

Estimator independent of σ^2.

Proposition:

MSE of $\hat{\boldsymbol{\beta}}_{\text{ridge}}$.

Consider a linear model with $\boldsymbol{\varepsilon} \sim N(0, \sigma^2\boldsymbol{I}_n)$, $\sigma^2 > 0$ known, and $\boldsymbol{X}^\top\boldsymbol{X} = n\boldsymbol{I}_p$ (orthonormal design).
Let $\mathcal{J} := \langle\boldsymbol{\beta}, \boldsymbol{v}\rangle$ for $\boldsymbol{v} \in \mathbb{R}^p$, and:

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

Then:

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

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

We have:

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

where $\boldsymbol{\varepsilon} \sim N(0, \sigma^2\boldsymbol{I}_n)$.

$$\begin{aligned}&= \frac{1}{1+\frac{\lambda}{n}}(\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{X}^\top\boldsymbol{\varepsilon}),\\ &= \frac{1}{1+\frac{\lambda}{n}}\boldsymbol{\beta} + \frac{1}{1+\frac{\lambda}{n}}\boldsymbol{X}^\top\boldsymbol{\varepsilon}.\end{aligned}$$

Bias-Variance Decomposition:

$$\begin{aligned}\mathbb{E}\left[(\hat{\boldsymbol{\beta}}_{\text{ridge}} - \mathcal{J})^2\right] &= (\mathbb{E}[\hat{\boldsymbol{\beta}}_{\text{ridge}}] - \mathcal{J})^2 + \operatorname{Var}(\hat{\boldsymbol{\beta}}_{\text{ridge}}).\\ &= ((1+\frac{\lambda}{n})^{-1}\langle\boldsymbol{\beta}, \boldsymbol{v}\rangle)^2 + \frac{\lambda^2}{(1+\lambda)^2}\operatorname{Var}(\boldsymbol{X}^\top\boldsymbol{\varepsilon}, \nu).\end{aligned}$$

Observe:

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

Also:

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

Corollary:

Under the same assumptions:

$$\begin{aligned}\mathbb{E}[\|\hat{\boldsymbol{\beta}}_{\text{ridge}} - \boldsymbol{\beta}\|^2] &= \mathbb{E}\left[\sum_{k=1}^p(\langle\boldsymbol{\beta}, \boldsymbol{e}_k\rangle - \beta_k)^2\right].\\ &= \frac{1}{(1+\frac{\lambda}{n})^2}\|\boldsymbol{\beta}\|^2 + \frac{p\sigma^2}{n(1+\frac{\lambda}{n})^2}.\end{aligned}$$

Remark:

For small $\|\boldsymbol{\beta}\|$, Ridge $\to$ OLS. The optimal choice of λ depends on $\|\boldsymbol{\beta}\|$.

1.7 Confidence Sets & Tests in Linear Model:

The estimators we studied are independent of σ^2, but uncertainty quantification will depend on σ^2!

Assume $\boldsymbol{\varepsilon} \sim N(0, \sigma^2 \boldsymbol{I}_n)$ throughout.

Easy Case:

For $\sigma^2 > 0$ known:

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

Indeed:

$$\operatorname{Cov}((\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{\varepsilon}) = (\boldsymbol{X}^\top \boldsymbol{X})^{-1}.$$

And for $\mathcal{J} = \langle \boldsymbol{\beta}, \boldsymbol{\nu} \rangle$,

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

Then a 95% confidence set for $\mathcal{J}$ is:

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

Notes on t- and F-distributions:

BUT: Normally, σ^2 is unknown. Replace σ by its estimator $\hat{\sigma}$. We need the t- and F-distributions.

Definitions:

Definition (t-distribution): The t-distribution with $n \geq 1$ degrees of freedom on $\mathbb{R}$ has density:

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

where C_n is the normalizing constant.

Note: For $n = 1$:

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

which corresponds to the **Cauchy distribution**.

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

Why is this useful?

Lemma: Let $X_1, \ldots, X_n, Y_1, \ldots, Y_n$ be 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:

1. The t-distribution arises when considering the "empirical mean" and "empirical variance."

2. For $n \to \infty$, $T_n \xrightarrow{d} N(0,1)$.

Proof:

(b) Observe:

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

By a change of measure ($y \mapsto y^2$ in $(0,\infty)$):

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

Since t is symmetric around 0, we obtain for all $x \in \mathbb{R}$:

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

It remains to show the claim for $F_{m,n}$.
Let:

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

Then:

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

where the density of χ_m^2 is:

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

Derivation:

Writing $W = \frac{X}{Y}$, we have:

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

Substituting $x = wy$, we get:

$$= \int_0^\infty \int_0^z 1\, f_X(wy) f_Y(y)\, y\, dw\, dy$$

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

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

Change of Variable:

Let $a = \frac{z}{z+1} y$, then:

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

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

It follows:

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

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

Change of Variable:

Let $F = \frac{X}{Y}$, given $f_F(z) = \frac{m}{n} f_{X,Y}\left(\frac{m}{n}z\right) = f_{m,n}(z)$.

9 Lecture 10

t-distribution $\cdot$ $t_n(x) \propto \left(\frac{n^2}{n} + 1\right)^{-(n+1)/2}$

F - $\cdot$ $f_{m,n}(x) \propto \frac{x^{m/2-1}}{(n+mx)^{(m+n)/2}}$

In the linear model $Y = X\beta + \varepsilon$, $\varepsilon \sim N(0, \sigma^2 I_n)$,

$$\hat{\beta} = (X^T X)^{-1} X^T Y \sim N(0, \sigma^2 A)$$

where

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

We now have

$$t_n = \frac{N(0,1)}{\chi^2(n)n} \qquad f_{m,n} = \frac{n\chi^2(m)}{m\chi^2(n)}$$

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

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 from $R\xi$*
3. *If $S \in \mathbb{R}^{n\times n}$ is another orthogonal projection, rank$(S) = s \le n$ 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 othrogonal projection, there exists an orthogonal matrix $T^T = T^{-1}$ such that

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

Then we have $T^T \sim N(0, T^T T) = N(0, I_n)$.

