Class Notes

Econometrics of High-Dimensional Models - ECON231C

Mauricio Vargas-Estrada
Master in Quantitative Economics
University of California - Los Angeles

1 Markov's Inequality

Being X a random variable such that $X \geq 0$, then:

$$\mathbb{P}(X \geq t) \leq \frac{\mathbb{E}[X]}{t}, \quad \forall t > 0$$

Proof. Whe can rewrite the left-hand side of the inequality using the indicator function:

$$X \geq X\mathbf{1}_{\{X \geq t\}}$$

The left-hand side would be greater when $X < t$ and equal when $X \geq t$. Given that:

$$X \geq X\mathbf{1}_{\{X \geq t\}} \geq t\mathbf{1}_{\{X \geq t\}}$$

In this case, $X\mathbf{1}_{\{X \geq t\}} > t\mathbf{1}_{\{X \geq t\}}$ when $X > t$, and $X\mathbf{1}_{\{X \geq t\}} = t\mathbf{1}_{\{X \geq t\}}$ when $X \leq t$ because $X = t$ or the indicator function is zero.

Taking the expectation of the inequality:

$$\begin{aligned}
\mathbb{E}[X] &\geq \mathbb{E}\left[X\mathbf{1}_{\{X \geq t\}}\right] \geq \mathbb{E}\left[t\mathbf{1}_{\{X \geq t\}}\right] \\
\mathbb{E}[X] &\geq \mathbb{E}\left[X\mathbf{1}_{\{X \geq t\}}\right] \geq t\mathbb{E}\left[\mathbf{1}_{\{X \geq t\}}\right] \\
\frac{\mathbb{E}[X]}{t} &\geq \frac{\mathbb{E}\left[X\mathbf{1}_{\{X \geq t\}}\right]}{t} \geq \mathbb{E}\left[\mathbf{1}_{\{X \geq t\}}\right]
\end{aligned}$$

But $\mathbb{E}\left[\mathbf{1}_{\{X \geq t\}}\right] = \mathbb{P}(X \geq t)$, so:

$$\frac{\mathbb{E}[X]}{t} \geq \mathbb{P}(X \geq t)$$

□

2 Chevyshev's Inequality

Given a random variable X with mean μ and variance σ^2, then:

$$\mathbb{P}\left(|X-\mu| \geq t\right) \leq \frac{\sigma^2}{t^2}, \quad \forall t > 0$$

Proof. We are going to use the fact that a strictly increasing function of a random variable does not change the probability of an event. Let $Y = (|X-\mu|)^2$. Then,

$$\mathbb{P}\left(|X-\mu|^2 \geq t^2\right) = \mathbb{P}\left(Y \geq t^2\right)$$

Using Markov's inequality, we have:

$$\mathbb{P}\left(Y \geq t^2\right) \leq \frac{\mathbb{E}\left[Y\right]}{t^2}$$

Given a random variable Z, the variance of Z is $\mathrm{Var}\left(Z\right) = \mathbb{E}\left[(Z - \mathbb{E}\left[Z\right])^2\right]$. Therefore,

$$\begin{aligned}
\mathbb{P}\left((X-\mu)^2 \geq t^2\right) &\leq \frac{\mathbb{E}\left[(X-\mu)^2\right]}{t^2} \\
\mathbb{P}\left((X-\mu)^2 \geq t^2\right) &\leq \frac{\mathrm{Var}\left(X\right)}{t^2} \\
\mathbb{P}\left((X-\mu)^2 \geq t^2\right) &\leq \frac{\sigma^2}{t^2}
\end{aligned}$$

□

3 Weak Law of Large Numbers

Given a collection of i.i.d. random variables $\{X_i\}_{i=1}^n$, with mean μ and variance σ^2.
Defining $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$, the weak law of large numbers states that for any $\epsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}\left(|\bar{X}_n - \mu| \geq \epsilon\right) = 0$$

or equivalently,

$$\bar{X}_n \xrightarrow{\mathbb{P}} \mu$$

Proof. Calculating the variance of $\bar{X}_n$,

$$\begin{aligned}
\operatorname{Var}\left(\bar{X}_n\right) &= \operatorname{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \\
&= \frac{1}{n^2}\sum_{i=1}^{n} \operatorname{Var}\left(X_i\right) \\
&= \frac{1}{n^2}\cdot n\sigma^2 \\
&= \frac{\sigma^2}{n}
\end{aligned}$$

Then, by Chebyshev's inequality,

$$\begin{aligned}
\mathbb{P}\left(|\bar{X}_n - \mu| \geq \epsilon\right) &\leq \frac{\operatorname{Var}\left(\bar{X}_n\right)}{\epsilon^2} \\
&= \frac{\sigma^2}{n\epsilon^2}
\end{aligned}$$

Taking the limit as $n \to \infty$,

$$\begin{aligned}
\lim_{n\to\infty} \mathbb{P}\left(|\bar{X}_n - \mu| \geq \epsilon\right) &\leq \lim_{n\to\infty} \frac{\sigma^2}{n\epsilon^2} \\
\lim_{n\to\infty} \mathbb{P}\left(|\bar{X}_n - \mu| \geq \epsilon\right) &\leq 0
\end{aligned}$$

□

4 Hoeffding's Inequality

If $\{X_i\}_{i=1}^n$ is a random sample from a distribution with mean μ such that, for a number [1] $a > 0$, we have:

$$|X_i - \mu| \leq a, \quad \forall i = 1, 2, \ldots, n$$

Then, for any $t > 0$, the following inequality holds:

$$\mathbb{P}\left(|\bar{X}_n - \mu| \geq t\right) \leq 2\exp\left(-\frac{nt^2}{2a^2}\right)$$

where $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ is the sample mean.

