SD-TSIA204
Properties of Ordinary Least Squares

Ekhine Irurozki
Télécom Paris

Model I : The fixed design model

$$
\begin{aligned}
y_i &= \theta_0^\star + \sum_{k=1}^{p} \theta_k^\star x_{i,k} + \varepsilon_i \\
x_i^\top &= (1, x_{i,1}, \ldots, x_{i,p}) \in \mathbb{R}^{p+1} \\
\varepsilon_i &\overset{i.i.d}{\sim} \varepsilon, \text{ for } i = 1, \ldots, n \\
\mathbb{E}(\boldsymbol{\varepsilon}) &= 0, \text{ Var}(\epsilon) = \sigma^2
\end{aligned}
$$

- x_i is deterministic
- σ^2 is called the noise level

Example :

- Physical experiment when the analyst is choosing the design *e.g.*,temperature of the experiment
- Some features are not random *e.g.*,time, location.

Model I with Gaussian noise : The fixed design Gaussian model

$$
\begin{aligned}
y_i &= \theta_0^\star + \sum_{k=1}^{p} \theta_k^\star x_{i,k} + \varepsilon_i \\
x_i^\top &= (1, x_{i,1}, \dots, x_{i,p}) \in \mathbb{R}^{p+1} \\
\varepsilon_i &\overset{i.i.d}{\sim} \mathcal{N}(0, \sigma^2), \text{ for } i = 1, \dots, n
\end{aligned}
$$

- Parametric model : specified by the two parameters $(\boldsymbol{\theta}, \sigma)$
- Strong assumption

Model II : The random design model

$$
\begin{aligned}
y_i &= \theta_0^\star + \sum_{k=1}^{p} \theta_k^\star x_{i,k} + \varepsilon_i \\
x_i^\top &= (1, x_{i,1}, \dots, x_{i,p}) \in \mathbb{R}^{p+1} \\
(\varepsilon_i, x_i) &\overset{i.i.d}{\sim} (\varepsilon, x), \text{ for } i = 1, \dots, n \\
\mathbb{E}(\varepsilon|x) &= 0, \ \mathrm{Var}(\varepsilon|x) = \sigma^2
\end{aligned}
$$

Rem: here, the features are modelled as random (they might also suffer from some noise)

The ordinary least squares (OLS) estimator

$$\hat{\boldsymbol{\theta}} \in \underset{\boldsymbol{\theta} \in \mathbb{R}^{p+1}}{\arg\min} \sum_{i=1}^{n} \left(y_i - \theta_0 - \sum_{k=1}^{p} \theta_k x_{i,k} \right)^2$$

How to deal with these two models ?

- The estimator is the same for both models
- The mathematics involved are different for each case
- The study of the fixed design case is easier as many closed formulas are available
- The two models lead to the same estimators of the variance σ^2

The OLS estimator, $\hat{\boldsymbol{\theta}} = (X^\top X)^{-1} X^\top Y$, how good it is ?

$$
\begin{aligned}
\hat{\boldsymbol{\theta}} &= \boldsymbol{\theta} + (X^\top X)^{-1} X^\top \boldsymbol{\varepsilon} \\
\hat{\boldsymbol{\theta}} &\sim N(\boldsymbol{\theta}, \sigma^2 (X^\top X)^{-1})
\end{aligned}
\tag{1}
$$

Its unbiased when $\mathbb{E}(\boldsymbol{\varepsilon}) = 0$

Its not very useful in practice since σ is not known

Exercise: Give the proof for Eq.(1). How is θ_i distributed ?

Expectation and covariance

$$\begin{aligned}\mathbb{E}[\hat{\boldsymbol{\theta}}] &= \mathbb{E}[\boldsymbol{\theta}^* + (X^\top X)^{-1}X^\top \boldsymbol{\varepsilon}] \\ &= \boldsymbol{\theta}^* + (X^\top X)^{-1}X^\top \mathbb{E}[\boldsymbol{\varepsilon}] \\ &= \boldsymbol{\theta}^*\end{aligned} \tag{2}$$