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

2. Let $A_1 = B\xi$, $A_2 = R\xi$, then

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

3. By (2), we know $S\xi$ and $R\xi$ are independent. By (1), $\xi^T S\xi \sim \chi^2(s)$, $\xi R\xi \sim \chi^2(r)$. The claim follows from the definition of $F(r, s)$.

□

Theorem 4. *Linear Model Confidence Sets -unknown σ^2 Assume regular linear model, $Y = X\beta + \varepsilon$, rank$(X) = p \le n$, $\varepsilon \sim N(0, \sigma^2 I_n)$. Let $\alpha \in (0, 1)$*

1. *Let $q_{F_{p,n-p},1-\alpha}$ be the $1-\alpha$ quantile of $F_{p,n-p}$ distribution. Then $C(Y, X) = \left\{\beta \in \mathbb{R}^p \mid \frac{||X(\beta - \hat{\beta}_{OLS})||^2}{p\hat{\sigma}^2} \le q_{F_{p,n-p},1-\alpha}\right\}$ is a level $1-\alpha$ confidence set.*
2. *Let $\alpha = \langle\beta, v\rangle$, for some $v \in \mathbb{R}^p$. Then a $1-\alpha$ confidence set is*

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

where $\hat{\alpha} = \langle\hat{\beta}_{OLS}, v\rangle$ and q is the $1-\alpha/2$ quantile of t_n.

Proof. 1. We know $X\hat{\beta}_{OLS} = \Pi_U Y = \Pi_U X\beta + \Pi_U \varepsilon = X\beta + \Pi_U\varepsilon$ Moreover,

$$\hat{\sigma}^2 = \frac{||X(\beta - \hat{\beta}_{OLS})||^2}{n-p} = \frac{(||I_n - \Pi_U)Y||^2}{n-p} = \frac{||\Pi_{U^\perp} Y||^2}{n-p} = frac||\Pi_{U^\perp}\varepsilon||^2 n - p$$

This implies

$$\frac{||X\beta - X\hat{\beta}_{OLS}}{p\hat{\sigma}^2} = \frac{|n-p|||\Pi_U\varepsilon||^2}{p||\Pi_{U^\perp}\varepsilon||^2} \sim \frac{(n-p)\sigma^2\chi^2(p)}{p\sigma^2\chi^2(n-p)} \sim F(p, n-p).$$

2. We know

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

And this implies

$$\frac{\alpha - \hat{\alpha}}{\sigma\sqrt{v^T(X^TX)^{-1}v}} \sim N(0,1).$$

Finally, also, as in (1), $\hat{\sigma}^2 \sim \sigma^2\chi^2(n-p)$. This implies

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

□

9.1 The $t-$ and $F-$test

Remark 9 (Method (t-test))**.** *In a regular linear model with $\varepsilon \sim N(0, \sigma I_n)$, consider $H_0 : \gamma = \gamma_0$ vs $H_1 \gamma \neq \gamma_0$ $(\gamma = \langle \beta, v\rangle)$. The two sided $t-$test is*

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

where

$$T_{\alpha_0, n-p} = \frac{\alpha_0 - \hat{\alpha}}{\hat{\sigma}\sqrt{v^T(X^TX)^{-1}v}}$$

and q is the $1 - \alpha/2$-quantile of t_{n-p}.

Remark 10 (Method (F-test))**.** *Same setting as before for t-test, $H_0 : \beta = \beta_0$ vs $H_1\beta \neq \beta_0$ since $\beta_0 \in \mathbb{R}^p$. Then the F-test is*

$$\varphi_{\beta_0}(Y, X) = \not{K}(|F_{\beta_0, n-p}(Y, X)| > q)$$

where

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

and $q = (1 - \alpha)$-quantile of $F_{p, n-p}$.

9.2 General linear hypothesis testing problems

Definition 9. *A linear hypothesis testomng pb. is of the form $H_0 : K\beta = d$ vs $H_1 : K\beta \neq d$, where $K \in \mathbb{R}^{r\times p}$ with rank$(K) = r \leq p$, $d \in \mathbb{R}^p$. In other words "r linear constrations on β"*

K os called the "contrast matrix"

Theorem 5. *Assume regular linear model, with $\varepsilon \sim N(0, \sigma^2 I_n)$, and consider $H_0 : K\beta = d$ vs. $K\beta \neq d$.*

Defin residual sum of squares as $RSS = ||Y - X\beta_{OLS}||^2$ and $RSS_{H_0} = ||Y - X\beta_{H_0}||^2$ and $\hat{\beta}_{H_0}$ over $\{\beta \colon K\beta = d\}$.

Proof. 1.

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

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

3. Define

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

under H_0.

□

10 Lecture 11

Theorem:
Assume regular LM, $\varepsilon \sim N(0, \sigma^2 I_n)$, and consider:

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

Define:

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

where β_{H_0} is the OLS estimator over $K\beta = d$:

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

1.