Proof. Lets define $Z_i = X_i - \mu, \forall i = 1, 2, \ldots, n$. Then, we have:

[1] In the field of statistics, it is common to consider bounded random variables, which naturally leads to the assumption that all moments exist. However, in econometrics, it's often more pragmatic to soften this assumption, focusing instead on the existence of only a select subset of moments. This approach allows for greater flexibility in dealing with real-world data, where the behavior of economic variables can't always be neatly bounded, and full moment conditions may not hold.

$$\bar{Z}_n = \frac{1}{n}\sum_{i=1}^{n} Z_i = \bar{X}_n - \mu$$

and,

$$|Z_i| \leq a, \forall i = 1, 2, \ldots, n$$

Consider the events:

$$A = \{\bar{Z}_n \geq t\}$$
$$B = \{\bar{Z}_n \leq -t\}$$

then, the probability of a event $C = \{|\bar{Z}_n| \geq t\}$, can be written as:

$$\mathbb{P}(C) = \mathbb{P}(A \cup B) \leq \mathbb{P}(A) + \mathbb{P}(B)$$

by the union bound. Now, we can write:

$$\mathbb{P}\left(|\bar{Z}_n| \geq t\right) \leq \mathbb{P}\left(\bar{Z}_n \geq t\right) + \mathbb{P}\left(\bar{Z}_n \leq -t\right)$$
$$\mathbb{P}\left(|\bar{Z}_n| \geq t\right) \leq \mathbb{P}\left(\sum_{i=1}^{n} Z_i \geq nt\right) + \mathbb{P}\left(\sum_{i=1}^{n} Z_i \leq -nt\right)$$

for any $\lambda > 0$, we have [2]:

$$\mathbb{P}\left(|\bar{Z}_n| \geq t\right) \leq \mathbb{P}\left(\lambda\sum_{i=1}^{n} Z_i \geq \lambda nt\right) + \mathbb{P}\left(\lambda\sum_{i=1}^{n} Z_i \leq -\lambda nt\right)$$

and, by Markov's inequality:

$$\mathbb{P}\left(|\bar{Z}_n| \geq t\right) \leq \frac{\mathbb{E}\left[\exp(\lambda\sum_{i=1}^{n} Z_i)\right]}{\exp(\lambda nt)} + \frac{\mathbb{E}\left[\exp(-\lambda\sum_{i=1}^{n} Z_i)\right]}{\exp(-\lambda nt)}$$

since Z_i are independent and identically distributed, we can write:

[2] Given that $f(x) = \lambda x$ is a monotonically increasing function when $\lambda > 0$

$$\mathbb{P}\left(|\bar{Z}_n| \geq t\right) \leq \frac{\prod_{i=1}^{n} \mathbb{E}\left[\exp(\lambda Z_i)\right]}{\exp(\lambda n t)} + \frac{\prod_{i=1}^{n} \mathbb{E}\left[\exp(-\lambda Z_i)\right]}{\exp(-\lambda n t)}$$
$$\mathbb{P}\left(|\bar{Z}_n| \geq t\right) \leq \frac{\prod_{i=1}^{n} \mathbb{E}\left[\exp(\lambda Z_i)\right]}{\exp(\lambda n t)} + \frac{\prod_{i=1}^{n} \exp(-1)\mathbb{E}\left[\exp(\lambda Z_i)\right]}{\exp(-\lambda n t)}$$
$$\mathbb{P}\left(|\bar{Z}_n| \geq t\right) \leq \frac{2\prod_{i=1}^{n} \mathbb{E}\left[\exp(\lambda Z_i)\right]}{\exp(\lambda n t)}$$

Applying the Hoeffding's lemma [3] to the above expression, we have:

$$\mathbb{P}\left(|\bar{Z}_n| \geq t\right) \leq \frac{2\prod_{i=1}^{n} \exp\left(\frac{\lambda^2 a^2}{2}\right)}{\exp(\lambda n t)}$$

simplifying the above expression:

$$\mathbb{P}\left(|\bar{Z}_n| \geq t\right) \leq \frac{2\exp\left(\frac{n\lambda^2 a^2}{2}\right)}{\exp(\lambda n t)}$$
$$\mathbb{P}\left(|\bar{Z}_n| \geq t\right) \leq 2\exp\left(\frac{n\lambda^2 a^2}{2} - \lambda n t\right)$$

Because the above inequality holds for any $\lambda > 0$, we can optimize the right-hand side with respect to λ.

$$\lambda^* = \arg\min_{\lambda > 0} \left\{ \frac{n\lambda^2 a^2}{2} - \lambda n t \right\}$$

Calculating the F.O.C. with respect to λ, we get:

$$n a^2 \lambda^* - n t = 0 \Rightarrow \lambda^* = \frac{t}{a^2}$$

Substituting λ^* back into the inequality:

$$\mathbb{P}\left(|\bar{Z}_n| \geq t\right) \leq 2\exp\left(\frac{n\left(\frac{t}{a^2}\right)^2 a^2}{2} - \frac{t}{a^2} n t\right)$$
$$\mathbb{P}\left(|\bar{Z}_n| \geq t\right) \leq 2\exp\left(\frac{nt^2}{2a^2} - \frac{nt^2}{a^2}\right)$$
$$\mathbb{P}\left(|\bar{Z}_n| \geq t\right) \leq 2\exp\left(-\frac{nt^2}{2a^2}\right)$$

[3] If X is a random variable such that $|X| \leq a$, then for any $\lambda > 0$, we have:

$$\mathbb{E}\left[\exp(\lambda X)\right] \leq \exp\left(\frac{\lambda^2 a^2}{2}\right)$$

replacing $\bar{Z}_n$ by $\bar{X}_n - \mu$:

$$\mathbb{P}\left(|\bar{X}_n - \mu| \geq t\right) \leq 2\exp\left(-\frac{nt^2}{2a^2}\right)$$

□

5 Maximal Inequality

Being $\{X_i\}_{i=1}^n$ a random sample, where $\dim X_i = p$, from a distribution with mean $\mu = [\mu_1, \cdots, \mu_p]'$, such that:

$$|X_{i,j} - \mu_j| \leq a, \quad , \forall i = 1, \cdots, n, \quad \forall j = 1, \cdots, p, \quad \forall a > 0$$

then:

$$\mathbb{P}\left(\max_{1\leq j\leq p}\left|\frac{1}{n}\sum_{i=1}^{n} X_{i,j} - \mu_j\right| \geq a\sqrt{2}\sqrt{\frac{\ln p + \ln\frac{2}{\varepsilon}}{n}}\right) \leq \varepsilon, \quad \forall \varepsilon \in (0,1)$$

$X_i = [X_{i,1}, \cdots, X_{i,p}]'$, where $X_{i,j}$ is the j-th component of the i-th random vector.

The term $\left|\frac{1}{n}\sum_{i=1}^{n} X_{i,j} - \mu_j\right|$ represent the deviation of the sample mean of the j-th component from the population mean of the j-th component.

The maximal inequality is a generalization of the weak law of large numbers, and it is used to bound the probability of the maximum deviation of the sample mean from the population mean.