Under model I, whenever the matrix X has full rank, we have

$$\begin{aligned}Cov(\hat{\boldsymbol{\theta}}) &= Cov(\boldsymbol{\theta}^* + (X^\top X)^{-1}X^\top \boldsymbol{\varepsilon}) \\ &= Cov((X^\top X)^{-1}X^\top \boldsymbol{\varepsilon}) \\ &= ((X^\top X)^{-1}X^\top)Cov(\boldsymbol{\varepsilon})((X^\top X)^{-1}X^\top)^\top \\ &= (X^\top X)^{-1}X^\top Cov(\boldsymbol{\varepsilon})X(X^\top X)^{-1} \\ &= (X^\top X)^{-1}X^\top \sigma^2 I X(X^\top X)^{-1} \\ &= \sigma^2(X^\top X)^{-1}X^\top X(X^\top X)^{-1} = \sigma^2(X^\top X)^{-1}\end{aligned} \tag{3}$$

Bias

Proposition: Under model I, whenever the matrix X has full rank, the least squares estimator is unbiased, i.e.,

$$\mathbb{E}(\hat{\boldsymbol{\theta}}) = \boldsymbol{\theta}^{\star}$$

Proof :

$$\begin{aligned}
B =& \mathbb{E}(\hat{\boldsymbol{\theta}}) - \boldsymbol{\theta}^{\star} = \mathbb{E}((X^{\top}X)^{-1}X^{\top}\mathbf{y}) - \boldsymbol{\theta}^{\star} \\
B =& \mathbb{E}((X^{\top}X)^{-1}X^{\top}(X\boldsymbol{\theta}^{\star} + \boldsymbol{\varepsilon})) - \boldsymbol{\theta}^{\star} \\
B =& (X^{\top}X)^{-1}X^{\top}X\boldsymbol{\theta}^{\star} + (X^{\top}X)^{-1}X^{\top}\mathbb{E}(\boldsymbol{\varepsilon}) - \boldsymbol{\theta}^{\star} = 0
\end{aligned}$$

The trace of a matrix

Let $A \in \mathbb{R}^{n\times n}$ denote a matrix. The **trace** of A is the sum of the diagonal elements of A and is denoted by $\mathrm{tr}(A)$:

$$\mathrm{tr}(A) = \sum_{i=1}^{n} A_{i,i}$$

Several properties :

- $\mathrm{tr}(A) = \mathrm{tr}(A^\top)$
- For any $A, B \in \mathbb{R}^{n\times n}$, and $\alpha \in \mathbb{R}$, $\mathrm{tr}(\alpha A + B) = \alpha\,\mathrm{tr}(A) + \mathrm{tr}(B)$ (linearity)
- $\mathrm{tr}(A^\top A) = \sum_{i=1}^{n}\sum_{j=1}^{n} A_{i,j}^2 := \|A\|_F^2$
- For any $A, B \in \mathbb{R}^{n\times n}$, $\mathrm{tr}(AB) = \mathrm{tr}(BA)$
- $\mathrm{tr}(PAP^{-1}) = \mathrm{tr}(A)$, hence if A is diagonalisable, the trace is the sum of the eigenvalues
- If H is an orthogonal projector $\mathrm{tr}(H) = \mathrm{rank}(H)$

Estimation risk $R(\boldsymbol{\theta}^\star, \hat{\boldsymbol{\theta}}) = \mathbb{E}\|\boldsymbol{\theta}^\star - \hat{\boldsymbol{\theta}}\|^2$

Under model I, whenever the matrix X has full rank, we have

$$R(\boldsymbol{\theta}^\star, \hat{\boldsymbol{\theta}}) = \mathbb{E}\left[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star)^\top (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star)\right] = \sigma^2 \operatorname{tr}\left((X^\top X)^{-1}\right)$$