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

2. Under H_0:

$$RSS_{H_0} \sim \chi^2(n)$$

3. Define:

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

Under H_0:

$$F \sim F_{n-p}$$

Proof:
1. To show $K\hat{\beta}_{H_0} = d$, note that β_{H_0} is the minimizer.
Observe:

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

Second part: Let $Y \in \mathbb{R}^n$, $K\beta = d$. By Pythagoras:

$$\|Y - X\hat{\beta}\|^2 = \|Y - X\beta_{H_0}\|^2 + \|X(\beta - \beta_{H_0})\|^2$$

where:

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

This implies:

$$(K\hat{\beta} - d)^\top \left(K(X^\top X)^{-1} K^\top\right)^{-1} (K\hat{\beta} - d) = 0$$

Overall:

$$\|Y - X\beta_{H_0}\|^2 = \|Y - X\hat{\beta}\|^2 + \|X(\hat{\beta} - \beta_{H_0})\|^2 \geq 0$$

Continuation:
2) Under H_0:

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

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

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

$$\mathbb{E}[Z] = \mathbb{E}[K\hat{\beta} - d] = K\mathbb{E}[\hat{\beta}] - d = K(X^\top X)^{-1} X^\top Y - d$$

(Substitute $K\beta = d$ into the expectation)

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

Thus:

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

Finally:

$$RSS_{H_0} - RSS = \|X(\hat{\beta} - \beta_{H_0})\|^2 = Z^\top (\sigma^2 K(X^\top X)^{-1} K^\top)^{-1} Z$$

$$\sim \chi^2(p)$$

$$RSS \sim \sigma^2\chi^2(n-p), \quad RSS_{H_0} \sim \sigma^2\chi^2(n).$$

3) We know:

$$\frac{RSS_{H_0} - RSS}{\sigma^2} \sim \chi^2(p), \quad \frac{RSS}{\sigma^2} \sim \chi^2(n-p)$$

To show independence: We have:

$$RSS_{H_0} \perp Y \quad \text{while} \quad RSS_{H_0} - RSS \text{ only depends on } \hat{\beta}.$$

(Since $\hat{\beta} \propto X^\top Y$ and $T_n = 0$ by the lemma from last time, independence follows.)

ANOVA (Analysis of Variance)

Motivation: We have data from k different groups. Are the means equal?

Definition: ANOVA

We are given data:

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

Assume:

$$\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$$

ANOVA (Analysis of Variance)

Index: $i = 1, \dots, k$ is called the factor.
The model is a *factor model* with 1 categorical variable.
$n = \sum_{i=1}^k n_i$ is the total sample size.
The model is balanced (design) if $n_1 = n_2 = \dots = n_k$.

Remark: ANOVA is a linear model:

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

Hypothesis Testing:

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

Basic Idea: Compare variation within groups vs. variation across groups.

Theorem (Decomposition of RSS): Define the group means:

$$\bar{Y}_{i.} = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij}, \quad i = 1, \dots, k$$

and the overall mean:

$$\bar{Y}_{..} = \frac{1}{n} \sum_{i,j} Y_{ij}.$$

Furthermore, let:

$$SSB = \sum_{i=1}^k n_i (\bar{Y}_{i.} - \bar{Y}_{..})^2 \quad \text{(Sum of squares between groups)},$$

$$SSW = \sum_{i=1}^k \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i.})^2 \quad \text{(Sum of squares within groups)}.$$

Then:

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

where SST is the total sum of squares.

Proof:

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

$$= \sum_{i,j}(Y_{ij} - \bar{Y}_{i.})^2 + (\bar{Y}_{i.} - \bar{Y}_{..})^2 + 2(Y_{ij} - \bar{Y}_{i.})(\bar{Y}_{i.} - \bar{Y}_{..})$$

$$= SSB + SSW + C,$$

where:

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

By construction:

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

so:

$$C = 0.$$

Theorem:

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

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

11 Lecture 12

ANOVA

linear model, factor/category, F-test for equality of means, $Y_{i,j} = \mu_i + \varepsilon_{ij}$ for $i = 1, \dots, k$ and $j = 1, \dots, k$.

First a note, $X^T X = ||X||^2_{\mathbb{R}^n}$

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

1. *The OLS estimate is*

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

2.

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

3. *Under H_0: $\mu_0 = \mu_1 = \ldots = \mu_k$, $\frac{SSB}{\sigma^2} = \frac{1}{\sigma^2}\sum_i n_i(\bar{y}_{i,\cdot} - \bar{y}) \sim \chi^2(k-1)$*
4. *SSW and SSB are indepdendent and under H_0,*

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

Proof. (a) We have $\hat{\mu} = (X^TX)^{-1}X^TY$, with

$$(X^TX)^{-1} = \begin{pmatrix} \frac{1}{n_1} & \ldots & 0 \\ 0 & \ldots & 0 \\ 0 & \ldots & \frac{1}{n_k} \end{pmatrix}$$

and this implies that

$$\hat{\mu} = \begin{pmatrix} \frac{1}{n_1} & \ldots & 0 \\ 0 & \ldots & 0 \\ 0 & \ldots & \frac{1}{n_k} \end{pmatrix} \begin{pmatrix} [] & \ldots & 0 \\ 0 & \ldots & 0 \\ 0 & \ldots & [] \end{pmatrix} Y = \begin{pmatrix} \frac{1}{n_1} & \ldots & 0 \\ 0 & \ldots & 0 \\ 0 & \ldots & \frac{1}{n_k} \end{pmatrix} \begin{pmatrix} \sum_j Y_{1,j} \\ \vdots \\ \sum_j Y_{k,j} \end{pmatrix} = \begin{pmatrix} \bar{Y_{1,j}} \\ \vdots \\ \bar{Y_{k,j}} \end{pmatrix}$$

(b)

$$SSW = ||Y - X\hat{\mu}||^2_{\mathbb{R}^n} = \sum_i \sum_j (y_j - y_{i,\cdot})^2 = RSS \implies \frac{SSW}{\sigma^2} \sim \chi^2(n-k).$$

(c) We know that $SSW = RSS$ and we know that $SSW + SSB = SST = \sum_{ij}(y_{ij} - \bar{y}_{i,\cdot})^2$. We also know $\frac{RSS_{H_0} - RSS}{\sigma^2} \sim \chi^2(k-1)$ from before, it suffices to show $SST = RSS_{H_0}$,

$$RSS_{H_0} = \min_{\mu \in \mathbb{R}} ||Y - \mu||^2_{\mathbb{R}^2} = ||Y - \bar{Y}_{\cdot\cdot}{}^2_{\mathbb{R}^n} = SST.$$

(d) Follows from general lin hypotheses testing theorem, Theorem 2.2.30 in Methoden der Statistik book.