Proof. Applying the union bound, we have:

$$\mathbb{P}\left(\max_{1\leq j\leq p}\left|\frac{1}{n}\sum_{i=1}^{n} X_{i,j} - \mu_j\right| \geq a\sqrt{2}\sqrt{\frac{\ln p + \ln\frac{2}{\varepsilon}}{n}}\right) \leq \sum_{j=1}^{p}\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_{i,j} - \mu_j\right| \geq a\sqrt{2}\sqrt{\frac{\ln p + \ln\frac{2}{\varepsilon}}{n}}\right)$$

By the Hoeffding's inequality, we have:

$$\sum_{j=1}^{p}\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_{i,j} - \mu_j\right| \geq a\sqrt{2}\sqrt{\frac{\ln p + \ln\frac{2}{\varepsilon}}{n}}\right) \leq \sum_{j=1}^{p} 2\exp\left(-\ln p + \ln\frac{2}{\varepsilon}\right)$$

Simplifying the expression, we have:

$$\mathbb{P}\left(\max_{1\leq j\leq p}\left|\frac{1}{n}\sum_{i=1}^{n}X_{i,j}-\mu_j\right|\geq a\sqrt{2}\sqrt{\frac{\ln p+\ln\frac{2}{\varepsilon}}{n}}\right)\leq\sum_{j=1}^{p}2\exp\left(-\ln p\right)\exp\left(-\ln\frac{2}{\varepsilon}\right)$$

$$\mathbb{P}\left(\max_{1\leq j\leq p}\left|\frac{1}{n}\sum_{i=1}^{n}X_{i,j}-\mu_j\right|\geq a\sqrt{2}\sqrt{\frac{\ln p+\ln\frac{2}{\varepsilon}}{n}}\right)\leq\sum_{j=1}^{p}2\exp\left(\ln\frac{1}{p}\right)\exp\left(\ln\frac{\varepsilon}{2}\right)$$

$$\mathbb{P}\left(\max_{1\leq j\leq p}\left|\frac{1}{n}\sum_{i=1}^{n}X_{i,j}-\mu_j\right|\geq a\sqrt{2}\sqrt{\frac{\ln p+\ln\frac{2}{\varepsilon}}{n}}\right)\leq\frac{2p\varepsilon}{2p}$$

$$\mathbb{P}\left(\max_{1\leq j\leq p}\left|\frac{1}{n}\sum_{i=1}^{n}X_{i,j}-\mu_j\right|\geq a\sqrt{2}\sqrt{\frac{\ln p+\ln\frac{2}{\varepsilon}}{n}}\right)\leq\varepsilon$$

□

Whe have proved that the rate of convergence of the maximal inequality is $\sqrt{\ln p/n}$, or in big-O notation:

$$\max_{1\leq j\leq p}\left|\frac{1}{n}\sum_{i=1}^{n}X_{i,j}-\mu_j\right|=O_p\left(\sqrt{\frac{\ln p}{n}}\right)$$

6 High-Dimensional Linear Regression

Consider the regression:

$$Y=\boldsymbol{X}'\boldsymbol{\beta}+\epsilon$$

Where Y is a scalar, $\boldsymbol{X}$ is a p-dimensional vector of regressors, $\boldsymbol{\beta}$ is a p-dimensional vector of coefficients, and ϵ is a scalar error term. And suppose we have a random sample:

$$(\boldsymbol{X}_1,Y_1),\cdots,(\boldsymbol{X}_n,Y_n)\sim\text{i.i.d.}(\boldsymbol{X},Y)$$

We are interested in case where p is large, meaning that: $p\sim n$, $p>n$ or $p>>n$.

Claim. OLS *linear estimator does not exist when* $p>n$.

Proof. The OLS linear regression estimator is given by:

$$\hat{\boldsymbol{\beta}}=(\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'\boldsymbol{Y}$$

Where $\mathbb{X}=(\boldsymbol{X_1},\cdots,\boldsymbol{X_n})'$ is the $n\times p$ design matrix, and $\boldsymbol{Y}=(Y_1,\cdots,Y_n)'$ is the $n\times 1$ vector of dependent variables.

Lemma. *If* $\mathbb{A}$ *is degenerated, then* $\mathbb{A}^{-1}$ *does not exist.*

Definition. *A matrix $\mathbb{A}$ is degenerated if there exists a non-zero vector $\boldsymbol{v}$ such that $\mathbb{A}\boldsymbol{v} = \mathbf{0}$.*

Proof. Asume that $\mathbb{A}$ is invertible, then there exist $\mathbb{A}\mathbb{A}^{-1} = \mathbb{I}_p$. If $\mathbb{A}$ is degenerated, then there exists a non-zero vector $\boldsymbol{v}$ such that $\mathbb{A}\boldsymbol{v} = \mathbf{0}$. Multiplying both sides by $\mathbb{A}^{-1}$ we get:

$$\mathbb{A}^{-1}\mathbb{A}\boldsymbol{v} = \mathbb{A}^{-1}\mathbf{0} \implies \boldsymbol{v} = \mathbf{0}$$

Which is a contradiction. Therefore, $\mathbb{A}$ is not invertible. □

Lemma. *The matrix $\mathbb{A} = \mathbb{X}'\mathbb{X}$ is degenerated when $p > n$.*

Proof. Consider the linear system of equations:

$$\mathbb{X}\boldsymbol{b} = \mathbf{0}, \quad \boldsymbol{b} \in \mathbb{R}^p$$

Where $\mathbb{X}$ is a $n \times p$ matrix. This system has a non unique solution when $p > n$. therefore, there existe a non-zero vector $\boldsymbol{b}$ such that $\mathbb{X}\boldsymbol{b} = \mathbf{0}$, meaning that $\mathbb{X}'\mathbb{X}$ is degenerated. □

From the previous lemma, we know that $\mathbb{X}'\mathbb{X}$ is degenerated when $p > n$. Therefore, $\mathbb{X}'\mathbb{X}$ is not invertible, and the OLS linear estimator does not exist. □

6.1 Expected Value of the OLS Linear Regression Estimator when $p > n$

Assuming that the linear model is homoscedastic, that is:

$$\mathbb{E}\left[\epsilon^2 | \boldsymbol{X}\right] = \sigma^2$$

then:

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left(\boldsymbol{X}_i'(\hat{\boldsymbol{\beta}}^{OLS} - \boldsymbol{\beta})\right)^2\right] = \frac{p\sigma^2}{n}$$

Proof. Given the linear regression in matrix form:

$$\boldsymbol{Y} = \mathbb{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

Where $\boldsymbol{Y} = (Y_1, \cdots, Y_n)'$ and $\boldsymbol{\epsilon} = (\epsilon_1, \cdots, \epsilon_n)'$ are $n \times 1$ column vectors; $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)'$ is a $p \times 1$ column vector; and $\mathbb{X} = (\boldsymbol{X}_1, \cdots, \boldsymbol{X}_n)'$ is a $n \times p$ matrix.

