Chapter 1

Gauss-Markov Estimation

Let $(V, \langle \cdot, \cdot \rangle)$ be an Euclidean space. We assume that the random vector $Y \in V$ is isotropic, that is , $\mathrm{Cov}(Y) = \sigma^2 I_V$, and that the unknown expectation $\mu = \mathbb{E}[Y]$ belongs to some subspace $M \subset V$.

Our goal is to estimate μ knowing that $\mu \in M$. We know that μ is uniquely defined by the given of $\psi(\mu)$ for any linear functional $\psi : M \to \mathbb{R}$, that is $\psi(\cdot) = \langle u, \cdot \rangle$ for some $u \in M$ by the representation theorem.

1.1 Linear functionals

For any linear functional $\psi : M \to \mathbb{R}$, there exists a unique vector $cv(\psi) \in M$ such that

$$\psi(m) = \langle cv(\psi), m \rangle, \quad m \in M.$$

$cv(\psi)$ is called the coefficient vector of ψ. We insist that $cv(\psi) \in M$.

Example 1.1. *[The triangle problem] Assume that $V = \mathbb{R}^3$ and*

$$M = \left\{ \sum_{1 \leq j \leq 3} \beta_j e^{(j)} \,:\, \beta_1 + \beta_2 + \beta_3 = 0 \right\} = \left\{ x \in \mathbb{R}^3 \,:\, x_1 + x_2 + x_3 = 0 \right\} = e^{\perp},$$

where $e = e_1 + e_2 + e_3$ and e_1, e_2, e_3 is the canonical basis of $\mathbb{R}^3$.

We consider the linear functional ψ_j on M defined by

$$\psi_j \left(\sum_i \beta_i e_i \right) = \beta_j = \langle \sum_i \beta_i e_i, e_j \rangle,$$

Note that e_j is not the coefficient vector of ψ_j since $e_j \notin M$. The coefficient vector is $P_M(e_j)$ the orthogonal projection of e_j onto M, that is

$$P_M(e_j) = (I - P_e)(e_j) = e_j - \frac{1}{\|e\|^2}\langle e, e_j\rangle e = e_j - \frac{1}{3}e.$$

Proposition 1.1. *Let M be a subspace of V. Assume that M is generated by linearly independent vectors $x_1, \ldots, x_d$. That is $M := \text{l.s.}\,(x_1, \cdots, x_d)$.Set*

$$M_j = \text{l.s.}\,(\{\, x_1, \ldots, x_d \,\} \backslash \{x_j\}) \,,$$

and $P^{\perp}_{M_j}$ the orthogonal projection onto $M_j^{\perp}$. Then, for any j the coefficient vector of the linear functional $\psi_j(\sum_i \beta_i x_i) = \beta_j$ is

$$cv(\psi_j) = \frac{P^{\perp}_{M_j}(x_j)}{\|P^{\perp}_{M_j}(x_j)\|^2}.$$

Proof. Any vector $m \in M$ admits the following UNIQUE decomposition onto this basis

$$m = \sum_{i=1}^{d} \beta_i x_i, \quad \beta_1, \ldots, \beta_d \in \mathbb{R}.$$

We consider the linear functional

$$\psi_j\left(\sum_i \beta_i x_i\right) = \beta_j.$$

We now want to determine the coefficient vector of ψ_j. Note that the basis $x_1, \ldots, x_d$ is not orthogonal in general. We need to pay a little attention to obtain the right coefficient vector. Set $m_j = \sum_{i=1:i\neq j}^{d} \beta_i x_i$. For brevity, set $v = cv(\psi_j)$. We have for any $m \in M$ that

$$\begin{aligned} \beta_j = \langle v, m\rangle &= \langle v, \beta_j x_j\rangle + \langle v, m_j\rangle \\ &= \beta_j + \langle v, m_j\rangle \end{aligned}$$

The above display imply that $v \in M_j^{\perp}$. Take

$$v = \frac{P^{\perp}_{M_j}(x_j)}{\|P^{\perp}_{M_j}(x_j)\|^2},$$

where $P^{\perp}_{M_J}$ is the orthogonal projection onto $M_j^{\perp}$. We have indeed that $P^{\perp}_{M_j}(x_j) \neq 0$ since $x_1, \ldots, x_d$ is a basis of M.

Then, we have for any $m \in M$ that

$$\langle v, m\rangle = \beta_j \frac{\langle P_{M_j}^{\perp}(x_j), x_j\rangle}{\|P_{M_j}^{\perp}(x_j)\|^2} + \frac{\langle P_{M_j}^{\perp}(x_j), m_j\rangle}{\|P_{M_j}^{\perp}(x_j)\|^2} = \beta_j.$$

$\square$

Exercise 1.1. *In simple linear regression $Y = \alpha e + \beta x + \epsilon$, we have $V = \mathbb{R}^n$ and $\mu = (\alpha, \beta)^\top$ in the basis e, x of*

$$M = \text{l.s.}(e, x),$$

where $e = \sum_{i=1}^n e_i$ and $e_1, \ldots, e_n$ is the canonical basis of $\mathbb{R}^n$ and $x \in \mathbb{R}^n$ is a given vector distinct from e. We want to estimate the slope β, that is the linear functional $\psi_\beta(\alpha e + \beta x) = \beta$. Determine the coefficient vector of ψ_β.

Exercise 1.2. *In multiple linear regression $Y = \alpha e + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon$, we have $V = \mathbb{R}^n$ and $\mu = (\alpha, \beta_1, \cdots, \beta_p)^\top$ in the basis $e, x_1, \cdots, x_p$ of*

$$M = \text{l.s.}(e, x_1, \cdots, x_p),$$

where $e = \sum_{i=1}^n e_i$ and $e_1, \ldots, e_n$ is the canonical basis of $\mathbb{R}^n$ and $e, x_1, \ldots, x_p$ is a basis of M (for simplicity, we assume that the family is linearly independent). We want to estimate β_j, that is the linear functional $\psi_j(\alpha e + \sum_{i=1}^p \beta_i x_i) = \beta_j$. Determine the coefficient vector of ψ_j.

1.2 Estimation of linear functionals of μ

When we observe $Y \in V$ with $\mathbb{E}(Y) = \mu$ and we assume that $\mu \in M$, then $P_M(Y)$ is a natural candidate estimator of μ and similarly for any linear functional ψ, $\psi(P_M(Y))$ is natural estimator of $\psi(\mu)$. We know investigate some basic statistical properties of these estimators.

Definition 1.1. *The Gauss-Markov Estimator (GME) $\hat{\psi}(Y)$ of a linear functional $\psi(\mu)$ of μ is*

$$\hat{\psi}(Y) = \psi(P_M(Y)) = \langle cv(\psi), Y\rangle.$$

(since $cv(\psi) \in M$ by definition of the coefficient director)

For any $x \in V$, the GME of the linear functional $\mu \to \langle x, \mu\rangle$ is $\langle P_M(x), Y\rangle$.

Example 1.2 (Continuation of 1.1)**.** *The GME of β_j is*

$$\hat{\beta}_j = \langle cv(\psi_j), Y\rangle = \langle e_j - \frac{1}{3}e, Y\rangle = Y_j - \frac{Y_1 + Y_2 + Y_3}{3}.$$

Exercise 1.3. *In the simple linear regression model, explicit the GME of the slope* β

Exercise 1.4. *In the multiple linear regression model, write down the GME of* ψ_j.