□

11.1 Exponential Families

General Model$(P_\theta : \theta \in \Theta) \supseteq$ Exp. families $\supseteq$ Linear Model

Regularity Assumptions:
Let $(P_\theta : \theta \in \Theta)$ be a statistical model

1. Dominated, there exists μ such that $P_\theta << \mu$ for all $\theta \in \Theta$
2. $\Theta \in \mathbb{R}^p$ is an open set $p \geq 1$.
3. Likelihood $p_\theta(x) > 0$ for all $\theta \in \Theta, x \in X$, in particular $\log p_\theta(x)$ is well defined.

Definition 10. *Score The score vector is* $U_\theta(x) = \nabla_\theta \log p_\theta(x) = \begin{pmatrix} \frac{\partial}{\partial\theta_1}\log p_\theta(x) \\ \vdots \\ \frac{\partial}{\partial\theta_p}\log p_\theta(x) \end{pmatrix}$ *whenever it exists*

Definition 11. *Fisher Information For $\theta \in \Theta$, the FI, whenever it exists, is* $I(\theta) = E(U_\theta(x)U_\theta(x)^T) \in \mathbb{R}^{p \times p}$

More Regularity Assumptions

1. $p_\theta(x)$ is twice differentiable, in particular, $U_\theta(x)$ is well defined.
2. $E_\theta[||U_\theta(x)||^2_{\mathbb{R}^p}] < \infty$ for all $\theta \in \Theta$, so $I(\theta)$ is well defined.
3. $$\int h(x)\nabla_\theta p_\theta(x)\mu(dx) = \nabla_\theta \int h(x)$$

   for relevant $h(x)$.

Lemma 3. *($P_\theta \in \Theta$) regular model (as above),*

1. $E_\theta(U_\theta(x)) = 0$
2. $I(\theta) = Cov(U_\theta)$

Proof.

$$E_\theta[U_\theta(x)] = \int_X \frac{\nabla_\theta p_\theta(x)}{p_\theta(x)} p_\theta(x) d\mu(x) = \int_X \nabla_\theta p_\theta(x) d\mu(x) = \nabla_\theta[\theta \mapsto 1] = 0$$

□

Definition 12. *Uniform minimum variance unbiased estimators (UMVUE) Let $p(\theta) \in \mathbb{R}$ be some quantity of interest, $T\colon X \to \mathbb{R}$ is a UMVUE if $E_\theta[T(x)] = g(\theta)$ and for all other unbiased estimators, $S\colon X \to \mathbb{R}$, $Var_\theta(T) \leq Var_\theta(S)$ for all $\theta \in \Theta$.*

Remark 11.

1. *UMUVE are the best possible among unbiased estimators.*
2. *Compare to Gauss Markov, $\hat{\beta}_{OLS}$ is UMVUE*
3. $$E_\theta[||T(x) - \rho(\theta)||^2] = Bias^2 + Var(T) = Var(T) \leq E_\theta[|S(x) - \rho(\theta)|^2]$$

Theorem 7. *Let $(P_\theta \mid \theta \in \theta)$ be regular. Let $\rho\colon \Theta \to \mathbb{R}$ continuous differentiable, then for any unbiased estimator T of S, $E_\theta[T] = S(\theta)$,*

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

Remark 12. *If $\Theta = \mathbb{R}$, $\rho(\theta) = \theta$, $Var_\theta(T) \geq I(\theta)^{-1}$*

Proof. Let us assume $\Theta \subseteq \mathbb{R}$,

$$\mathrm{Cov}_\theta(U_\theta, T) = E_\theta[U_\theta T] - E_\theta[U_\theta]E_\theta[T] = E_\theta[U_\theta T]$$

More over, by Cauchy Swartz,

$$\mathrm{Cov}_\theta(U_\theta T) \leq \mathrm{Var}_\theta(U_\theta)^{1/2}\mathrm{Var}_\theta(T)^{1/2} I(\theta)^{1/2}\mathrm{Var}_\theta(T)^{1/2}$$

But then

$$\begin{aligned} E_\theta[U_\theta T] &= \int_X \nabla_\theta \log p_\theta(x) T(x) p_\theta(x) d\mu(x) \\ &= \int_X T(x) \nabla_\theta p_\theta(x) d\mu(x) \\ &= \nabla_\theta \int \int_X T(x) p_\theta(x) d\mu(x) \\ &= E_\theta[T] && = \rho'(\theta) \end{aligned}$$

Thus, $\mathrm{Var}_\theta(T) \geq I(\theta)^{-1}\rho'(\theta)^2$

□

Another regularity condition, $I(\theta)$ is invertible.

Lecture 13

Regular Stat Model

- $\Theta \subset \mathbb{R}^p$ open
- $p_\vartheta(x) > 0$ for all $\vartheta \in \Theta,\ x \in \mathcal{X}$ and p_ϑ is continuously differentiable.

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

exists $\forall \vartheta \in \Theta$, and $I(\vartheta)$ is positive definite ($\Rightarrow I(\vartheta)^{-1}$ exists).

Interchange ∇_ϑ and $\int$.