Recall that the OLS linear estimator is given by:

$$\hat{\boldsymbol{\beta}}^{OLS} = (\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'\boldsymbol{Y}$$

Substituting the linear regression in matrix form into the OLS linear estimator we get:

$$\begin{aligned}\hat{\boldsymbol{\beta}}^{OLS} &= (\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'(\mathbb{X}\boldsymbol{\beta}+\boldsymbol{\epsilon}) \\ \hat{\boldsymbol{\beta}}^{OLS}-\boldsymbol{\beta} &= (\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'\boldsymbol{\epsilon} \\ \mathbb{X}(\hat{\boldsymbol{\beta}}^{OLS}-\boldsymbol{\beta}) &= \mathbb{X}(\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'\boldsymbol{\epsilon}\end{aligned}$$

Where:

$$\boldsymbol{e} = \widehat{\boldsymbol{Y}} - \mathbb{E}\left[\boldsymbol{Y}|X\right] = \mathbb{X}(\hat{\boldsymbol{\beta}}^{OLS}-\boldsymbol{\beta})$$

Therefore, the expected value of the mean squared error for the OLS estimator is:

$$\begin{aligned}\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left(\boldsymbol{X}_i'(\hat{\boldsymbol{\beta}}^{OLS}-\boldsymbol{\beta})\right)^2\right] &= \frac{1}{n}\mathbb{E}\left[\boldsymbol{e}'\boldsymbol{e}\right] \\ &= \frac{1}{n}\mathbb{E}\left[(\mathbb{X}(\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'\boldsymbol{\epsilon})'(\mathbb{X}(\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'\boldsymbol{\epsilon})\right] \\ &= \frac{1}{n}\mathbb{E}\left[\boldsymbol{\epsilon}'\mathbb{X}(\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'\mathbb{X}(\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'\boldsymbol{\epsilon}\right]\end{aligned}$$

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left(\boldsymbol{X}_i'(\hat{\boldsymbol{\beta}}^{OLS}-\boldsymbol{\beta})\right)^2\right] = \frac{1}{n}\mathbb{E}\left[\boldsymbol{\epsilon}'\mathbb{X}(\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'\boldsymbol{\epsilon}\right] \tag{1}$$

Lemma. *If $\mathbb{A}$ is a $n \times p$ matrix, and $\mathbb{B}$ is a $p \times n$ matrix, then:*

$$tr(\mathbb{A}\mathbb{B}) = tr(\mathbb{B}\mathbb{A})$$

The expected value in equation 1 is an scalar, therefore:

$$\frac{1}{n}\mathbb{E}\left[\boldsymbol{\epsilon}'\mathbb{X}(\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'\boldsymbol{\epsilon}\right] = \frac{1}{n}\mathrm{tr}\left(\mathbb{E}\left[\boldsymbol{\epsilon}'\mathbb{X}(\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'\boldsymbol{\epsilon}\right]\right)$$

Because the trace and the expected value are linear operators, we interchange the order of the trace and the expected value:

$$\begin{aligned}\frac{1}{n}\mathbb{E}\left[\boldsymbol{\epsilon}'\mathbb{X}(\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'\boldsymbol{\epsilon}\right] &= \frac{1}{n}\mathbb{E}\left[\mathrm{tr}\left(\boldsymbol{\epsilon}'\mathbb{X}(\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'\boldsymbol{\epsilon}\right)\right] \\ &= \frac{1}{n}\mathbb{E}\left[\mathrm{tr}\left(\boldsymbol{\epsilon}\boldsymbol{\epsilon}'\mathbb{X}(\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'\right)\right] \\ &= \frac{1}{n}\mathrm{tr}\left(\mathbb{E}\left[\boldsymbol{\epsilon}\boldsymbol{\epsilon}'\mathbb{X}(\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'\right]\right)\end{aligned}$$

By the law of iterated expectations:

$$
\begin{aligned}
\frac{1}{n}\mathbb{E}\left[\boldsymbol{\epsilon}'\mathbb{X}(\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'\boldsymbol{\epsilon}\right] &= \frac{1}{n}\text{tr}\left(\mathbb{E}\left[\mathbb{E}\left[\boldsymbol{\epsilon}\boldsymbol{\epsilon}'\mathbb{X}(\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'|\mathbb{X}\right]\right]\right) \\
&= \frac{1}{n}\text{tr}\left(\mathbb{E}\left[\mathbb{E}\left[\boldsymbol{\epsilon}\boldsymbol{\epsilon}'|\mathbb{X}\right]\mathbb{X}(\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'\right]\right) \\
&= \frac{1}{n}\text{tr}\left(\mathbb{E}\left[\sigma^2\mathbb{I}_n\mathbb{X}(\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'\right]\right) \\
&= \frac{\sigma^2}{n}\text{tr}\left(\mathbb{E}\left[\mathbb{X}(\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'\right]\right) \\
&= \frac{\sigma^2}{n}\mathbb{E}\left[\text{tr}\left(\mathbb{X}(\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'\right)\right] \\
&= \frac{\sigma^2}{n}\mathbb{E}\left[\text{tr}\left(\mathbb{X}'\mathbb{X}(\mathbb{X}'\mathbb{X})^{-1}\right)\right] \\
&= \frac{\sigma^2}{n}\mathbb{E}\left[\text{tr}\left(\mathbb{I}_p\right)\right] \\
&= \frac{\sigma^2}{n}\mathbb{E}\left[p\right]
\end{aligned}
$$

$$
\frac{1}{n}\mathbb{E}\left[\boldsymbol{\epsilon}'\mathbb{X}(\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'\boldsymbol{\epsilon}\right] = \frac{p\sigma^2}{n}
$$

□

Meaning that the expected value of the mean squared error of the OLS linear estimator does not converge to zero when $p >> n$.

7 Sparsity

Given the linear regression model:

$$
Y = \boldsymbol{X}'\boldsymbol{\beta} + \epsilon, \quad \mathbb{E}\left[\epsilon|\boldsymbol{X}\right] = 0
$$

where Y is the response variable, $\boldsymbol{X}$ is the p-dimensional feature vector, $\boldsymbol{\beta}$ is the p-dimensional coefficient vector, and ϵ is the error term. We assume that ϵ is independent of $\boldsymbol{X}$.

Definition (Sparsity Index)**.**

$$
\begin{aligned}
S &= \sum_{j=1}^{p} \mathbf{1}_{\{\beta_j \neq 0\}} \\
S &= ||\boldsymbol{\beta}||_0
\end{aligned}
$$

It is often called the l_0 norm of $\boldsymbol{\beta}$, even though does not satisfy all the properties of a norm[4].

[4]See Norms in the appendix for more details.

We are going to assume that S is small, i.e., $S \ll p$. This assumption is often referred to as the *sparsity assumption*.