Exercise 1.5 (ANOVA). *We observe real-valued random variables* $Y_{i,j}$ *with* $1 \leq i \leq p$, $1 \leq j \leq n_i$ *and* $\sum_i n_i = n$ *and such that* $\mathbb{E}(Y_{i,j}) = \mu_i$ *for any* $1 \leq j \leq n_i$ *and any* $1 \leq i \leq j$. *Take* $V = \mathbb{R}^n$ *and consider* $Y = (Y_{i,j})_{1\leq i\leq p,\, 1\leq j\leq n_i}$ *as a random vector with values in* V.

$$M = \text{l.s.}\,\{v_1, \cdots, v_p\}$$

with $(v_1)_j = 1$ *if* $1 \leq j \leq n_1$ *and zero otherwise,* $(v_2)_j = 1$ *if* $n_1 + 1 \leq j \leq n_2$, *etc,* $(v_p)_j = 1$ *if* $n_{p-1} + 1 \leq j \leq n_p$. *Determine the coefficient vector of the linear functional* $\psi_i : x \to \beta_i$ *for any* $x \in M$ *and the GME estimator.*

We recall now some basic properties of the GME.

Proposition 1.2. *Let* Y *be a isotropic random vector wih values in* V. *The GME* $\hat{\psi}(Y)$ *of a linear functional* $\psi(\mu)$ *is a linear transformation of* Y *and an unbiased estimator of* $\psi(\mu)$, *that is* $\mathbb{E}_\mu \hat{\psi}(\mu) = \psi(\mu), \quad \forall \mu \in M$. *(We have indeed that* $\mathbb{E}_\mu \hat{\psi}(Y) = \mathbb{E}_\mu \langle cv(\psi), Y\rangle = \langle cv(\psi), \mu\rangle = \psi(\mu)$ *and*

$$\text{Var}(\hat{\psi}(Y)) = \sigma^2 \|cv(\psi)\|^2 = \sigma^2 \|\psi\|^2.$$

We now state the main result of this chapter that says that $\hat{\psi}(Y)$ is the best linear unbiased estimator of $\psi(\mu)$ in the following sense.

Theorem 1.1. *[Gauss-Markov theorem] For each linear functional* ψ *of* μ, *the GME* $\hat{\psi}(Y)$ *is the unique estimator having minimum variance in the class of linear unbiased estimators of* $\psi(\mu)$.

Proof. Suppose for a given $x \in V$, $\langle x, Y\rangle$ unbiasedly estimates $\psi(\mu)$, so that $\psi(\mu) = \mathbb{E}_\mu(\langle x, Y\rangle) = \langle x, \mu\rangle$ for every element μ of M. Then, we have $cv(\psi) = P_M(x)$ and

$$\text{Var}(\langle x, Y\rangle) = \sigma^2\|x\|^2 \geq \sigma^2 \|P_M(x)\|^2 = \sigma^2\|\psi\|^2 = \text{Var}(\hat{\psi}(Y)),$$

with equality if and only if $x = P_M(x)$, that is, $\langle x, Y\rangle = \hat{\psi}(Y)$. □

Proposition 1.3. *The covariance between two GMEs* $\hat{\psi}_1$ *and* $\hat{\psi}_2$ *is given by*

$$\text{Cov}(\hat{\psi}_1(Y), \hat{\psi}_2(Y)) = \sigma^2 \langle cv(\psi_1), cv(\psi_2)\rangle = \sigma^2\langle \psi_1, \psi_2\rangle,$$

In particular, $\hat{\psi}_1$ *and* $\hat{\psi}_2$ *are uncorrelated if and only if* $cv(\psi_1)$ *and* $cv(\psi_2)$ *are orthogonal.*

Example 1.3 (Continuation of 1.1)**.** *In the triangle problem, the covariance between $\hat{\beta}_i$ and $\hat{\beta}_j$ is*

$$\begin{aligned}\operatorname{Cov}\left(\hat{\beta}_i, \hat{\beta}_j\right) &= \sigma^2 \langle P_{[e]}^{\perp}(e_i), P_{[e]}^{\perp}(e_j)\rangle \\ &= \sigma^2 \left(\langle e_i, e_j\rangle - \langle P_e(e_i), P_e(e_j)\rangle\right) \\ &= \sigma^2 \left(\delta_{ij} - 1/3\right).\end{aligned}$$

1.3 Estimation of μ and σ^2

We note that $P_M(Y)$ is an unbiased estimator of μ. In addition, in view of Gauss-Markov theorem, we can deduce that it has minimum variance in the class of linear unbiased estimators. More precisely, for any linear estimator DY of μ, we have

$$\Sigma(DY) \geq \Sigma(P_M(Y)).$$

where $\geq$ refers to the ordering of symmetric matrices. We can say that this result is the vector version of the Gauss-Markov theorem we initially stated for linear functionals.

We recall that $P_{M^\perp}(Y)$ has in $M^\perp$ an isotropic distribution with zero mean and covariance operator $\sigma^2 I_{M^\perp}$. Thus, when $d(M) = dim(M) < d(V)$, the following estimator of σ^2:

$$\hat{\sigma}^2 = \frac{\|P_M^\perp(Y)\|^2}{d(M^\perp)} = \frac{\|Y\|^2 - \|P_M(Y)\|^2}{d(V) - d(M)}$$

is unbiased. The natural estimator of the variance of the GME $\hat{\psi}(Y)$ is

$$\sigma_\psi^2 = \hat{\sigma}^2 \|\psi\|^2.$$

Example 1.4. *We consider again the simple linear regression, $V = \mathbb{R}^n$, $Y_1, \ldots, Y_n$ are uncorrelated random variables with equal variance σ^2 and $\mathbb{E}(Y_i) = \alpha + \beta(x_i - \bar{x})$ for $1 \leq i \leq n$ with $x_1, \ldots, x_n$ known constants. The regression manifold in this case is $M = l.s.(e, v)$ with $e = e_1 + \cdots + e_n$ and $v = x - \bar{x}$ where $x = (x_1, \cdots, x_n)^\top$. Because $e \perp v$, we have*

$$P_M(Y) = \hat{\alpha} e + \hat{\beta} v,$$

where $\hat{\alpha} = \frac{\langle Y, e\rangle}{\|e\|^2}$ and $\hat{\beta} = \frac{\langle Y, v\rangle}{\|v\|^2}$. Hence, the estimator of σ^2 is

$$\hat{\sigma}^2 = \frac{\|P_M^\perp(Y)\|^2}{n-2} = \frac{1}{n-2}\left(\sum_i Y_i^2 - n\hat{\alpha}^2 - \hat{\beta}^2 \sum_i (x_i - \bar{x})^2\right).$$

The estimators of the variances of $\hat{\alpha}$ *and* $\hat{\beta}$ *are*

$$\hat{\sigma}_\alpha^2 = \frac{\hat{\sigma}^2}{\|e\|^2} = \frac{\hat{\sigma}^2}{n}$$

and

$$\hat{\sigma}_\beta^2 = \frac{\hat{\sigma}^2}{\|v\|^2} = \frac{\hat{\sigma}^2}{\sum_i (x_i - \bar{x})^2}.$$

Chapter 2

Normal estimation

Let $(V, \langle\cdot,\cdot\rangle)$ be an Euclidean space, M is a linear subspace of V and Y is weakly spherical random vector with values in V and expectation $\mu \in M$. We assume in addition throughout the chapter that

$$Y \sim N_V(\mu, \sigma^2 I_V), \quad \mu \in M,\ \sigma^2 > 0.$$

We will show that the GME $P_M(Y)$ of μ enjoys some remarkable statistical properties:

1. This is also the Maximum Likelihood Estimator (MLE)
2. It has minimum variance among the class of linear unbiased estimator of μ
3. It is minimax with respect to the mean square error.