Theorem

($p_\vartheta, \vartheta \in \Theta$ regular.) Let $g : \Theta \to \mathbb{R}$ be continuously differentiable.

Let $T : \mathcal{X} \to \mathbb{R}$ be an unbiased estimator, $\mathbb{E}_\vartheta[T] = g(\vartheta)\ \forall \vartheta \in \Theta$.

Then

$$\mathrm{Var}_\vartheta(T) \geq \left(g'(\vartheta)\right)^T I(\vartheta)^{-1} g'(\vartheta) \quad \forall \vartheta \in \Theta.$$

Cramér-Rao / Information Inequality

Score Vector

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

Fisher Information Matrix

Remarks

- If $I(\vartheta)$ is large, better estimation seems possible: "more information contained in the data."
- Another interpretation.

Derivation

Let $\Theta \subseteq \mathbb{R}$. Suppose p_ϑ is twice differentiable in ϑ:

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

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

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

Thus,

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

Theorem 8. *Let $(P_\theta, \theta \in \Theta)$ be a regular model, $\Theta \subseteq \mathbb{R}$ and let $\rho : \Theta \to \mathbb{R}$, be a continuous differentiable function, an unbiased estimator T, $E_\theta[T] = \rho(\theta)$ attains equality in the CR-bound iff and only if*

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

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

Proof. Define $v(\theta) = \rho'(\theta) I(\theta)^{-1}$, then let T as above,

$$0 \leq \mathrm{var}(T - v(\theta) U_\theta) = \mathrm{var}(T) + v(\theta)^2 E_\theta[U_\theta^2] - 2v(\theta) \underbrace{\mathrm{Cov}_\theta(T, U_\theta)}_{\rho'(\theta)}$$

$$= \mathrm{Var} - \rho'(\theta)^2 I(\theta)^{-1} = 0$$

This implies

$$T - v(\theta) U_\theta = Constant$$

Since T is unbiased we have $E_\theta[T] = \rho(\theta)$ so, $T = \rho(\theta) + v(\theta) U_\theta$ almost surely. This shows $\Longrightarrow$, $\Longleftarrow$ is a straightforward computation. $\square$

Remark 13. *1. $T(x)$ is nto always a measureable feature of x in the equation above.*

2. If T attains the CR-bound, we say that T is the Cramer-Rao coefficent

Corollary 2. *Assume previous scaling and assume $\rho(\theta) \neq 0$ for all $\theta \in \Theta$ then the likelihood can be written in the form*

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

where $n : \theta \to \mathbb{R}$, such that $n'(\theta) = \frac{I(\theta)}{\rho'(\theta)}$ and $c(x)$ and $\Psi(\theta)$ are invertible.

Proof. By the above equation, from the last theorem, we have

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

and this implies

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

and then we get

$$T(x)\int_{\theta_0}^{\theta} \frac{I(\theta)}{\rho(\theta)}dt + \Psi(\theta) = \log(p_\theta(\theta)) + constant$$

which implies

$$p_\theta(x) = \exp(constant(x)) = \exp(n(\theta)T(x) - \Psi(\theta))$$

□

Definition 13. *Exponential Families A regular model $(P_\theta : \theta \in \Theta)$ is called the k-parameter Exponential family $(k \geq 1)$ if there exists measurable functions*

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

such that

$$p_\theta(x) = \frac{dP_\theta}{d\mu}(x) = c(x)\exp(\langle n(\theta)T(x)\rangle_{\mathbb{R}^k} - \Psi(\theta))$$

for all θ, x where

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

Remark 14.

1. *Key features is the factorization of $\langle n(\theta)T(x)\rangle_{\mathbb{R}^k}$.*
2. *Exponential forms are motivated by finding general models in which CR-efficent procedures exists.*

Example 2. *Binomial $p_\theta = Bin(n, \theta)$*

$$\begin{aligned} p_\theta(k) = \binom{n}{k}\theta^k(1-\theta)^{n-k} \\ &= \binom{n}{k}\exp(k\log\theta + (n+k)\log(1-\theta)) \\ &= \underbrace{\binom{n}{k}}_{(} c(k))\exp(\underbrace{k}_{T(k)}\underbrace{\log\frac{\theta}{n-\theta}}_{n(\theta)} + \underbrace{n\log(n-k))}_{\Psi(\theta)} \end{aligned}$$

Example 3. *Normal $\theta = (\mu, \sigma^2) \in \mathbb{R} \times (0, \infty)$, $p_\theta = N(\mu, \sigma^2)$*

$$p_{\mu,\sigma^2} = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{x^2 - 2x\mu - \mu^2}{2\sigma^2}\right)$$

Take $T(x) = \binom{x^2}{x}$, $n(\theta) = \binom{-\frac{1}{2\sigma^2}}{\frac{\mu}{\sigma^2}}$ for $k = 2$, then

$$p_{\mu,\sigma^2}(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(\langle n(\theta), T(x)\rangle - \frac{\mu^2}{2\sigma^2}\right)$$

Example 4. *Poisson* $X = \{0, 1, \ldots\}$, $p_\theta = Poisson(\theta)$, $\theta > 0$,

$$p_\theta(k) = e^{-\theta}\frac{\theta^k}{k!} = \exp(-\theta + k\log\theta)\frac{1}{k!}$$

Take $n(v) = \log\theta$, $T(k) = k$, $c(k) = \frac{1}{k!}$, $\Psi(\theta) = \theta$.