A more relaxed version of the sparsity assumption is the *approximate sparsity assumption*, where most $\boldsymbol{\beta}_j$ are close to zero, but not exactly zero.

7.1 Best Subset Selection

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{b} \in \mathbb{R}^p}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} (Y_i - \boldsymbol{X}_i' \boldsymbol{b})^2 + \lambda ||\boldsymbol{b}||_0$$

where λ is the penalty parameter, and $\lambda||b||_0$ is the penalty term that encourages sparsity in the solution.

In order to solve this optimization problem, we need to consider all possible subset of features $\binom{n}{k}$ for $k = 0, 1, \ldots, p$. This is computationally infeasible for large p, and the ℓ_0 penalty is non-convex, which makes the optimization problem even harder.

7.2 Least Absolute Shrinkage and Selection Operator (LASSO)

Introduced by Tibshirani (1996), the LASSO estimator is defined as:

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{b} \in \mathbb{R}^p}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} (Y_i - \boldsymbol{X}_i' \boldsymbol{b})^2 + \lambda ||\boldsymbol{b}||_1$$

7.2.1 Main LASSO Result

If λ is chosen appropriately, and some regularization conditions are satisfied (including sparsity), then the LASSO estimator $\hat{\boldsymbol{\beta}}^{LASSO}$ satisfies:

$$||\hat{\boldsymbol{\beta}}^{LASSO} - \boldsymbol{\beta}||_2 \leq C \sqrt{\frac{s \ln p}{n}}$$

and,

$$||\hat{\boldsymbol{\beta}}^{LASSO} - \boldsymbol{\beta}||_1 \leq CS \sqrt{\frac{\ln p}{n}}$$

with probability approaching 1, where C is some constant. Intuitively, the LASSO estimator is consistent in terms of the ℓ_2 norm, if $\frac{s \ln p}{n} \to 0$.

8 Main LASSO Result: Proof of Theorem

For clarification on notation, consult appendix C.

Theorem 1 (Main LASSO Result)**.** $\forall c > 1$*, if for some* $\lambda > 0$*:*

$$\lambda > \frac{2c}{n} \max_{1\leq j\leq p} \left| \sum_{i=1}^{n} X_{i,j}\epsilon_i \right| = \frac{2c}{n} \max_{1\leq j\leq p} \left| \boldsymbol{X}'_{(j)}\boldsymbol{\epsilon} \right|$$

then, the following inequalities hold:

$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_{2,n} \leq \left(1 + \frac{1}{c}\right) \frac{\lambda\sqrt{S}}{k_{\bar{c}}}$$
$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_{1} \leq (1 + \bar{c}) \left(1 + \frac{1}{c}\right) \frac{\lambda S}{k_{\bar{c}}}$$

where $\bar{c} = \dfrac{c+1}{c-1}$

Recall the linear regression:

$$\boldsymbol{Y} = \mathbb{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{2}$$

The *LASSO* estimator is defined as:

$$\hat{\boldsymbol{\beta}} = \underset{b\in\mathbb{R}^p}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} (Y_i - \boldsymbol{X}'_i b)^2 + \lambda \sum_{j=1}^{p} |b_j| = \underset{b\in\mathbb{R}^p}{\arg\min} \frac{1}{n} \|\boldsymbol{Y} - \mathbb{X}b\|_2^2 + \lambda\|b\|_1 \tag{3}$$

And lets denote $\boldsymbol{\delta} = \hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \in \mathbb{R}^p$.

We are intending to prove the main Lasso result (theorem 1). In order to do so, we will introduce the following claims that, when combined, are sufficient to prove the theorem.

Claim 1. *For all λ which satisfies the condition of theorem 1, the following inequality holds:*

$$\|\boldsymbol{\delta}_{T^\complement}\|_1 \leq \bar{c}\|\boldsymbol{\delta}_T\|_1$$

Claim 2. *For all λ which satisfies the condition of theorem 1, the following inequality holds:*

$$\|\boldsymbol{\delta}\|_{2,n}^2 \leq \left(1 + \frac{1}{c}\right) \lambda \|\boldsymbol{\delta}_T\|_1$$

Where $\|\boldsymbol{\delta}\|_{2,n}^2 = MSE\left(\mathbb{X}\hat{\boldsymbol{\beta}}\right)$.

Proof of Claim 2. The LASSO estimator (equation 3) is the solution of the minimization problem. Then:

$$\frac{1}{n}\|\boldsymbol{Y}-\mathbb{X}\hat{\boldsymbol{\beta}}\|_2^2+\lambda\|\hat{\boldsymbol{\beta}}\|_1 \leq \frac{1}{n}\|\boldsymbol{Y}-\mathbb{X}\boldsymbol{\beta}\|_2^2+\lambda\|\boldsymbol{\beta}\|_1$$
$$\frac{1}{n}\|\mathbb{X}\boldsymbol{\beta}+\boldsymbol{\epsilon}-\mathbb{X}\hat{\boldsymbol{\beta}}\|_2^2+\lambda\|\hat{\boldsymbol{\beta}}\|_1 \leq \frac{1}{n}\|\mathbb{X}\boldsymbol{\beta}+\boldsymbol{\epsilon}-\mathbb{X}\boldsymbol{\beta}\|_2^2+\lambda\|\boldsymbol{\beta}\|_1$$
$$\frac{1}{n}\|\boldsymbol{\epsilon}-\mathbb{X}\boldsymbol{\delta}\|_2^2+\lambda\|\hat{\boldsymbol{\beta}}\|_1 \leq \frac{1}{n}\|\boldsymbol{\epsilon}\|_2^2+\lambda\|\boldsymbol{\beta}\|_1$$

Expanding the cuadratic norm on the left side:

$$\frac{1}{n}\|\boldsymbol{\epsilon}\|_2^2-\frac{2}{n}\boldsymbol{\epsilon}'\mathbb{X}\boldsymbol{\delta}+\frac{1}{n}\|\mathbb{X}\boldsymbol{\delta}\|_2^2+\lambda\|\hat{\boldsymbol{\beta}}\|_1 \leq \frac{1}{n}\|\boldsymbol{\epsilon}\|_2^2+\lambda\|\boldsymbol{\beta}\|_1$$
$$\frac{1}{n}\|\mathbb{X}\boldsymbol{\delta}\|_2^2 \leq \frac{2}{n}\boldsymbol{\epsilon}'\mathbb{X}\boldsymbol{\delta}+\lambda\|\boldsymbol{\beta}\|_1-\lambda\|\hat{\boldsymbol{\beta}}\|_1$$
$$\|\boldsymbol{\delta}\|_{2,n}^2 \leq \frac{2}{n}\boldsymbol{\epsilon}'\mathbb{X}\boldsymbol{\delta}+\lambda\|\boldsymbol{\beta}\|_1-\lambda\|\hat{\boldsymbol{\beta}}\|_1 \leq \left|\frac{2}{n}\boldsymbol{\epsilon}'\mathbb{X}\boldsymbol{\delta}\right|+\lambda\|\boldsymbol{\beta}\|_1-\lambda\|\hat{\boldsymbol{\beta}}\|_1$$