2.1 Maximum likelihood estimation

Relative to the Lebesgue measure on V, Y has density

$$\begin{aligned} f_{\mu,\sigma^2}(y) &= \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} e^{-\frac{1}{2\sigma^2}\langle y-\mu, \Sigma^{-1}(y-\mu)\rangle} \\ &= \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\frac{1}{2\sigma^2}\|P_M(Y)-\mu\|^2} e^{-\frac{1}{2\sigma^2}\|P_M^{\perp}(y)\|^2}, \end{aligned}$$

where $n = \dim(V)$ and $\Sigma = \Sigma(Y)$. The MLE estimator $(\hat{\mu}_{MLE}, \hat{\sigma}^2_{MLE})$ satisfies

$$\hat{\mu}_{MLE} = P_M(Y), \quad \hat{\sigma}^2_{MLE} = \frac{\|P_M^{\perp}(Y)\|^2}{n}.$$

Proof as an exercise.

Thus we get that

$$\hat{\mu}_{MLE} \sim N_V(\mu, \sigma^2 I_M), \quad \hat{\sigma}^2_{MLE} \sim \sigma^2 \chi^2_{d(M^\perp)}/n.$$

and $\hat{\sigma}^2_{MLE}$ is independent of $\hat{\mu}_{MLE}$.

Exercise 2.1. *Suppose that $\psi(\mu)$ is a linear functional on M, $\hat{\psi}(Y)$ is its GME, and $\hat{\sigma}_{\hat{\psi}} = \hat{\sigma}\|\psi\|$ is its estimated standard error. Prove that*

$$\frac{\hat{\psi}(Y) - \psi(\mu)}{\hat{\sigma}_{\hat{\psi}}}$$

has a t distribution with $d(M^\perp)$ degrees of freedom.

2.2 Minimum variance unbiased estimation

Set

$$\hat{\mu} = \hat{\mu}_{MLE} = P_M(Y)$$

and

$$\hat{\sigma}^2 = \frac{n}{d(M^\perp)} \hat{\sigma}^2_{MLE} = \frac{\|P_M^\perp(Y)\|^2}{d(M^\perp)}.$$

We are going to show that they have minimum dispersion in the class of all unbiased estimators because they are functions of a complete sufficient statistic (Lehmann-Scheffe's theorem). This result was already established in Gauss-Markov theorem for linear functionals of weakly spherical random vectors. This alternative approach of proof for Gaussian vector also gives optimality of the variance estimator.

Definition 2.1. *A statistic $T(Y)$ is sufficient for μ and σ^2 if for each possible value t of T, the conditional distribution of Y given $T(Y) = t$ does not depend on the parameters μ, σ^2.*

A statistic $T(Y)$ is complete if whenever g is a function such that

$$\mathbb{E}_{\mu,\sigma^2} g(T(Y)) = 0, \quad \forall \mu \text{ and } \sigma^2,$$

then

$$\mathbb{P}_{\mu,\sigma^2}\left(g(T(Y)) \neq 0\right) = 0, \quad \forall \mu \text{ and } \sigma^2.$$

Theorem 2.1 (Lehmann-Scheffe)**.** *Let $T(Y)$ be a sufficient and complete statistic of Y. Then, each function of $T(Y)$ is the minimum variance unbiased estimator of its expected value.*

We will apply this result to prove that $\hat{\mu}$ and $\hat{\sigma}^2$ admit minimum variance within the class of unbiased estimators. To this end, we need to exhibit a sufficient and complete statistic of Gaussian vector Y.

We essentially exploit the fact that the Gaussian distribution belongs to the family of exponential distributions. Note that the density of Y can be rewritten as

$$f_{\mu,\sigma^2}(y) = C(\theta_1, \cdots, \theta_p, \theta_{p+1})e^{\sum_{1\leq i\leq p+1} T_i(y)\theta_i}$$

where, with $b_1, \ldots, b_p$ denoting an orthonormal basis of M,

$$T_i(y) = \langle P_M y, b_i\rangle, \quad \theta_i = \langle \mu/\sigma^2, b_i\rangle, \quad \text{for } i = 1, \ldots, p,$$

$$T_{p+1}(y) = \|y\|^2, \quad \theta_{p+1} = -\frac{1}{2\sigma^2},$$

and

$$C(\theta_1, \cdots, \theta_p, \theta_{p+1}) = \frac{1}{\pi^{n/2}}(-\theta_{p+1})^{n/2}e^{-\frac{1}{2}\sum_{1\leq i\leq p}\theta_i^2}.$$

Notice that as (μ, σ^2) ranges over $M \times (0, \infty)$, $\theta = (\theta_1, \cdots, \theta_p, \theta_{p+1})$ ranges over

$$\Theta = \mathbb{R}^p \times (-\infty, 0).$$

If follows from the factorization above that the statistic

$$T(Y) = (T_1(Y), \cdots, T_p(Y), T_{p+1}(Y))$$

is sufficient; moreover, $T(Y)$ is complete because the possible distributions of $T(Y)$ constitute an exponential family and Θ has a nonempty interior as a subset of $\mathbb{R}^{p+1}$.

Finally, we note that

$$\hat{\mu} = P_M(Y) = \sum_{1\leq i\leq p} T_i(Y)b_i,$$

and

$$\hat{\sigma}^2 = \frac{\|P_M^\perp(Y)\|^2}{d(M^\perp)} = \frac{\|Y\|^2 - \|P_M(Y)\|^2}{d(M^\perp)} = \frac{T_{p+1}(Y) - \sum_{1\leq i\leq p} T_i^2(Y)}{d(M^\perp)}.$$

are indeed measurable functions of $T(Y)$.

2.3 Minimaxity of $P_M(Y)$

We assume now that σ^2 is known. For simplicity, we take $\sigma = 1$.

$$Y \sim N_V(\mu, I_V), \quad \mu \in M.$$

We define the mean square risk of an estimator $\hat{\mu} = \hat{\mu}(Y)$ as :

$$R(\hat{\mu}, \mu) = \mathbb{E}_{\mu}\|\hat{\mu} - \mu\|^2 = \mathrm{tr}(\Sigma(\hat{\mu})) + \|\mathbb{E}_{\mu}\hat{\mu} - \mu\|^2$$

We have

$$R(\hat{\mu}, \mu) = \mathrm{tr}(\Sigma(\hat{\mu})) + \|\mathbb{E}_{\mu}\hat{\mu} - \mu\|^2$$

Consider now the estimator $\hat{\mu} = P_M(Y)$. Since this estimator is unbiased and admits covariance P_M, we have

$$R(P_M(Y), \mu) = \dim(M).$$

Definition 2.2. *The minimax risk of an estimator $\tilde{\mu}$ of $\mu \in M$ is defined as*

$$\bar{R}_M(\tilde{\mu}) = \sup_{\mu \in M} R(\tilde{\mu}, \mu).$$

We say that an estimator $\hat{\mu}$ is minimax if

$$R(\hat{\mu}) = \inf_{\tilde{\mu}} \bar{R}(\tilde{\mu}),$$

where the infimum is taken on all estimators $\hat{\mu}$ which are measurable functions of Y with values in M.

