1 Warm-up

Let X be a random vector in $\mathbb{R}^d$ with mean $\mu \in \mathbb{R}^d$ and covariance matrix $\Sigma \in \mathbb{R}^{d\times d}$ and A a symmetric matrix in $\mathbb{R}^{d\times d}$. Then,

$$\mathbb{E}[X^\top AX] = \mu^\top A\mu + \mathrm{Trace}(A\Sigma)\,.$$

2 Student's t-statistics

We assume that for all $1 \leqslant i \leqslant n$, $Y_i = X_i^\top \theta_\star + \varepsilon_i$ for some unknown $\theta_\star \in \mathbb{R}^d$ where the $(\varepsilon_i)_{1\leqslant i\leqslant n}$ are i.i.d. random variables with distribution $\mathcal{N}(0, \sigma_*^2)$. Let $\varepsilon \in \mathbb{R}^n$ be the random vector such that for all $1 \leqslant i \leqslant n$, the i-th component of ε is ε_i. The model is then written $Y = X\theta_\star + \varepsilon$. Assume that X has full rank and that $\widehat{\theta}_n = (X^\top X)^{-1}X^\top Y$ and $\widehat{\sigma}_n^2 = \|Y - X\widehat{\theta}_n\|^2/(n-d)$.

1. For all $1 \leqslant j \leqslant d$, show that

$$\frac{\widehat{\theta}_{n,j} - \theta_{\star,j}}{\widehat{\sigma}_n\sqrt{(X^TX)^{-1}_{j,j}}} \sim \mathcal{S}(n-d)\,,$$

   where $\mathcal{S}(n-d)$ is the Student's t-distribution with $n-d$ degrees of freedom, i.e. the law of $X/\sqrt{Y/(n-d)}$ where $X \sim \mathcal{N}(0,1)$ is independent of $Y \sim \chi^2(n-d)$.

2. Provide a confidence interval with confidence level $1-\alpha$ for $\theta_{\star,j}$.

3 Random design

Consider the regression model given by

$$Y = X\theta_\star + \varepsilon\,,$$

where $X \in \mathbb{R}^{n\times d}$, the $(\varepsilon_i)_{1\leqslant i\leqslant n}$ are i.i.d. centered Gaussian random variables with variance $\sigma_\star^2$ and independent of $(X_i)_{1\leqslant i\leqslant n}$ which are assumed to be random. Assume that $X^\top X$ has full rank and that $\theta_\star$ is estimated by

$$\widehat{\theta}_n = (X^\top X)^{-1}X^\top Y\,.$$

1. Compute the excess risk $\mathsf{R}(\theta) - \mathsf{R}(\theta_\star)$, where $\mathsf{R}(\theta) = n^{-1}\mathbb{E}[\|Y - X^\top\theta\|_2^2]$.
2. Compute then the excess risk $\mathbb{E}[\mathsf{R}(\widehat{\theta}_n) - \mathsf{R}(\theta_\star)]$.

4 Fisher statistics (bonus)

Consider the regression model given by

$$Y = X\theta_\star + \varepsilon\,,$$

where $X \in \mathbb{R}^{n\times d}$ and the $(\varepsilon_i)_{1\leqslant i\leqslant n}$ are i.i.d. centered Gaussian random variables with variance $\sigma_\star^2$. Assume that $X^\top X$ has full rank and that $\theta_\star$ and $\sigma_\star^2$ are estimated by

$$\widehat{\theta}_n = (X^\top X)^{-1}X^\top Y \quad\text{and}\quad \widehat{\sigma}_n^2 = \frac{\|Y - X\widehat{\theta}_n\|^2}{n-d}\,.$$

1. Let L be a $\mathbb{R}^{q\times d}$ matrix with rank $q \leqslant d$. Show that

$$\frac{(\widehat{\theta}_n - \theta_\star)^\top L^\top (L(X^\top X)^{-1}L^\top)^{-1}L(\widehat{\theta}_n - \theta_\star)}{q\widehat{\sigma}_n^2} \sim \mathcal{F}(q, n-d)\,,$$

   where $\mathcal{F}(q, n-d)$ is the Fisher distribution with q and $n-d$ degrees of freedom, i.e. the law of $(X/q)/(Y/(n-d))$ where $X \sim \chi^2(q)$ is independent of $Y \sim \chi^2(n-d)$.

2. Using the previous question, build a confidence region with confidence level $1-\alpha \in (0,1)$ for $\theta_\star$.