We define $\mathcal{S}=\frac{1}{n}\boldsymbol{\epsilon}'\mathbb{X}$ and, by the Asymmetric Hölder inequality, we have:

$$|\mathcal{S}'\boldsymbol{\delta}| \leq \|\mathcal{S}\|_\infty\|\boldsymbol{\delta}\|_1$$

where $\|\mathcal{S}\|_\infty = \max_{1\leq j\leq p}|\sum_{i=1}^n X_{i,j}\epsilon_i| = \max_{1\leq j\leq p}\left|\boldsymbol{X}'_{(j)}\boldsymbol{\epsilon}\right|$.

Substituting this inequality in the previous expression:

$$\|\boldsymbol{\delta}\|_{2,n}^2 \leq |2\mathcal{S}'\boldsymbol{\delta}|+\lambda\|\boldsymbol{\beta}\|_1-\lambda\|\hat{\boldsymbol{\beta}}\|_1 \leq 2\|\mathcal{S}\|_\infty\|\boldsymbol{\delta}\|_1+\lambda\|\boldsymbol{\beta}\|_1-\lambda\|\hat{\boldsymbol{\beta}}\|_1$$

Because we asume that λ satisfies the condition of theorem 1, we have:

$$\|\mathcal{S}\|_\infty \leq \frac{\lambda}{2c}$$

Substituting in the previous result:

$$\|\boldsymbol{\delta}\|_{2,n}^2 \leq 2\|\mathcal{S}\|_\infty\|\boldsymbol{\delta}\|_1+\lambda\|\boldsymbol{\beta}\|_1-\lambda\|\hat{\boldsymbol{\beta}}\|_1 \leq 2\frac{\lambda}{2c}\|\boldsymbol{\delta}\|_1+\lambda\|\boldsymbol{\beta}\|_1-\lambda\|\hat{\boldsymbol{\beta}}\|_1$$
$$\|\boldsymbol{\delta}\|_{2,n}^2 \leq \frac{\lambda}{c}\|\boldsymbol{\delta}\|_1+\lambda\|\boldsymbol{\beta}\|_1-\lambda\|\hat{\boldsymbol{\beta}}\|_1$$
$$\|\boldsymbol{\delta}\|_{2,n}^2 \leq \frac{\lambda}{c}\|\boldsymbol{\delta}_T\|_1+\frac{\lambda}{c}\|\boldsymbol{\delta}_{T^\complement}\|_1+\lambda\left(\|\boldsymbol{\beta}\|_1-\|\hat{\boldsymbol{\beta}}\|_1\right)$$
$$\|\boldsymbol{\delta}\|_{2,n}^2 \leq \frac{\lambda}{c}\|\boldsymbol{\delta}_T\|_1+\frac{\lambda}{c}\|\boldsymbol{\delta}_{T^\complement}\|_1+\lambda\left(\|\boldsymbol{\beta}_T\|_1-\|\hat{\boldsymbol{\beta}}_T+\hat{\boldsymbol{\beta}}_{T^\complement}\|_1\right)$$
$$\|\boldsymbol{\delta}\|_{2,n}^2 \leq \frac{\lambda}{c}\|\boldsymbol{\delta}_T\|_1+\frac{\lambda}{c}\|\boldsymbol{\delta}_{T^\complement}\|_1+\lambda\left(\|\boldsymbol{\beta}_T\|_1-\|\hat{\boldsymbol{\beta}}_T\|_1-\|\hat{\boldsymbol{\beta}}_{T^\complement}\|_1\right)$$

By the triangle inequality:

$$\|\boldsymbol{\delta}\|_{2,n}^2 \leq \frac{\lambda}{c}\|\boldsymbol{\delta}_T\|_1+\frac{\lambda}{c}\|\boldsymbol{\delta}_{T^\complement}\|_1+\lambda\left(\|\boldsymbol{\beta}_T\|_1-\|\hat{\boldsymbol{\beta}}_T\|_1-\|\hat{\boldsymbol{\beta}}_{T^\complement}\|_1\right) \leq \frac{\lambda}{c}\|\boldsymbol{\delta}_T\|_1+\frac{\lambda}{c}\|\boldsymbol{\delta}_{T^\complement}\|_1+\lambda\left(\|\boldsymbol{\beta}_T-\hat{\boldsymbol{\beta}}_T\|_1-\|\hat{\boldsymbol{\beta}}_{T^\complement}\|_1\right)$$
$$\|\boldsymbol{\delta}\|_{2,n}^2 \leq \frac{\lambda}{c}\|\boldsymbol{\delta}_T\|_1+\frac{\lambda}{c}\|\boldsymbol{\delta}_{T^\complement}\|_1+\lambda\left(\|\boldsymbol{\beta}_T-\hat{\boldsymbol{\beta}}_T\|_1-\|\boldsymbol{\beta}_{T^\complement}-\hat{\boldsymbol{\beta}}_{T^\complement}\|_1\right)$$
$$\|\boldsymbol{\delta}\|_{2,n}^2 \leq \frac{\lambda}{c}\|\boldsymbol{\delta}_T\|_1+\frac{\lambda}{c}\|\boldsymbol{\delta}_{T^\complement}\|_1+\lambda\left(\|\boldsymbol{\delta}_T\|_1-\|\boldsymbol{\delta}_{T^\complement}\|_1\right)$$
$$\|\boldsymbol{\delta}\|_{2,n}^2 \leq \lambda\left(1+\frac{1}{c}\right)\|\boldsymbol{\delta}_T\|_1-\lambda\left(1-\frac{1}{c}\right)\|\boldsymbol{\delta}_{T^\complement}\|_1 \leq \lambda\left(1+\frac{1}{c}\right)\|\boldsymbol{\delta}_T\|_1$$

Therefore,

$$\|\boldsymbol{\delta}\|_{2,n}^2 \leq \lambda \left(1+\frac{1}{c}\right) \|\boldsymbol{\delta}_T\|_1$$

□

Proof of Claim 1. Following the proof 8 for the claim 2, we have that:

$$\|\boldsymbol{\delta}\|_{2,n}^2 \leq \lambda \left(1+\frac{1}{c}\right) \|\boldsymbol{\delta}_T\|_1 - \lambda \left(1-\frac{1}{c}\right) \|\boldsymbol{\delta}_{T^\complement}\|_1 \leq \lambda \left(1+\frac{1}{c}\right) \|\boldsymbol{\delta}_T\|_1$$