Definition 2.3. *The bayesian risk of an estimator $\tilde{\mu}$ of μ w.r.t a prior Π on M is given by*

$$R_{\Pi}(\tilde{\mu}, \mu) = \int_M R(\hat{\mu}, \mu)\Pi(d\mu).$$

We say that an estimator $\hat{\mu}$ is bayes optimal w.r.t the prior Π if

$$R_{\Pi}(\hat{\mu}, \mu) = \inf_{\tilde{\mu}} R_{\Pi}(\tilde{\mu}, \mu) =: B(\Pi).$$

We say that an estimator $\hat{\mu}$ is ϵ-Bayes optimal w.r.t to Π if $R_{\Pi}(\hat{\mu}, \mu) - B(\Pi) \leq \epsilon$.

We say $\hat{\mu}$ is extended bayes if for each n there exists a prior Π_n such that $\hat{\mu}$ is $1/n$-Bayes w.r.t Π_n.

We now prove that $P_M(Y)$ is minimax. To this end, we use the following Lemma.

Lemma 2.1. *Assume that $\hat{\mu}$ is an estimator of μ having finite constant risk, say r. If there exists a sequence $(\Pi_n)_n$ of priors on μ such that $B(\Pi_n) \to r$, then $\hat{\mu}$ is extended Bayes and minimax and r is the minimax risk of estimation of μ.*

Proof. We have

$$\bar{R}(\tilde{\mu}) = \sup_{\mu \in M} R(\tilde{\mu}, \mu) \geq \int_M R(\tilde{\mu}, \mu)\Pi_n(d\mu) \geq B(\Pi_n), \quad \forall \tilde{\mu},\ n.$$

Taking $n \to \infty$, we get

$$\bar{R}(\tilde{\mu}) \geq r = \bar{R}(\hat{\mu}).$$

□

We now need to build this sequence Π_n. Let Π be a prior on μ. Let Θ be a random variable distributed as Π and such that $Y|\Theta = \mu \sim N_V(\mu, I_V)$.

Lemma 2.2. *For any prior Π on μ, the Bayes estimator w.r.t. Π is $\rho(Y)$, where $\rho(y) = \mathbb{E}(\Theta|Y = y)$ is the mean of the posterior distribution of μ given that $Y = y$ and where*

$$B(\Pi) = R(\rho, \Pi) = \int_M R(\rho, \mu)\Pi(d\mu).$$

Proof. For the sake of completeness, we exclude from this proof the details of measurability and integrability. For any estimator δ, we have

$$\begin{aligned}
R(\tilde{\mu}, \Pi) &= \int_M \mathbb{E}_\mu \left(\|\tilde{\mu}(Y) - \mu\|^2\right) \Pi(d\mu) \\
&= \int_M \mathbb{E}_\mu \left(\|\tilde{\mu}(Y) - \Theta\|^2 |\Theta = \mu\right) \Pi(d\mu) \\
&= \mathbb{E}(\|\tilde{\mu}(Y) - \Theta\|^2) \\
&= \int_V \mathbb{E}_\mu \left(\|\tilde{\mu}(Y) - \Theta\|^2 | Y = y\right) P(dy) \\
&= \int_V \mathbb{E}_y \left(\|\Theta - \tilde{\mu}(y)\|^2\right) P(dy),
\end{aligned}$$

where P denotes the marginal distribution of Y. Set now $\rho(y) = \mathbb{E}_y(\Theta)$. We have

$$\mathbb{E}(\|\Theta - \tilde{\mu}(y)\|^2) = \mathbb{E}_y(\|\Theta - \rho(y)\|^2) + \|\rho(y) - \tilde{\mu}(y)\|^2.$$

Thus, we get that

$$R(\tilde{\mu}, \Pi) \geq R(\rho, \Pi).$$

□

Assume now that Θ has marginal distribution $N_M(0, \lambda I_M)$ for some $\lambda > 0$ and for each $\mu \in M$, the conditional distribution of Y given $\Theta = \mu$ is $N_V(\mu, I_V)$. We can prove that (Θ, Y) are normally distributed (use of characteristic function). Then, we get that

1. $\mathbb{E}(\Theta) = 0$.
2. $\Sigma_{\Theta\Theta} = \lambda I_M$
3. $\mathbb{E}(Y) + \Sigma_{Y\Theta}\Sigma_{\Theta,\Theta}^{-1}(\mu - \mathbb{E}\Theta) = \mu$ for all $\mu \in M$.
4. $\Sigma_{YY} - \Sigma_{Y\Theta}\Sigma_{\Theta,\Theta}^{-1}\Sigma_{\Theta Y} = I_V$.

Exploiting the above relations, we get that the marginal distribution of Y is $N_V(0, I_V + \lambda P_M)$ and for each $y \in V$, the conditional distribution of Θ given $Y = y$ is

$$N_N(\frac{\lambda}{1+\lambda}P_M(y), \frac{\lambda}{1+\lambda}I_M).$$

We know that

$$\tilde{\mu}_\lambda(Y) = \frac{\lambda}{1+\lambda}P_M(Y) + \frac{1}{1+\lambda}0$$

is the Bayes estimator w.r.t $\Pi_\lambda = N_M(0, \lambda I_M)$.

Then, we get that

$$R\left(\frac{\lambda}{1+\lambda}P_M(Y), \mu\right) = \left(\frac{\lambda}{1+\lambda}\right)^2 R(P_M(Y), \mu) + \frac{1}{(1+\lambda)^2}R(0, \mu).$$

and

$$\begin{aligned} B(\Pi_\lambda) &= R(\tilde{\mu}_\lambda, \Pi_\lambda) = \mathbb{E}R(\tilde{\mu}_\lambda, \Theta) \\ &= \left(\frac{\lambda}{1+\lambda}\right)^2 \dim(M) + \frac{\lambda}{(1+\lambda)^2}\dim(M) \\ &= \frac{\lambda}{1+\lambda}\dim(M), \end{aligned}$$

since $\mathbb{E}R(0, \mu) = \mathbb{E}\mathbb{E}_\Theta(R(0, \mu)|\Theta = \mu) = \mathbb{E}(\|\Theta\|^2) = \lambda \dim(M)$.

Taking $\lambda \to \infty$, we get that $B(\Pi_\lambda) \to \dim(M)$.

2.4 James-Stein

The minimax criterion guarantees that $P_M(Y)$ is the best estimator with regards to the worst possible risk. We now wonder if $P_M(Y)$ is the best estimator for any value of μ. Unfortunately, we will answer this question by the negative.

Definition 2.4. *An estimator $\tilde{\mu}$ of μ is admissible if there exists no other estimator $\tilde{\mu}^*$ such that*

$$\begin{aligned} R(\tilde{\mu}^*, \mu) &\leq R(\tilde{\mu}, \mu), \quad \textit{for all } \mu \in M \\ R(\tilde{\mu}^*, \mu) &< R(\tilde{\mu}, \mu), \quad \textit{for some } \mu \in M. \end{aligned}$$

Proposition 2.1. *The estimator $P_M(Y)$ is admissible if and only if If* $\dim(M) \leq 2$.