<u>Proof</u> :

$$\begin{aligned}
R(\boldsymbol{\theta}^\star, \hat{\boldsymbol{\theta}}) &= \mathbb{E}\left[(\hat{\boldsymbol{\theta}} - \mathbb{E}\hat{\boldsymbol{\theta}})^\top (\hat{\boldsymbol{\theta}} - \mathbb{E}\hat{\boldsymbol{\theta}})\right] = \mathbb{E}\left[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star)^\top (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star)\right] \\
&= \mathbb{E}\left[((X^\top X)^{-1} X^\top (X\boldsymbol{\theta}^\star + \boldsymbol{\varepsilon}) - \boldsymbol{\theta}^\star)^\top ((X^\top X)^{-1} X^\top (X\boldsymbol{\theta}^\star + \boldsymbol{\varepsilon}) - \boldsymbol{\theta}^\star)\right] \\
&= \mathbb{E}\left[((X^\top X)^{-1} X^\top \boldsymbol{\varepsilon})^\top ((X^\top X)^{-1} X^\top \boldsymbol{\varepsilon})\right] = \mathbb{E}(\boldsymbol{\varepsilon}^\top X (X^\top X)^{-2} X^\top \boldsymbol{\varepsilon}) \\
&= \operatorname{tr}[\mathbb{E}(\boldsymbol{\varepsilon}^\top X (X^\top X)^{-1} (X^\top X)^{-1} X^\top \boldsymbol{\varepsilon})] \text{ (thx to } \operatorname{tr}(u^\top u) = u^\top u) \\
&= \mathbb{E}\left(\operatorname{tr}\left[(X^\top X)^{-1} X^\top \boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top X (X^\top X)^{-1}\right]\right) \\
&= \operatorname{tr}[(X^\top X)^{-1} X^\top \mathbb{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top) X (X^\top X)^{-1}] \\
&= \sigma^2 \operatorname{tr}((X^\top X)^{-1})
\end{aligned}$$

Prediction risk (normalized) $R_{\text{pred}}(\boldsymbol{\theta}^\star, \hat{\boldsymbol{\theta}}) = \mathbb{E}\|X\boldsymbol{\theta}^\star - \hat{\mathbf{y}}\|^2/n$

Under model I, whenever the matrix X has full rank, we have

$$R_{\text{pred}}(\boldsymbol{\theta}^\star, \hat{\boldsymbol{\theta}}) = \mathbb{E}\left[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star)^\top \left(\frac{X^\top X}{n}\right)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star)\right] = \sigma^2 \frac{\operatorname{rank}(X)}{n}$$

Because X has full rank, $\operatorname{rank}(X) = p + 1$.

<u>Proof</u> : As before

$$\begin{aligned} n \cdot R_{\text{pred}}(\boldsymbol{\theta}^\star, \hat{\boldsymbol{\theta}}) &= \mathbb{E}\left[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star)^\top (X^\top X)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star)\right] \\ &= \mathbb{E}(\boldsymbol{\varepsilon}^\top X(X^\top X)^{-1}(X^\top X)(X^\top X)^{-1}X^\top \boldsymbol{\varepsilon}) \\ &= \mathbb{E}(\boldsymbol{\varepsilon}^\top X(X^\top X)^{-1}X^\top \boldsymbol{\varepsilon}) \\ &= \operatorname{tr}[\mathbb{E}(\boldsymbol{\varepsilon}^\top H \boldsymbol{\varepsilon})] = \operatorname{tr}[\mathbb{E}(\boldsymbol{\varepsilon}^\top H^\top H \boldsymbol{\varepsilon})] \\ &= \operatorname{tr}[\mathbb{E}(H\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top H^\top)] = \operatorname{tr}\left(H\mathbb{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top)H^\top\right) \\ &= \sigma^2 \operatorname{tr}(H) = \sigma^2 \operatorname{rank}(H) = \sigma^2 \operatorname{rank}(X) \end{aligned}$$

Best linear unbiased estimator

Under the fixed design model, among all the unbiased linear estimators AY, $\hat{\theta}_n$ is the one with minimal variance, i.e.,

$$\operatorname{cov}(\hat{\theta}_n) \leqslant \operatorname{cov}(AY),$$

with equality if and only if $A = (X^T X)^{-1} X^T$.

proof First note that AY is unbiased if and only if $(A - (X^T X)^{-1} X^T) X \theta^\star = 0$ for all $\theta^\star$, equivalently, $BX = 0$ with $B = (A - (X^T X)^{-1} X^T)$. Consequently, using that $E[\epsilon\epsilon^T] = \sigma^2 I_n$, $\operatorname{cov}(BY, \hat{\theta}_n) = 0$. Then, just write

$$\begin{aligned}
\operatorname{cov}(AY) &= \operatorname{cov}(BY + \hat{\theta}_n) \\
&= \operatorname{cov}(BY) + \operatorname{cov}(\hat{\theta}_n) \\
&= \sigma^2 B B^T + \operatorname{cov}(\hat{\theta}_n) \geqslant \operatorname{cov}(\hat{\theta}_n).
\end{aligned}$$

The previous inequality is an equality if and only if $B = 0$.