Definition 14. *Natural Exponential Family Define*

$$\Xi = \left\{n \in \mathbb{R}^k \mid \int_X c(x)\exp\left(\langle n(\theta)T(x)\rangle_{\mathbb{R}^k}\right)d\mu(x) < \infty\right\}$$

A model $(P_n \mid n \in \theta)$ *is called an natural Exponential family if* $\Theta = \Xi$,

$$\frac{dP_n}{d\mu} = c(x)\exp(\langle n, T(x)\rangle_{\mathbb{R}^k} - \Psi(n))$$

for all $x \in X, n \in \Xi$.

Remark 15. *A natural Exponential family is specificed by* $c(x)$, $T(x)$, *and* Ξ.

Lemma 4. *Let* $(P_n \mid n \in \Xi)$ *be a* 1*-parameter natural Exponential form. For all* $n \in int(\Xi)$

1. $\Psi'(n) = E_n[T]$
2. $\Psi''(n) = var_n[T]$

Proof. Let $n \in \text{int}(\Xi)$, and define

$$\gamma(n) = e^{\Psi(n)} = \int_X c(x)\exp(nT(x))d\mu(x).$$

We show that γ is infinitely differentiable at n. Observe that

$$\frac{d}{dn}(c(x)\exp(nT(x))) = c(x)T(x)\exp(nT(x))$$

To use dominated convergence theorem, we want

$$\sup_t |c(x)T(x)\exp(n+t)T(x)|$$

is integrable for some $\varepsilon > 0$. But for some ε small enough, $n \neq t \in textint(\Xi)$ and

$$T(x)\exp(nT(x)) \leq C\exp((n+\varepsilon)T(x))$$

for some constant $C > 0$, using that $x \leq Ce^{\varepsilon x}$ for all ε.

Thus, by using DCT

$$\frac{d}{d\mu}\gamma(n) - \int_X c(x)T(x)\exp(nT(x))d\mu(x),$$
$$\frac{d}{dn}\Psi(n) = \frac{d}{dn}\log\gamma(n) = \frac{\gamma'(n)}{\gamma(n)} = E_n[T].$$

□

12 Lecture 14

Exponential Families: Assume $\Theta \subseteq \mathbb{R}$ open,

1. $p_\theta(x) = c(x)\exp\{n(\theta)T(x) - \Psi(\theta)\}$
2. Natural EF: $p_n(x) = c(x)\exp\{nT(x) - \Psi(n)\}$
3. Natural Parameter Space: $\Xi = \{t \mid \underbrace{\int_X c(x)\exp(tT(x))d\mu(x)}_{e^{\Psi(n)}} < \infty\}$ we need to check that Ξ is an open interval.

4. $p_\theta(x)$ satisfies the regularity assumptions.

Lemma 5. *$(P_n)_{n\in\Xi}$ a natural and regular EF, then for every $n \in \Xi = \Xi_0$,*

1. $\varphi'(n) = E_n[T]$

2. $\varphi''(n) = Var_n(T)$

Proof. 1. Done in last Lecture

2. Recall that $\alpha(n) = \int_X c(x)\exp(nT(x))d\mu(x) = e^{\Psi(n)}$ we had shown that α is C^∞ on this natural parameter space Ξ. as well as $\Psi'(n) = \frac{\alpha'(n)}{\alpha(n)} = E_n[T]$. Similarly,

$$\Psi''(n) = \frac{\alpha''(n)}{\alpha(n)} - \frac{\alpha'(n)^2}{\alpha(n)^2} = \int_X c(x)T(x)^2\exp(nT(x))d\mu(x) - E_n[T]^2 = E_n[T^2] - E_n[T]^2 = \mathrm{Var}_n(T)$$

□

Example 5. *Often $T(x) = x$, like with the Poisson, Normal, etc.*

1. $P_\theta = Bin(n,\theta)$, $T(k) = k$, $n = \log\frac{\theta}{1-\theta}$

Recall $p_\theta(k) = \binom{n}{k}\theta^k(1-\theta)^{n-k} = \binom{n}{k}\exp\left(k\log\frac{\theta}{1-\theta} + n\log(1-\theta)\right)$

$$\Psi(n) = -n\log(1-\theta) = -n\log\left(1 - \frac{e^n}{1+e^n}\right)$$
$$\theta = \frac{e^n}{1+e^n}$$
$$= -n\log\left(\frac{1}{1+e^n}\right) = n\log(1+e^n)$$

Hence,

$$\Psi'(n) = n\frac{e^n}{1+e^n} = n\theta = E_n[T] = \textit{Mean of } Bin(n,\theta)$$

$\Psi(n) = \ Var_n[T] = n\theta(1-\theta)$

2. $P_\lambda = Poisson(\lambda)$,

$$P_\lambda(k) = e^{-\lambda}\frac{\lambda^k}{k!} = \frac{1}{k!}\exp(\underbrace{k}_{T(k)}\underbrace{\log\lambda}_{n} - \underbrace{\lambda}_{\Psi(n)})$$

with $\Psi(n) = e^n$, then $\Psi' = \Psi'' = e^n$. Hence,

$$E_n[T] = e^n = \lambda,\ Var_n(T) = e^n = \lambda$$

Theorem 9. *MLEs in Internal EF and UMVUE Estimators Let $(P_\theta)_{n\in\Xi}$ be a natural 1-parameter regular EF, then*

1. If a unique MLE $\hat{n}_{MLE}$ exists then $(\Psi')^{-1}(T) = \hat{n}_{MLE}$

2. Define $\rho(n) = E_n[T]$ where T is the UMVUE for $\rho(n)$

Proof. 1.

$$\hat{n}_{MLE} = \hat{n} = \arg\max_{n\in\Xi} p_n(x)$$
$$= \arg\max_n \log p_n(x)$$
$$= \arg\max_n nT(x) - \Psi(n)$$

Therefore,

$$\frac{d}{d\mu}\left(nT(x) - \Psi(n)\right)\Bigg|_{n=\hat{n}} = 0 \implies T(x) = \Psi'(\hat{n})$$

Moreover, $\Psi''(n) = \mathrm{Var}_n(T) > 0$ results from inverting Ψ'.