We consider the Bayesian setting of the previous section where

$$\mu \sim N_M(0, \lambda I_M), \quad Y|\mu \sim N_V(\mu, I_V).$$

We recall that $\frac{\lambda}{1+\lambda} P_M(Y)$ is the bayes estimator of μ provided that λ is known. In the opposite case, we can estimate λ from the data or more precisely $\frac{1}{1+\lambda}$. We have that $X = P_M(Y) \sim N_M(0, (1+\lambda)I_M)$ and

$$S = \|X\|^2 \sim (1+\lambda)\chi_p^2,$$

where $p = \dim(M)$. We also have that

$$\mathbb{E}\left(\frac{1}{\|X\|^2}\right) = \frac{1}{p-2}, \quad \text{if } p \geq 3,$$

and $\mathbb{E}\left(\frac{1}{\|X\|^2}\right) = \infty$ if $p \leq 2$.

Thus we can estimate $\frac{1}{1+\lambda}$ by $\frac{p-2}{S}$ if $p \geq 3$. We obtain the following estimator

$$\hat{\mu}_{JS} = \left(1 - \frac{p-2}{\|P_M(Y)\|^2}\right) P_M(Y). \tag{2.1}$$

This estimator is known as James-Stein estimator

Theorem 2.2. *Assume that $p \geq 3$. Then, the James-Stein estimator $\hat{\mu}_{JS}$ admits the risk*

$$R(\hat{\mu}_{JS}, \mu) = \mathbb{E}_\mu \|\hat{\mu}_{JS} - \mu\|^2 = p - (p-2)^2 \mathcal{E}_p(\|\mu\|^2),$$

where for any $t \geq 0$,

$$\mathcal{E}_p(\|\mu\|^2) = \sum_{0 \leq k < \infty} e^{-t/2} \frac{(t/2)^k}{k!} \frac{1}{p-2+2k} = \mathbb{E}\left(\frac{1}{p-2+2K}\right),$$

where K is a Poisson random variable with parameter $t/2$.

We will prove a simplified version of this result. See the complementary note.

This striking result shows that there exist nonlinear and maybe biased estimators of μ with better mean risk than the linear projection $P_M(Y)$. Looking at the shape of $\hat{\mu}_{JS}$, we see that this estimator is obtained by a perturbation of the $P_M(Y)$ where $P_M(Y)$ is shrunk to 0 in the neighborhood where $\|\mu\|^2$ is small (where we replaced this quantity by its estimator $\|X\|^2 - p$). We will further explore this shrinkage idea in the high-dimensional framework where there are more parameters to estimate than available observations ($p > n$) and how see it yields extremely interesting results when combined with additional low complexity (sparsity) conditions on μ.

Chapter 3

Statistical Testing and Confidence Intervals

Throughout this chapter, we consider again an Euclidean space $(V, \langle\cdot,\cdot\rangle)$ and observation $Y \sim N_V(\mu, \sigma^2 I_V)$, with $\mu \in M_1$ a linear subspace of V and $\sigma^2 > 0$ may be unknown.

3.1 Likelihood Ratio and Fisher Testing

We are interested in the following testing problem

$$H_0 : \mu \in M_0 \quad H_1 : \mu \notin M_0,$$

where M_0 is a linear subspace of M_1.

Likelihood Ratio Test

Recall that Y admits the density

$$f_{\mu,\sigma^2}(y) = \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\frac{1}{2\sigma^2}\|y-\mu\|^2}, \quad n = \dim(V),$$

w.r.t. the Lebesgue measure in V.

Definition 3.1. *The Likelihood Ratio Test (LRT) associated to the hypothesis $H_0 : \mu \in M_0$ versus $H_1 : \mu \in M_1$ where $M_0 \subset M_1 \subset V$ is*

$$\Lambda(Y) = \frac{\sup_{\mu\in M_0,\sigma^2>0} f_{\mu,\sigma^2}(Y)}{\sup_{\mu\in M_1,\sigma^2>0} f_{\mu,\sigma^2}(Y)}.$$

Heuristic: If $\Lambda(Y)$ is small then the null hypothesis is less likely than the hypothesis H_1. Conversely, under the alternative, we expect the value of $\Lambda(Y)$ to be larger.

Let P_{M_i} be the orthogonal projection onto M_i, $i = 1, 2$. We have

$$f_{\mu,\sigma^2}(Y) = \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\frac{1}{2\sigma^2}\|P_{M_i}Y-\mu\|^2} e^{-\frac{1}{2\sigma^2}\|P_{M_i^\perp}Y\|^2}, \quad n = \dim(V).$$

Now, the maximum of $f_{\mu,\sigma^2}(Y)$ over $\mu \in M_i$ and $\sigma^2 > 0$ is obtained at $\hat{\mu}_{MLE} = P_{M_i}(Y)$ and $\hat{\sigma}^2 = \|P_{M_i^\perp}Y\|^2/n$. We get that

$$f_{\hat{\mu}_{MLE},\hat{\sigma}^2_{MLE}}(Y) = \frac{1}{(2\pi)^{n/2}} \left(\frac{n}{\|P_{M_i^\perp}Y\|^2} \right)^{n/2} e^{-n/2}.$$

Thus

$$\Lambda(Y) = \left(\frac{\|P_{M_1^\perp}Y\|^2}{\|P_{M_0^\perp}Y\|^2} \right)^{n/2}.$$

The LRT test rejects the null hypothesis for small values of the ratio of the distance of Y from M over the distance of Y from M_0.

Fisher Test

The Fisher Test is closely related to the LRT. For the purpose of obtaining a test statistic with a known distribution, we manipulate the LRT statistic. The condition that the ratio $\Lambda(Y)$ is small is equivalent to

$$\frac{\|P_{M_0^\perp}Y\|^2}{\|P_{M_1^\perp}Y\|^2} \quad \text{is large}$$

which is also equivalent to

$$\frac{\|P_{M_0^\perp}Y\|^2 - \|P_{M_1^\perp}Y\|^2}{\|P_{M_1^\perp}Y\|^2} \quad \text{is large.}$$

Pythagora's theorem gives

$$\|P_{M_0^\perp}Y\|^2 = \|P_{M_1^\perp}Y\|^2 + \|P_{M_1 \cap M_0^\perp}Y\|^2.$$

Combining the last two displays, we get that the LRT is equivalent to

$$\frac{\|P_{M_1 \cap M_0^\perp}Y\|^2}{\|P_{M_1^\perp}Y\|^2} \quad \text{is large.}$$

We renormalize by the dimensions of the involved subspaces

$$\frac{\dim(M_1 \cap M_0^\perp)}{\dim(M_1^\perp)}.$$

We finally obtain the following Fisher Test statistic

$$T := \frac{\|P_{M_1 \cap M_0^\perp} Y\|^2 / \dim(M_1 \cap M_0^\perp)}{\|P_{M_1^\perp} Y\|^2 / \dim(M_1^\perp)}. \tag{3.1}$$

Next, we have for $\mu \in M_1$ that

$$\|P_{M_1 \cap M_0^\perp} Y\|^2 \sim \sigma^2 \, \chi^2_{\dim(M_1 \cap M_0^\perp), \|P_{M_1 \cap M_0^\perp} \mu\|/\sigma},$$

and

$$\|P_{M_1^\perp} Y\|^2 \sim \sigma^2 \, \chi^2_{\dim(M_1^\perp)}$$

are independent.

Thus

$$T \sim F(\dim(M_1 \cap M_0^\perp), \dim(M_1^\perp), \|P_{M_1 \cap M_0^\perp} \mu\|/\sigma).$$

- If H_1 is true, then the numerator is a biased estimator of the variance σ^2 since

$$\mathbb{E}\|P_{M_1 \cap M_0^\perp} Y\|^2 / \dim(M_1 \cap M_0^\perp) = \sigma^2 + \frac{\|P_{M_1 \cap M_0^\perp} \mu\|^2}{\dim(M_1 \cap M_0^\perp)}.$$

  Thus the test statistic T is larger than 1.