Maximum Likelihood Estimation (MLE)

Explanation of the principle of maximum likelihood :

- Maximum Likelihood Estimation (MLE) is a widely used method to estimate unknown parameters.
- It is based on the idea of finding the parameter values that make the observed data most probable under a given statistical model.

Illustration of Maximum Likelihood Estimation (MLE)

MLE as finding the parameter value that maximizes likelihood :

- Consider a statistical model with unknown parameter θ and observed data X.
- The likelihood function $L(\theta; X)$ measures how probable the data is under the parameter θ as a product of their densities, $L(\theta; X) = \displaystyle\prod_{k=1}^{n} p(X_k; \theta)$.
- MLE seeks to find $\hat{\theta}$ that maximizes $L(\theta; X)$:
$$\hat{\theta} = \arg\max_{\theta} L(\theta; X)$$

Example : MLE for Coin Flip Model

Coin Flip Model : Probability of getting heads in a coin flip

- Model : Bernoulli
- Parameter : p_H (probability of getting heads, $0 \leqslant p_H \leqslant 1$)
- Fair coin : $p_H = 0.5$

Observations : "HH" (two heads in a row)

Likelihood for $p_H = 0.5$: $L(p_H = 0.5 \mid \text{HH}) = 0.5^2 = 0.25$

Likelihood for $p_H = 0.3$: $L(p_H = 0.3 \mid \text{HH}) = 0.3^2 = 0.09$

General Observation : For each observed value $s \in S$, we can calculate the corresponding likelihood as $\prod_{s \in S} p(s; \theta)$.

Note : Likelihoods need not integrate or sum to one over the parameter space.

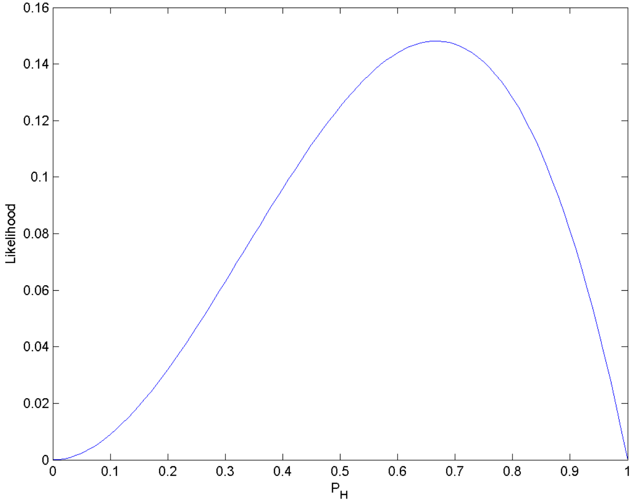

Figure – Likelihood function for different p_H values when we observe HHT

Definition of Likelihood Function and Log-Likelihood Function

Likelihood Function :

- Measures how well the observed data fit the model parameterized by θ.
- Denoted by $L(\theta; X)$, where θ is the parameter and X is the observed data.
- Provides a probability distribution for the observed data given the parameter.
- For independent and identically distributed random variables, it will be the product of univariate density functions :

$$L(\theta; X) = \prod_{k=1}^{n} p(X_k; \theta) \ .$$

Log-Likelihood Function :

- Definition : $\mathcal{L}(\theta; X) = \log L(\theta; X)$.
- Log-transform simplifies calculations and often leads to mathematical convenience.
- Useful for optimization techniques to find the MLE.
- The MLE can be found by maximizing the log-likelihood.