2. Recall CR lower bound $\mathrm{Var}_n(S) \geq \rho'(n)^2 I'(n)$ for any unbiased estimator S. It holds that

$$\rho'(n) = \frac{d}{dn}(E_n[T]) = \Psi''[n]$$

where $I(n) = E_n[\log p_n(x)'^2] = -E_n[\log p_n(x)'']$ Then

$$(\log p_n(x))' = (nT(x) - \Psi(n))' = T(x) - \Psi'(n) = T(x) - E_n[T(x)]$$

$$E_n[(\log p_n(x))'^2] = \mathrm{Var}_n(T) = \Psi''(n).$$

Therefore the CR-bound $\mathrm{Var}_n(S) \geq \Psi(n) = \mathrm{Var}_n(T)$

□

13 Generalized Linear Models (GLMs)

$$\underbrace{\text{Linear models}}_{\text{establish relations}, x_i \to y_i, X\beta + \varepsilon = Y} + \underbrace{\text{Exponential Families}}_{\text{flexible classescontinuous (normal)discrete (Poisson)Binomial (Bernoulli)}} \to \text{GLMs.}$$

Note, General (X, F) are allowed, but Θ will still be an open set in $\mathbb{R}^p$

Example 6. *Suppose we have binary data, $Y_i \in \{0, 1\}$, $i = 1, \dots, n$*

1. *Covariates $X_i \in \mathbb{R}^p$*
2. *Logistic Regression for $\beta \in \mathbb{R}^p$,*

$$p_\beta(Y_i = 1|X_i) = \frac{\exp(X_i^T \beta)}{1 + \exp(X_i^T \beta)}$$

Equivalently,

$$X_i^T \beta = \log\left(\frac{P_\beta(Y_i = 1|X_i)}{1 - P_\beta(Y_i = 1|X_i)}\right) = \log it(P_\beta(Y_i = 1|X)).$$

Example 7. *Poisson Regression*

1. *$Y_i \sim Poission(\lambda_i)$, $\lambda_i > 0$.*
2. *$\log \lambda_i = X_i^T \beta$*
3. *$X_i \in \mathbb{R}^p$, $\beta \in \mathbb{R}^p$*
4. *Used for count data*

Definition 15. *Generalized Linear Models We have the data $(X_i, Y_i) \in \mathbb{R}^p \times \mathbb{R}$ In a GLM,*

$$dP_{n_i}^{Y_i}(y_i) = c(y_i) \exp(n_i y_i - \Psi(n_i))$$

where n_i and x_i are linked through some link function

$$g \colon \mathbb{R} \to \mathbb{R},$$

$$g(E_{n_i}[Y_i]) = X_i^T \beta.$$

Intuition, $n_i \iff E_{n_i}[Y_i] = \mu_i \iff X_i^T \beta$ with $g' > 0$.
When $g(E_{n_i}[Y_i]) = n_i$, then g is called the canonical link, or the natural link function.

Remark 16. 1. *Under the canonical link,*

$$p_\beta^{Y_i}(Y_i) = c(y_i) \exp(Y_i X_i^T \beta - \Psi(X_i^T \beta)).$$

2. *Link function links the linear predictors X_i to the mean of the outcome.*

$$E[Y_i|X_i, \beta] = g^{-1}(X_i^T \beta)$$

Lecture 15: GLMs & Model Selection

Natural Exponential Family (Nat. EF)

$$P_\eta(x) = c(x)\exp\big(\eta T(x) - \Psi(\eta)\big), \quad \eta \in \mathbb{R}$$

- What is the Fisher information?
- Is the Cramer-Rao lower bound attained?

$$\mathbb{E}_\eta[T] = \Psi'(\eta)$$

- What happens if $T = \text{const. a.s.}$?

Generalized Linear Models (GLM)

- Parameter: $\beta \in \mathbb{R}^p$
- Design matrix: $X \in \mathbb{R}^{n\times p}$
- Data: $Y_i \sim c(y)\exp\big(yX_i^\top\beta - \Psi(X_i^\top\beta)\big)$
- $$g\big(\mathbb{E}[Y_i \mid X_i, \beta]\big) = X_i^\top\beta$$
- (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:

$$g(p_i) = \log\frac{p_i}{1-p_i}, \quad g^{-1}(X_i^\top\beta) = \frac{\exp(X_i^\top\beta)}{1+\exp(X_i^\top\beta)}$$

- $$\begin{aligned} P_\beta(Y_i) &= p_i^{Y_i}(1-p_i)^{1-Y_i} \\ &= \exp\big(Y_i\log p_i + (1-Y_i)\log(1-p_i)\big) \\ &= \exp\Big(Y_i\log\frac{p_i}{1-p_i} + \log(1-p_i)\Big) \\ &= \exp\big(Y_iX_i^\top\beta - \Psi(X_i^\top\beta)\big), \quad \Psi(X_i^\top\beta) = \log(1+\exp(X_i^\top\beta)) \end{aligned}$$

MLE in Logistic Regression

Goal: Find $\hat\beta \in \arg\max_{\beta\in\mathbb{R}^p} P_\beta(Y)$.