- If M_0 is true, then we have $\|P_{M_1 \cap M_0^\perp} \mu\|/\sigma = 0$. Numerator and denominator are hence unbiased estimators of the variance σ^2 and T is close to 1. Furthermore $T \sim F(\dim(M_1 \cap M_0^\perp), \dim(M_1^\perp))$.

Definition 3.2. *The Fisher test for*

$$\mathbf{H}_0 \ : \ \mu \in M_0 \quad \textit{versus} \quad \mathbf{H}_1 \ : \ \mu \in M_1$$

admits critical region $T \geq c$ where T is defined in (3.1).

- *For $\alpha \in (0,1)$, the type-I error test of level α is given by taking $c = c_\alpha$, the α quantile of the Fisher distribution $F(\dim(M_1 \cap M_0^\perp), \dim(M_1^\perp))$.*

- *The power of the F-test is*

$$\beta(\mu, \sigma^2) = \mathbf{P}_{\mu,\sigma^2}(H_1) = F_{\dim(M_1 \cap M_0^\perp), \dim(M_1^\perp); \gamma}([c_\alpha, \infty]),$$

  where c_α is defined above and the noncentrality parameter γ is given by

$$\gamma = \frac{\|P_{M_1 \cap M_0^\perp} \mu\|}{\sigma}.$$

Proposition 3.1. *For the testing problem* $\mathbf{H}_0 : \mu \in M_0$ *versus* $\mathbf{H}_1 : \mu \in M_1$*, the LRT coincides with the Fisher test.*

Proposition 3.2. *The power of the F-test is an increasing function of the noncentrality parameter*

$$\gamma = \frac{\|P_{M_1 \cap M_0^\perp} \mu\|}{\sigma}.$$

Example 3.1. *Assume* $V = \mathbb{R}^3$ *and* $\mathbb{E}(Y_i) = \beta_i$ *for* $i = 1, 2, 3$ *with* $\beta_1 + \beta_2 + \beta_3 = 0$*, so*

$$M_1 = \left\{ \sum_{j=1}^{3} \beta_j e_j : \ \beta_1 + \beta_2 + \beta_3 = 0 \right\},$$

where e_1, e_2, e_3 *is the canonical basis of* $\mathbb{R}^3$*. Note that* $M_1 = l.s.(e)^\perp$ *where* $e = e_1 + e_2 + e_3$*.. We want to test whether* $\beta_1 = \beta_2 = \beta_3$*. Under the constraint to belong to* M_1*, this is equivalent to*

$$H_0 \ : \ \beta_1 = \beta_2 = \beta_3 = 0,$$

i.e., $M_0 = 0$*. The* T *statistic takes the form*

$$T = \frac{\sum_{i=1}^{3} (Y_i - \bar{Y})^2/2}{3\bar{Y}^2} \sim F(2, 1, \frac{\|\mu\|}{\sigma}).$$

If H_0 *is true, then* $\mu = 0$*.*

Testing the utility of regressors with R. The outputs values are the estimated values of the parameters, the standard deviations and the test statistic under the null assumption H_0: $\beta_i = 0$. We reject H_0 for the two estimated parameters.

Multiple linear regression in R. In the model $Y = X\theta + \epsilon$ with $p + 1$ regressors and the first regressor is the constant $\mathbb{1}_n$. By convention we set $X = [\mathbb{1}_n, X_1, \cdots, X_p]$ and $\theta = (\beta, \theta_1, \cdots, \theta_p)$. We want to test the utility of a subset of the regressors, in other words, the null assumption is

$$H_0 : \{X_{q+1}, \ldots, X_p \, \text{are useless}\}, \quad \text{versus} \quad H_1 := \{\text{this is wrong}\}.$$

We can reformulate this testing problem in term of the α_j

$$H_0 : \{\theta_q = 0, \ldots, \theta_p = 0\}, \quad \text{versus} \quad H_1 := \{\text{at least 1 coefficient } \theta_j \neq 0\}.$$

If H_0 is true, the model becomes

$$Y = X_0\theta_0 + \epsilon, \quad X_0 = [1_n, X_1, \ldots, X_q], \quad \theta_0 = (\beta, \theta_1, \ldots, \theta_q)^\top.$$

The least squares estimator is

$$\hat{\theta}_0 = (X_0^\top X_0)^{-1} X_o^\top Y.$$

Define $M_1 = l.s.\{\mathbb{1}_n, X_1, \cdots, X_p\}$ and $M_0 = l.s.\{\mathbb{1}_n, X_1, \ldots, X_q\}$. We have $\dim(M_1) = p+1$ and $\dim(M_0) = q+1$. Then the test statistic becomes

$$T = \frac{\|P_{M_1}Y - P_{M_0}Y\|^2/(p-q)}{\|Y - P_{M_1}Y\|^2/(n-p-1)} \sim F(p-q, n-p-1), \quad \text{under } H_0.$$

The rejectance region for the test is

$$\{F > q_{1-\alpha}(F(p-q, n-p-1))\}.$$

Exercise 3.1. *We consider* 50 *daily measurements of the ozone concentration, noted* `O3`*, et the explicative variable is the temperature at noon, noted* `T12`*. The data are treated with R. Interpret the following R output.*

```
    > a ~ lm(O3   T12)
> summary(a)
Call:  lm(formula = O3   T12) Residuals:  Min 1Q Median 3Q Max -45.256
-15.326 -3.461 17.634 40.072 Coefficients :  Estimate Std.
Error t value Pr(>|t|) (Intercept) 31.4150 13.0584 2.406 0.0200 *
T12 2.7010 0.6266 4.311 8.04e-05 *** - Signif.  codes:  0 *** 0.001
** 0.01 * 0.05 .  0.1 1
Residual standard error:  20.5 on 48 degrees of freedom
Multiple R-Squared:  0.2791, Adjusted R-squared:  0.2641
F-statistic:  18.58 on 1 and 48 DF, p-value:  8.041e-05
```

3.2 Confidence Intervals for linear functionals of μ

We return to the old notation:

$$Y \sim N_V\left(\mu, \sigma^2 I_V\right), \quad \mu \in M, \quad \sigma^2 > 0. \tag{3.2}$$

Let $\psi(\mu) = \langle cv_\psi, \mu\rangle$ be a nonzero linear functional of μ. Recall that $cv_\psi \in M$ is the coefficient vector of ψ. The best unbiased estimator of $\psi(\mu)$ is

$$\hat{\psi}(\mu) = \langle cv_\psi, Y\rangle = \langle cv_\psi, P_M Y\rangle. \tag{3.3}$$

The standard deviation of $\hat{\psi}(Y)$ is $\sigma_{\hat{\psi}} = \sigma\|cv_\psi\|$, that may be estimated by

$$\hat{\sigma}_{\hat{\psi}} = \hat{\sigma}\|cv_\psi\|, \tag{3.4}$$

where

$$\hat{\sigma}^2 = \frac{\|P_{M^\perp} Y\|^2}{\dim(M^\perp)}. \tag{3.5}$$

Proposition 3.3. *For the usual model (3.2) and the linear functional $\psi : M \to \mathbb{R}$ with coefficient vector cv_ψ, we have for the estimators defined in (3.3) and (3.4) that*

$$\frac{\hat{\psi}(Y) - \psi(\mu)}{\hat{\sigma}_{\hat{\psi}}} \sim t_{\dim(M^\perp)}.$$

Consequently, if $t_m(\beta)$ denotes the quantile of level β of the t-distribution with m degrees of freedom, then we have

$$\mathbf{P}\left(\psi(\mu) \in [\hat{\psi}(Y) \pm t_{\dim(M^\perp)}\left(\frac{\alpha}{2}\right)\hat{\sigma}_{\hat{\psi}}]\right) = 1 - \alpha,$$

for all $\mu \in M$ and all $\sigma^2 > 0$. We say that $\hat{\psi}(Y) \pm t_{\dim(M^\perp)}\left(\frac{\alpha}{2}\right)\hat{\sigma}_{\hat{\psi}}$ is a $100(1-\alpha)\%$ CI for $\psi(\mu)$.