Log-Likelihood and Maximum

In practice, it is often convenient to work with the natural logarithm of the likelihood function, called the log-likelihood :

$$\mathcal{L}(\theta; \mathbf{y}) = \ln L_n(\theta; \mathbf{y}).$$

Since the logarithm is a monotonic function, the maximum of $\mathcal{L}(\theta; \mathbf{y})$ occurs at the same value of θ as does the maximum of $\mathcal{L}_n$. If $\mathcal{L}(\theta; \mathbf{y})$ is differentiable in Θ, the necessary conditions for the occurrence of a maximum (or a minimum) are :

$$\frac{\partial \mathcal{L}}{\partial \theta_1} = 0, \quad \frac{\partial \mathcal{L}}{\partial \theta_2} = 0, \quad \ldots, \quad \frac{\partial \mathcal{L}}{\partial \theta_k} = 0.$$

MLE for Different Distributions. Exercise : give the proofs

Bernoulli Distribution : MLE for success probability p :

$$\hat{p} = \frac{\text{number of successes}}{\text{total trials}}$$

Normal Distribution : MLE for mean μ and variance σ^2 :

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

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} (x_i - \hat{\mu})^2$$

Poisson Distribution : MLE for rate parameter λ : $\hat{\lambda} = \frac{1}{n}\sum_{i=1}^{n} x_i$

Exponential Distribution : MLE for rate parameter λ : $\hat{\lambda} = \frac{n}{\sum_{i=1}^{n} x_i}$

Multinomial Distribution : MLE for probabilities $p_1, p_2, \ldots, p_k$ of k categories in n trials : $\hat{p}_i = \frac{n_i}{n}$, where n_i is the count of category i

Poisson and Exponential Distributions

Poisson Distribution

- Discrete probability distribution.
- Models the number of rare events in a fixed interval.
- Parameter : λ (average rate of events).
- Probability mass function (PMF) :

$$P(X = k) = \frac{e^{-\lambda}\lambda^k}{k!}$$

- Mean : λ
- Variance : λ

Exponential Distribution

- Continuous probability distribution.
- Models the time between rare events.
- Parameter : λ (rate parameter).
- Probability density function (PDF) :

$$f(x|\lambda) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \geqslant 0 \\ 0 & \text{if } x < 0 \end{cases}$$

- Mean : $\frac{1}{\lambda}$
- Variance : $\frac{1}{\lambda^2}$

Estimation of the noise level

- An estimator of the noise level σ^2 is given by

$$\boxed{\frac{1}{n}\|\mathbf{y} - X\hat{\boldsymbol{\theta}}\|_2^2}$$

- Another estimator which is unbiased is defined by

$$\boxed{\hat{\sigma}^2 = \frac{1}{n - \mathrm{rank}(X)}\|\mathbf{y} - X\hat{\boldsymbol{\theta}}\|_2^2}$$

To show that this estimator is unbiased we need to give more properties of the Hat matrix and Cochran's lemma

Properties of the Hat matrix

Rem: the Hat matrix is defined as $H = X(X^\top X)^{-1}X^\top$

Proposition:

1. H is an orthogonal projection matrix
2. $(I - H)$ is an orthogonal projection matrix
3. $HX = X$
4. $(I - H)X = 0$

Statistical background, χ_k^2 distribution

Let $Z \sim \mathcal{N}(0,1)$, then the sum of their squares, $Q = \sum_{i=1}^{k} Z_i^2$, is distributed according to the chi-squared distribution with k degrees of freedom. This is denoted as $Q \sim \chi_k^2$. The chi-squared distribution has one parameter : a positive integer k that specifies the number of degrees of freedom (the number of random variables being summed, $_is$).

If $a \sim \chi_k^2$ then $\mathbb{E}[a] = k$ and $Var(a) = 2k$

Cochran's lemma

Let $\boldsymbol{\varepsilon} \sim N(0, \sigma^2 I)$ and $\hat{\sigma}^2 = \frac{1}{n-p-1}\|\mathbf{y} - X\hat{\boldsymbol{\theta}}\|_2^2$ and X full rank. Then

$$\begin{aligned}
&\hat{\theta}_n \text{ and } \hat{\sigma}_n^2 \text{ are independent,} \\
&\hat{\theta}_n \sim N\left(\theta^\star, \sigma^2 (X^T X)^{-1}\right), \\
&(n-p-1)\left(\frac{\hat{\sigma}_n^2}{\sigma^2}\right) \sim \chi^2_{n-p-1}.
\end{aligned} \tag{4}$$