$$\begin{aligned} \log p_\beta(Y) &= \sum_{i=1}^n \log p_i^{Y_i}(1-p_i)^{1-Y_i} \\ &= \sum_{i=1}^n Y_i\log\frac{p_i}{1-p_i}\log(1-p_i) \\ &= \sum_{i=1}^n y_i \underbrace{\eta_i}_{X_i^T\beta} + \log\frac{1}{1+e^{\eta^t}} \\ \nabla_\beta \log p_\beta(Y) &= \sum_{i=1}^n Y_i\nabla_\beta\eta_i - \frac{1}{1+e^{\eta_i}}e^{\eta_i}\nabla_\beta\eta_i \\ &= \sum_{i=1}^n \nabla_\beta\eta_i\left(Y_i - \underbrace{\frac{e^{\eta_i}}{1+e^{\eta_i}}}_{p_i}\right) = \sum_{i=1}^n (Y_i - p_i)X_i^T \\ &= X^T(y-p) \in \mathbb{R}^p \end{aligned}$$

MLE in Logistic regression,

$$X^T(Y-p) \neq 0$$

where $p = (p_1, \ldots, p_n)$ is $p_i = \frac{e^{X_i^T\beta}}{1+e^{X_i^T\beta}}$

Caustion, $\hat{\beta}$ 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 17. *$\hat{\beta}$ the MLE in GLMs are not given in closed form, and have to be computed using optimization methods (e.g. gradient decent, Newton's Method)*

Lemma 6. *Score and FI in GLMs Consider a GLM with canonical links. Then,*

$$\nabla_\beta \log p_\beta(Y) = \sum_{i=1}^n Y_i - \Psi'(X_i^T\beta)X_i^T$$

$$I(\beta) = \sum_{i=1}^n \Psi''(X_i^T\beta)X_iX_i^T \in \mathbb{R}^p$$

Proof.

$$\log p_\beta(Y) = \sum_{i=1}^n \log c(y_i) + y_iX_i^T - \Psi(X_i^T\beta)$$

$$\implies \nabla_\beta \log p_\beta(Y) = \sum_{i=1}^n Y_iX_i^T - \Psi'(X_i^T\beta)X_i^T = \sum_{i=1}^n (Y_i - \psi'(X_i^T\beta))X_i^T$$

Once checks that

$$\frac{\partial^2}{\partial\beta_k\beta_\ell}\log p_\beta(Y) = \sum_{i=1}^n -\Psi''\Psi''(X_i^T\beta)(x_i)_k(x_i)_\ell x$$

Since $I(\beta) = E_\beta[-\nabla_\beta(\log p_\beta(x))]$, the claim follows. □

Remark 18. *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 defininate, then $\hat{\beta}$ is unique (not clear in general)

13.1 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 $k = 1, \ldots, K \leq n$, 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,*

$$X^{(k)} = \textit{first } k \textit{ columns of } X \textit{ where } X \textit{ is a } k \textit{ by } n \textit{ matrix}$$

We can ask ourselves, what's the best model.

Fro example,

$$\hat{\beta}^{(k)} = ((X^k)^T X^{(k)})^{-1} {X^{(k)}}^T Y = \arg\min_{\beta^{(k)}\in\mathbb{R}^k} \|Y - X^{(k)}\beta^{(k)}\|$$

Which has the least MSE.

$$\|\underbrace{\mu - X^{(k)}\hat{\beta}^{(k)}}_{\hat{\mu}^{(k)}}\|^2$$

We calculate,

$$E\|\mu - \hat{\mu}^{(k)}\|^2_{\mathbb{R}^k} = (\mu - E[\hat{\mu}^{(k)}])^2 + E[(\mu^{(k)} - E[\hat{\mu}^{(k)}])^2] - 2\underbrace{E[\langle \mu - E[\hat{\mu}^{(k)}], \hat{\mu}^{(k)} - E[\hat{\mu}^{(k)}]\rangle]}_{=0}$$

Moreover,

$$E[m\hat{u}^{(k)}] = E[\Pi^{(k)}Y] = E[\Pi^{(k)}(\mu+\varepsilon)] = \Pi^{(k)}\mu$$

So then,

$$\hat{\mu}^{(k)} - E[\hat{\mu}^{(k)}] = \Pi^{(k)}\varepsilon$$

$$\implies E\|\mu - \hat{\mu}^{(k)}\|^2_{\mathbb{R}^k} = \|(I_n - \Pi^{(k)})\mu\|^2 + \underbrace{E[\Pi^{\varepsilon}]^2}_{\sigma^2\chi^2(k)}$$

and all of this is equal to

$$= \|(I_n - \Pi^{(k)})\mu\|^2 + k\sigma^2 = \text{ BIAS + VARIANCE with k}$$

How well is μ approximated by $\mathrm{col}(X^{(k)})$ We'd like to pick

$$\hat{k} = \arg\min_{k=1,\ldots K} \underbrace{\|(I_n - \Pi^{(k)})\mu\|}_{\text{Unknown}} + k\sigma^2$$

To estimate the first term, we consider RSS,

$$\begin{aligned} E\|\underbrace{Y - X^{(k)}\hat{\beta}^{(k)}}_{\text{Data Driven}}\|^2 &= E\|(I_n - \Pi^{(k)})(\mu+\varepsilon)\|^2 \\ &= E\|(I_n - \Pi^{(k)})(\mu)\|^2 + E\|(I_n - \Pi^{(k)})(\varepsilon)\|^2 + E[\langle \underbrace{(I_n - \Pi^{(k)})\mu}_{detm}, \underbrace{(I_n - \Pi^{(k)})\varepsilon}_{E[]=0} \rangle] \\ &= \|(I_n - \Pi^{(k)})\mu\|^2 + \sigma^2(n-k) \end{aligned}$$

Which implies

$$\|Y - X^{(k)}\hat{\beta}^{(k)}\|^2 - \sigma^2(n-k) + \sigma^2 k$$

is unbiased risk estimator for risk, $E\|\mu - \hat{\mu}^{(k)}\|^2$

Method (Mallow's Cp)

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 Alaike's Information Criterion. (AIC).