Proof. Note that $\hat{\psi}(Y) \sim N(\psi(\mu), \sigma^2\|cv_\psi\|^2)$ independently of $\hat{\sigma}^2 \sim \sigma^2\xi_{\dim(M^\perp)}/\dim(M^\perp)$. Then we get

$$\frac{\hat{\psi}(Y) - \psi(\mu)}{\hat{\sigma}_{\hat{\psi}}} = \frac{(\hat{\psi}(Y) - \psi(\mu))/(\sigma\|\psi\|)}{\hat{\sigma}/\sigma} \sim t_{\dim(M^\perp)}.$$

The rest follows trivially. □

Example 3.2. *Consider simple linear regression: $V = \mathbb{R}^n$, $Y_1, \ldots, Y_n$ are uncorrelated with equal variances σ^2, and $\mathbb{E}[Y_i] = \alpha + \beta x_i$ for $1 \le i \le n$, with $x_1, \ldots, x_n$ known constants. Here*

$$\mu = \mathbb{E}(\mathbf{Y}) = \alpha\mathbf{e} + \beta\mathbf{x},$$

where $(\mathbf{Y})_i = Y_i$, $(\mathbf{e})_i = 1$ and $(\mathbf{x})_i = x_i$, for $1 \le i \le n$. Fix an x_0 in $\mathbb{R}$ and consider the point on the population regression line above x_0:

$$\psi_{x_0}(\mu) = \alpha + \beta x_0 = \langle cv_{\psi_{x_0}}, \mu\rangle,$$

with

$$cv_{\psi_{x_0}} = \frac{\mathbf{e}}{\|\mathbf{e}\|^2} + (x_0 - \bar{x})\frac{\mathbf{v}}{\|\mathbf{v}\|^2}, \quad \mathbf{v} = \mathbf{x} - \bar{x}\mathbf{e}.$$

The GME of $\psi_{x_0}(\mu)$ is the corresponding point on the fitted line:

$$\widehat{\psi_{x_0}}(\mathbf{Y}) = \langle cv_{\psi_{x_0}}, \mathbf{Y}\rangle = \hat{\alpha} + \hat{\beta}(x_0 - \bar{x}),$$

where

$$\hat{\alpha} = \langle \frac{e}{\|e\|^2}, \mathbf{Y} \rangle = \bar{Y}, \quad \textit{and} \quad \bar{\beta} = \langle \frac{v}{\|v\|^2}, \mathbf{Y} \rangle = \frac{\sum_i (x_i - \bar{x}) Y_i}{\sum_i (x_i - \bar{x})^2}$$

are the GMEs of α and β, respectively. One has

$$\sigma_{\hat{\psi}_{x_0}} = \sigma \|cv_{\psi_{x_0}}\| = \sigma \sqrt{\frac{1}{\|e\|^2} + \frac{(x_0 - \bar{x})^2}{\|v\|^2}} = \sigma \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_i (x_i - \bar{x})^2}},$$

and

$$\hat{\sigma}^2 = \frac{\|\mathbf{Y} - P_M(\mathbf{Y})\|^2}{\dim(M^\perp)} = \frac{\sum_i (Y_i - (\hat{\alpha} + \hat{\beta}(x_i - \bar{x})))^2}{n-2}.$$

Thus

$$\hat{\alpha} + \hat{\beta}(x_0 - \bar{x}) \pm t_{n-2}\left(\frac{\alpha}{2}\right) \hat{\sigma} \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_i (x_i - \bar{x})^2}} \tag{3.6}$$

is a $100(1-\alpha)\%$ confidence interval for $\alpha + \beta(x_0 - \bar{x})$.

Up to now x_0 has been fixed. But it is often the case that one wants to estimate $\psi_{x_0}(\mu) = \alpha + \beta(x_0 - \bar{x})$ simultaneously for all, or at least many, values of x_0. The intervals (3.6) are then inappropriate for $100(1-\alpha)\%$ confidence, because

$$\mathbb{P}_{\mu,\sigma^2}\left(\psi_{x_0}(\mu) \in [\hat{\psi}_{x_0}(\mathbf{Y}) \pm t_{n-2}\left(\frac{\alpha}{2}\right) \hat{\sigma}_{\hat{\psi}_{x_0}}], \quad \forall x_0 \in \mathbb{R}\right) < 1 - \alpha.$$

We will develop a method to build simultaneous inferences on arbitrary familly of linear functionals $\psi(\mu)$. We consider again (3.2). Let $\mathcal{K}$ be a collection of linear functionals of μ and set

$$K = \{cv_\psi \,:\, \psi \in \mathcal{K}\}. \tag{3.7}$$

Let $\mathcal{L}$ be the subspace generated by $\mathcal{K}$ in the vector space M^o of all linear functionals on M and set

$$L = \{cv_\psi \,:\, \psi \in \mathcal{L}\} \subset M, \tag{3.8}$$

equivalently, $L = \text{l.s}(K)$. We note that $\mathcal{L}$ and L are isomorphic, so $\dim(\mathcal{L}) = \dim(L)$.

Let $\mathcal{F}_{f_1,f_2}(\alpha)$ denotes the upper α fractional point of $\mathcal{F}$ distribution with f_1, f_2 degrees of freedom. We set

$$S_{f_1,f_2}(\alpha) = \sqrt{f_1 \mathcal{F}_{f_1,f_2}(\alpha)}. \tag{3.9}$$

Theorem 3.1. *If $\mathcal{L}$ is a subspace of M^o and $L = \{cv(\psi) : \psi \in \mathcal{L}\} \subset M$ is the corresponding subspace of coefficient vectors, then the intervals*

$$\hat{\psi}(\mathbf{Y}) \pm S_{\dim(L),\dim(M^{\perp})}(\alpha)\hat{\sigma}_{\hat{\psi}}$$

cover the $\psi(\mu)$'s for $\psi \in \mathcal{L}$ with simultaneous confidence $100(1-\alpha)\%$.