Estimation of the noise level, $\hat{\sigma}^2$ is unbiased

Under model I, whenever the matrix X has full rank, we have

$$\mathbb{E}\hat{\sigma}^2 = \sigma^2$$

<u>Proof sketch</u> :

$$\|\mathbf{y} - \hat{\mathbf{y}}\|_2^2 = \mathbf{y}^\top(\mathrm{Id}_n - H)\mathbf{y} = \boldsymbol{\varepsilon}^\top(\mathrm{Id}_n - H)\boldsymbol{\varepsilon}$$

Gaussian case : if $\boldsymbol{\varepsilon}_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$, then $\|\mathbf{y} - \hat{\mathbf{y}}\|_2^2 \sim \chi^2$ à $n - \mathrm{rank}(X)$ degrés de liberté

Exercise: Complete the proof

Heteroscedasticity

Model I and Model II are homoscedastic models, *i.e.,*we assume that the noise level σ^2 does not depend on x_i

<u>Heteroscedastic Model</u> : we allow σ^2 to change with the observation i, we denote by $\sigma_i^2 > 0$ the associated variance

$$\hat{\boldsymbol{\theta}} \in \mathop{\arg\min}_{\boldsymbol{\theta} \in \mathbb{R}^{p+1}} \sum_{i=1}^{n} \left(\frac{y_i - \langle \boldsymbol{\theta}, x_i \rangle}{\sigma_i} \right)^2 = \mathop{\arg\min}_{\boldsymbol{\theta} \in \mathbb{R}^{p+1}} (y - X\boldsymbol{\theta})^\top \Omega (y - X\boldsymbol{\theta})$$

with $\Omega = \operatorname{diag}(\frac{1}{\sigma_1^2}, \dots, \frac{1}{\sigma_n^2})$

Exercise: give a closed formula for $\hat{\boldsymbol{\theta}}$ when $X^\top \Omega X$ has full rank

---

Exercise: give a necessary and sufficient condition for $X^\top \Omega X$ to be invertible

Bias and variance

Proposition: Under model II, whenever the matrix $X = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^\top$ has full rank, we have

$$\mathbb{E}(\hat{\boldsymbol{\theta}} \mid X) = \boldsymbol{\theta}^\star$$
$$\mathrm{Var}(\hat{\boldsymbol{\theta}} \mid X) = (X^\top X)^{-1}\sigma^2$$

Proof : The same as in the case of fixed design with the conditional expectation

Rem:We cannot compute the $\mathbb{E}(\hat{\boldsymbol{\theta}})$ nor $\mathrm{Var}(\hat{\theta})$ because the matrix X has full rank is now random !

Rem:One solution is to rely on asymptotic convergence

Asymptotics of $\hat{\boldsymbol{\theta}}$

Under model II, whenever the covariance matrix $\mathrm{cov}(X)$ has full rank, we have

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star) \xrightarrow{\mathrm{d}} \mathcal{N}(0, \sigma^2 S^{-1})$$

with $S = \mathbb{E}[\mathbf{x}\mathbf{x}^\top]$

Outline of the proof : It could happen that $\hat{\boldsymbol{\theta}}$ is not uniquely defined, so we put

$$\hat{\boldsymbol{\theta}} = \left(X^\top X\right)^+ X^\top Y$$

where A^+ is the generalized inverse of A

- With high probability, we have that $X^\top X$ is invertible because $\frac{X^\top X}{n} = \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^\top$ goes to S

Asymptotics

Outline of the proof :

- As a consequence, in the asymptotics we can replace $\left(X^\top X\right)^+$ by $\left(X^\top X\right)^{-1}$ (that we shall admit)

Then we use that

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star) = \left(\frac{X^\top X}{n}\right)^{-1} \left(\frac{X^\top \epsilon}{\sqrt{n}}\right)$$

- The term on the right $\frac{X^\top \varepsilon}{\sqrt{n}}$ converges to $\mathcal{N}(0, \mathbb{E}[\mathbf{x}\mathbf{x}^\top]\sigma^2)$ in distribution
- The term on the left $\left(\frac{X^\top X}{n}\right)^{-1}$ goes to S^{-1} in probability

Asymptotics

- In the random design model, since closed formulas for the bias and variance of $\boldsymbol{\theta}$ are lacking ; Asymptotics is used to validate the procedure and to build-up the variance estimator

By the previous Proposition, the **variance** to estimate is

$$\sigma^2 S^{-1}$$

a natural “Plug-in” estimator is

$$\hat{\sigma}^2 \hat{S}_n^+$$

with $\hat{\sigma}^2 = \frac{1}{n-\operatorname{rank}(X)} \|\mathbf{y} - X\hat{\boldsymbol{\theta}}\|_2^2$

<u>Rem</u>:It coincides with the estimator in the case of fixed design

Variance estimation

Noise level is conditionally unbiased : Under model II, whenever the matrix $X = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^\top$ has full rank, we have

$$\mathbb{E}(\hat{\sigma}^2 \mid X) = \sigma^2$$

Exercise: Write the proof

Convergence of the variance estimator : Under model II, if the covariance matrix $\mathrm{cov}(X)$ has full rank, we have

$$\hat{\sigma}^2 \hat{S}_n^+ \to \sigma^2 S^{-1}$$

in probability

Qualitative variables

A variable is qualitative, when its state space is discrete (non-necessarily numeric)

Exemple : colors, gender, cities, etc.

Classically : "One-hot encoder" consists in representing a qualitative variable with several dummy variables (valued in $\{0, 1\}$)

If each x_i is valued in $a_1, \ldots, a_K$, we define the following K explanatory variables :
$\forall k \in [\![1, K]\!], \mathbb{1}_{a_k} \in \mathbb{R}^n$ is given by

$$\forall i \in [\![1, n]\!], \quad (\mathbb{1}_{a_k})_i = \begin{cases} 1, \text{ if } x_i = a_k \\ 0, \text{ else} \end{cases}$$

Examples

Binary case : M/F, yes/no, I like it/I don't.

Client	Gender
1	H
2	F
3	H
4	F
5	F

$$\longrightarrow \begin{pmatrix} F & H \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 1 & 0 \end{pmatrix}$$

General case : colors, cities, etc.

Client	Colors
1	Blue
2	Blanc
3	Red
4	Red
5	Blue

$$\longrightarrow \begin{pmatrix} \text{Blue} & \text{Blanc} & \text{Red} \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}$$

Somme difficulties

Correlations : $\sum_{k=1}^{K} \mathbb{1}_{a_k} = \mathbf{1}_n$! We can drop-off one modality (*e.g.*,`drop_first=True` dans `get_dummies` de `pandas`)

Without intercept, with all modalities : $X = [\mathbb{1}_{a_1}, \dots, \mathbb{1}_{a_K}]$. If $x_{n+1} = a_k$ then $\hat{y}_{n+1} = \hat{\boldsymbol{\theta}}_k$

With intercept, with one less modality : $X = [\mathbf{1}_n, \mathbb{1}_{a_2}, \dots, \mathbb{1}_{a_K}]$, dropping-off the first modality

If $x_{n+1} = a_k$ then $\hat{y}_{n+1} = \begin{cases} \hat{\boldsymbol{\theta}}_0, & \text{if } k = 1 \\ \hat{\boldsymbol{\theta}}_0 + \hat{\boldsymbol{\theta}}_k, & \text{else} \end{cases}$

Rem: might give null column in Cross-Validation (if a modality is not present in a CV-fold)

Rem: penalization might help (*e.g.*,Lasso, Ridge)

What if $n < p$?

Many of the things presented before need to be adapted

For instance : if $\operatorname{rank}(X) = n$, then $H = \mathrm{Id}_n$ and $\hat{\mathbf{y}} = X\hat{\boldsymbol{\theta}} = \mathbf{y}$!

The vector space generated by the columns $[\mathbf{x}_0, \ldots, \mathbf{x}_p]$ is $\mathbb{R}^n$, making the observed signal and predicted signal are **identical**

Rem: typical kind of problem in large dimension (when p is large)

Possible solution : variable selection, *cf.*Lasso and greedy methods (coming soon)

Web sites and books

- Python Packages for OLS :
  statsmodels
  sklearn.linear_model.LinearRegression
- McKinney (2012) about python for statistics
- Lejeune (2010) about the Linear Model
- Delyon (2015) Advanced course on regression
  https://perso.univ-rennes1.fr/bernard.delyon/regression.pdf