The confidence intervals in Theorem 3.1 are called Scheffé intervals; $S_{\dim(L),\dim(M^{\perp})}$ is called the Scheffé multiplier. Note that $S_{1,f}(\alpha) = t_f\left(\frac{\alpha}{2}\right)$, so that when $\mathcal{L} = [\psi]$ is 1-dimensional, Theorem 3.1 reduces to the simple assertion that

$$\mathbb{P}\left(\psi(\mu) \in \left[\hat{\psi}(Y) \pm t_{\dim\left(M^{\perp}\right)}\left(\frac{\alpha}{2}\right)\hat{\sigma}_{\hat{\psi}}\right]\right) = 100(1-\alpha)\%.$$

Proof. Assume that some $\psi \in \mathcal{L}$ is nonzero, so $\dim(\mathcal{L}) \geq 1$. We will produce a constant C (depending on $\dim(L)$,$\dim\left(M^{\perp}\right)$ and α) such that

$$\mathbb{P}_{\mu,\sigma^2}\left(\psi(\mu \in [\hat{\psi}(Y) \pm C\hat{\sigma}_{\hat{\psi}}], \quad \forall \psi \in \mathcal{L}\right) = 1-\alpha \tag{3.10}$$

for all μ, σ^2. of course, this implies that

$$\mathbb{P}_{\mu,\sigma^2}\left(\psi(\mu) \in [\hat{\psi}(Y) \pm C\hat{\sigma}_{\hat{\psi}}], \quad \forall \psi \in \mathcal{K}\right) \geq 1-\alpha \tag{3.11}$$

for all μ, σ^2.

Now, we have $\psi(\mu) \in [\hat{\psi}(Y) \pm C\hat{\sigma}_{\hat{\psi}}]$ for all $\psi \in \mathcal{L}$ if and only if

$$\sup_{\psi \in \mathcal{L}\setminus 0} \frac{(\hat{\psi}(Y) - \psi(\mu))^2}{\hat{\sigma}^2_{\hat{\psi}}} \leq C^2.$$

$$\begin{aligned}
\sup_{\psi \in \mathcal{L}\setminus 0} \frac{(\hat{\psi}(Y) - \psi(\mu))^2}{\hat{\sigma}^2_{\hat{\psi}}} &= \sup_{\psi \in \mathcal{L}\setminus 0} \frac{\langle cv_\psi, Y-\mu\rangle^2}{\hat{\sigma}^2 \|\mathrm{cv}_\psi\|^2} \\
&= \frac{1}{\hat{\sigma}^2} \sup_{\psi \in \mathcal{L}\setminus 0} \frac{\langle cv_\psi, Y-\mu\rangle^2}{\|\mathrm{cv}_\psi\|^2} = \frac{1}{\hat{\sigma}^2} \sup_{x \in L\setminus\{0\}} \left(\left\langle \frac{x}{\|x\|}, P_L(Y-\mu)\right\rangle\right)^2 \\
&= \frac{\|P_L(Y-\mu)\|^2}{\hat{\sigma}^2} \equiv Q.
\end{aligned}$$

Since $Y - \mu \sim N_V\left(0, \sigma^2 I_V\right)$ and $L \perp M^{\perp}$, we have

$$\frac{Q}{\dim(L)} = \frac{\|P_L(Y-\mu)\|^2/\dim(L)}{\|P_{M^{\perp}}(Y-\mu)\|^2/\dim(M^{\perp})} \sim \mathcal{F}_{\dim(L),\dim\left(M^{\perp}\right)}.$$

It follows that for all μ, σ^2,

$$\mathbb{P}_{\mu,\sigma^2}\left(\psi(\mu) \in [\hat{\psi}(Y) \pm C\hat{\sigma}_{\hat{\psi}}] \,:\, \psi \in \mathcal{L}\right) = \mathbb{P}_{\mu,\sigma^2}\left(Q \le C^2\right) \sim \mathcal{F}_{\dim(L),\dim\left(M^{\perp}\right)}\left([0, C^2/\dim(L)].\right) \tag{3.12}$$

Finally, (3.10) holds valid with $C = S_{\dim(L),\dim\left(M^{\perp}\right)}(\alpha)$.

□

Example 3.3. *In the simple linear regression model, we put*

$$\mathcal{K} = \left\{\psi_{x_0} = \alpha + \beta(x_0 - \bar{x}) = \left\langle cv_{\psi_{x_0}}, \mu\right\rangle \,:\, x_0 \in R\right\},$$

so

$$K = \left\{\frac{e}{\|e\|^2} + (x_0 - \bar{x})\frac{v}{\|v\|^2} \,:\, x_0 \in \mathbb{R}\right\},$$

and

$$L = \text{l.s.}(K) = \text{l.s.}(e, v) = M \quad \textit{and} \quad \dim(L) = \dim(M) = 2.$$

From Theorem 3.1, the Scheffé intervals

$$\left[\hat{\psi}_{x_0}(Y \pm S_{2,n-2}(\alpha)\hat{\sigma}_{\hat{\psi}_{x_0}}\right] \tag{3.13}$$

covers the various $\psi_{x_0}(\mu) = \alpha + \beta(x_0 - \bar{x})$ for $x_0 \in \mathbb{R}$ with simultaneous confidence $100(1-\alpha)\%$. Note that $\mathcal{K}$ is properly contained in $\mathcal{L} = \text{l.s.}(\mathcal{K})$.

The Scheffé intervals have an interesting connection with the F-test. We consider the testing problem

$$\mathbb{H}_0 \,:\, \psi(\mu) = 0 \quad \forall \psi \in \mathcal{K} \quad \text{versus} \quad \mathbb{H}_1 \,:\, \psi(\mu) \neq 0 \quad \text{for some } \psi \in \mathcal{K}.$$

Indeed, we define

$$\begin{aligned} M_0 &= \{\mu \in M \,:\, \psi(\mu) = 0 \quad \forall \psi \in \mathcal{K}\} \\ &= \{\mu \in M \,:\, \psi(\mu) = 0 \quad \forall \psi \in \mathcal{L}\} \\ &= \{\mu \in M \,:\, \langle x, \mu\rangle\, \mu) = 0 \quad \forall x \in L\} \\ &= M - L \end{aligned}$$

is indeed a subspace of M. Note that

$$M - M_0 = L = \{cv_{\psi} \,:\, \psi \in \mathcal{L}\}$$

The hypothesis $\mu \in M_0$ fails if $\psi(\mu) \neq 0$ for some $\psi \in \mathcal{L}$. The canonical estimator of $\psi(\mu)$ is the GME $\hat{\psi}(Y)$. If $\hat{\psi}(Y)$ is significantly different from zero at level α

according to Theorem 3.1, then 0 is not in the $100(1-\alpha)\%$ confidence interval $[\hat{\psi}(Y) \pm S_{\dim(L),\dim\left(M^\perp\right)}(\alpha)\hat{\sigma}_{\hat{\psi}}]$. Equivalently,

$$\mathbb{P}_{\mu,\sigma^2}\left(|\hat{\psi}(Y)| \geq S_{\dim(L),\dim\left(M^\perp\right)}(\alpha)\hat{\sigma}_{\hat{\psi}}\right) \geq 1-\alpha.$$

Indeed, following the same argument as in the proof of Theorem 3.1, we get that $\hat{\psi}(Y)$ is SDFZ (statistically different from zero) at level α for some $\psi \in \mathcal{L}$ iff

$$\begin{aligned}
&\Leftrightarrow \sup_{\psi\in\mathcal{L}} |\hat{\psi}(Y)|^2/\hat{\sigma}^2_{\hat{\psi}} \geq \dim(L)\, \mathcal{F}_{\dim(L),\dim\left(M^\perp\right)}(\alpha) \\
&\Leftrightarrow \frac{\|P_L(Y)\|^2/\dim(L)}{\|P_{M^\perp}(Y)\|^2/\dim(M^\perp)} \geq \mathcal{F}_{\dim(M-M_0),\dim\left(M^\perp\right)}(\alpha) \\
&\Leftrightarrow \text{the size } \alpha\ F\text{-test rejects the hypothesis } \mathbf{H}_0 : \mu \in M_0 \ .
\end{aligned}$$