We know that $0 \leq \|\boldsymbol{\delta}\|_{2,n}^2$, then:

$$\begin{aligned} 0 &\leq \lambda \left(1+\frac{1}{c}\right) \|\boldsymbol{\delta}_T\|_1 - \lambda \left(1-\frac{1}{c}\right) \|\boldsymbol{\delta}_{T^\complement}\|_1 \leq \lambda \left(1+\frac{1}{c}\right) \|\boldsymbol{\delta}_T\|_1 \\ \|\boldsymbol{\delta}_{T^\complement}\|_1 &\leq \frac{1-\frac{1}{c}}{1+\frac{1}{c}} \|\boldsymbol{\delta}_T\|_1 \\ \|\boldsymbol{\delta}_{T^\complement}\|_1 &\leq \bar{c} \|\boldsymbol{\delta}_T\|_1 \end{aligned}$$

□

Claim 3. *If λ satisfies the condition of theorem 1, then the claims 1 and 2 implies the theorem 1.*

Proof. The claim 1 implies that $\boldsymbol{\delta} \in \mathcal{R}_{\bar{c}}$, where $\mathcal{R}_{\bar{c}} = \{\boldsymbol{\delta} \in \mathbb{R}^p : \|\boldsymbol{\delta}_{T^\complement}\| \leq \bar{c} \|\boldsymbol{\delta}_T\|_1\|\}$. Also, it implies that [5]:

$$k_{\bar{c}} \leq \frac{\sqrt{S}\|\boldsymbol{\delta}\|_{2,n}}{\|\boldsymbol{\delta}_T\|_1} \implies \|\boldsymbol{\delta}_T\|_1 \leq \frac{\sqrt{S}\|\boldsymbol{\delta}\|_{2,n}}{k_{\bar{c}}}$$

Substituting in the claim 2:

$$\begin{aligned} \|\boldsymbol{\delta}\|_{2,n}^2 &\leq \lambda \left(1+\frac{1}{c}\right) \|\boldsymbol{\delta}_T\|_1 \leq \lambda \left(1+\frac{1}{c}\right) \frac{\sqrt{S}\|\boldsymbol{\delta}\|_{2,n}}{k_{\bar{c}}} \\ \|\boldsymbol{\delta}\|_{2,n} &\leq \left(1+\frac{1}{c}\right) \frac{\lambda\sqrt{S}}{k_{\bar{c}}} \end{aligned}$$

Which is one of the inequalities of the theorem 1.

Knowing that $\boldsymbol{\delta} = \boldsymbol{\delta}_T + \boldsymbol{\delta}_{T^\complement}$, and by the triangle inequality, $\|\boldsymbol{\delta}\|_1 \leq \|\boldsymbol{\delta}_T\|_1 + \|\boldsymbol{\delta}_{T^\complement}\|_1$. Substituting in the claim 1:

$$\begin{aligned} \|\boldsymbol{\delta}_T\|_1 + \|\boldsymbol{\delta}_{T^\complement}\|_1 &\leq \|\boldsymbol{\delta}_T\|_1 + \bar{c}\|\boldsymbol{\delta}_T\|_1 \\ \|\boldsymbol{\delta}_T\|_1 + \|\boldsymbol{\delta}_{T^\complement}\|_1 &\leq (1+\bar{c})\|\boldsymbol{\delta}_T\|_1 \leq (1+\bar{c}) \frac{\sqrt{S}\|\boldsymbol{\delta}\|_{2,n}}{k_{\bar{c}}} \leq (1+\bar{c})(1+\frac{1}{c}) \frac{\lambda\sqrt{S}}{k_{\bar{c}}} \\ \|\boldsymbol{\delta}\|_1 &\leq (1+\bar{c})(1+\frac{1}{c}) \frac{\lambda S}{k_{\bar{c}}} \end{aligned}$$

Which is the other inequality of the theorem 1. □

[5] See appendix C for the definition of $k_{\bar{c}}$

A Big O Notation

The *Big O* notation in statistics deals with the convergence of sets of random variables, where convergence is in the sense of convergence in probability. The notation is used to describe the rate of convergence of a sequence of random variables to a limit.

For a set of random variables X_n, and a corresponding set of constants a_n (both indexed by n), the notation:

$$X_n = O_p(a_n), \quad \text{as} \quad n \to \infty$$

means that the set of values $\frac{X_n}{a_n}$ is stochastically bounded, That means:

$$\forall \varepsilon > 0, \exists M \in (0, \infty), \exists N_\varepsilon \in (0, \infty) : \forall n > N_\varepsilon \left(n > N_\varepsilon \implies \mathbb{P}\left(\left| \frac{X_n}{a_n} \right| > M \right) < \varepsilon \right)$$

equivalently, we can rewrite the above expression as:

$$\forall \varepsilon > 0, \exists \delta_\varepsilon \in (0, \infty), \exists N_\varepsilon \in (0, \infty) : \forall n > N_\varepsilon \left(n > N_\varepsilon \implies \mathbb{P}\left(|X_n| > \delta_\varepsilon \right) < \varepsilon \right)$$

B Norms

Definition. *A norm on a metric vector space V over a field F (usually $\mathbb{R}$ or $\mathbb{C}$) is a function $\|\cdot\|: V \to [0, \infty)$ that satisfies the following properties for all vectors $\boldsymbol{u}, \boldsymbol{v} \in V$ and all scalars $\alpha \in F$:*

1. $\|\boldsymbol{v}\| = 0$ *if and only if* $\boldsymbol{v} = \boldsymbol{0}$.
2. $\|\alpha \boldsymbol{v}\| = |\alpha| \|\boldsymbol{v}\|$ *(Absolute scalability).*
3. $\|\boldsymbol{u} + \boldsymbol{v}\| \leq \|\boldsymbol{u}\| + \|\boldsymbol{v}\|$ *(Triangle inequality).*

Some common norms Defined in the vector space $V = \mathbb{R}^p$ include:

- **Zero norm** (or l_0 norm):

$$\|\boldsymbol{v}\|_0 = \sum_{i=1}^{p} \mathbf{1}_{\{v_i \neq 0\}}$$

- **Manhattan norm** (or l_1 norm):

$$\|\boldsymbol{v}\|_1 = \sum_{i=1}^{p} |v_i|$$

- **Euclidean norm** (or l_2 norm):

$$\|\boldsymbol{v}\|_2 = \sqrt{\sum_{i=1}^{p} |v_i|^2}$$

- **General l_q norm**:

$$\|\boldsymbol{v}\|_q = \left(\sum_{i=1}^{p} |v_i|^q \right)^{1/q}$$

- **Maximum norm** (or l_∞ norm):

$$\|\boldsymbol{v}\|_\infty = \max_{i=1}^{n} |v_i|$$

- l_2 **Prediction norm**:

  For a matrix $\mathbb{X} \in \mathbb{R}^{n \times p}$

$$\|\boldsymbol{v}\|_{2,n} = \sqrt{\frac{1}{n}\boldsymbol{v}'\mathbb{X}'\mathbb{X}\boldsymbol{v}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\boldsymbol{X}_i'\boldsymbol{v}\right)^2}$$

Often the matrix $\mathbb{X}$ is defined as the *(*design matrix) in a linear regression model. From a random sample $\{(\boldsymbol{X}_1, Y_1), \cdots, (\boldsymbol{X}_n, Y_n)\} \overset{iid}{\sim} (\boldsymbol{X}, Y)$, where $\boldsymbol{X}_i \in \mathbb{R}^p$ is a column vector, and $Y_i \in \mathbb{R}$. The design matrix is defined as $\mathbb{X} = (\boldsymbol{X}_1, \cdots, \boldsymbol{X}_n)'$.

An specific case of the *prediction norm* is in the context of linear regression, where:

$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_{2,n} = \sqrt{\frac{1}{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\mathbb{X}'\mathbb{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})}$$

B.1 Useful inequalities Involving Norms

Theorem 2 (Cauchy-Schwarz Inequality)**.** *For any vectors $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^p$, and a norm $\|\cdot\|$,*

$$|\boldsymbol{u}'\boldsymbol{v}| \leq \|\boldsymbol{u}\|\|\boldsymbol{v}\|$$

Proof.

$$|\boldsymbol{u}'\boldsymbol{v}| = \left|\sum_{i=1}^{p} u_i v_i\right| \leq \sum_{i=1}^{p} |u_i v_i| \leq \sum_{i=1}^{p} |u_i||v_i|$$
$$|\boldsymbol{u}'\boldsymbol{v}| = \|\boldsymbol{u}\|\|\boldsymbol{v}\|$$

□

Theorem 3 (Asymetric Hölder's Inequality)**.** *For any vectors $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^p$,*

$$|\boldsymbol{u}'\boldsymbol{v}| \leq \|\boldsymbol{u}\|_1 \|\boldsymbol{v}\|_\infty$$

Proof.

$$|\boldsymbol{u}'\boldsymbol{v}| = \left|\sum_{i=1}^{p} u_i v_i\right| \leq \sum_{i=1}^{p} |u_i v_i| \leq \sum_{i=1}^{p} |u_i| \max_{i=1}^{p} |v_i|$$
$$|\boldsymbol{u}'\boldsymbol{v}| = \|\boldsymbol{u}\|_1 \|\boldsymbol{v}\|_\infty$$

□

Theorem 4 (Triangle Inequality)**.** *For any vectors $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^p$, and a norm $\|\cdot\|$,*

$$\|\boldsymbol{u} + \boldsymbol{v}\| \leq \|\boldsymbol{u}\| + \|\boldsymbol{v}\|$$

Proof. The proof is trivial given that the triangle inequality is a property of norms. □

C Helpful Notation for LASSO Theory

Recall the linear model:

$$Y = \boldsymbol{X}'\boldsymbol{\beta} + \epsilon, \quad \mathbb{E}\left[\epsilon|\boldsymbol{X}\right] = 0$$

Where:

- Vector of p covariates: $\boldsymbol{X} = (X_1, \cdots, X_j, \cdots, X_p)' \in \mathbb{R}^p$
- Dependent Variable: $Y \in \mathbb{R}$
- Vector of coefficients: $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_j, \cdots, \beta_p)' \in \mathbb{R}^p$
- Error term: $\epsilon \in \mathbb{R}$

and a random sample $(\boldsymbol{X}_1, Y_1), \cdots, (\boldsymbol{X}_n, Y_n) \overset{iid}{\sim} (\boldsymbol{X}, Y)$, where $\boldsymbol{X}_i \in \mathbb{R}^p$ and $Y_i \in \mathbb{R}$.

The linear regression (incorporating the random sample) is denoted as:

$$Y_i = \boldsymbol{X}_i'\boldsymbol{\beta} + \epsilon_i, \quad \mathbb{E}\left[\epsilon_i|\boldsymbol{X}_i\right] = 0, \quad \forall i = 1, \cdots, n$$

We can express the same regression in matrix form:

$$\boldsymbol{Y} = \mathbb{X}'\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \mathbb{E}\left[\boldsymbol{\epsilon}|\mathbb{X}\right] = 0$$

Where:

- Dependent Vector: $\boldsymbol{Y} = (Y_1, \cdots, Y_i, \cdots, Y_n)' \in \mathbb{R}^n$
- Design Matrix:

$$\mathbb{X} = \begin{bmatrix} X_{11} & \cdots & X_{1j} & \cdots & X_{1p} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ X_{i1} & \cdots & X_{ij} & \cdots & X_{ip} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ X_{n1} & \cdots & X_{nj} & \cdots & X_{np} \end{bmatrix} \in \mathbb{R}^{n \times p}$$

$$\mathbb{X} = \begin{bmatrix} \boldsymbol{X}_1' \\ \vdots \\ \boldsymbol{X}_i' \\ \vdots \\ \boldsymbol{X}_n' \end{bmatrix} = (\boldsymbol{X}_1, \cdots, \boldsymbol{X}_i, \cdots, \boldsymbol{X}_n)' \in \mathbb{R}^{n \times p}$$

$$\mathbb{X} = (\boldsymbol{X}_{(1)}, \cdots, \boldsymbol{X}_{(j)}, \cdots, \boldsymbol{X}_{(p)}) \in \mathbb{R}^{n \times p}$$

- Vector of coefficients: $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_j, \cdots, \beta_p)' \in \mathbb{R}^p$
- Error Vector: $\boldsymbol{\epsilon} = (\epsilon_1, \cdots, \epsilon_i, \cdots, \epsilon_n)' \in \mathbb{R}^n$

Potentially, $p \sim n$, $p > n$ or $p >> n$.

The estimated linear regression is:

$$\widehat{\boldsymbol{Y}} = \mathbb{X}'\hat{\boldsymbol{\beta}}$$

Where, for *OLS*, the vector estimator of coefficients is given by:

$$\hat{\boldsymbol{\beta}} = (\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'\boldsymbol{Y}$$

The error of estimation is given by:

$$\begin{aligned}\boldsymbol{e} &= \mathbb{E}\left[\boldsymbol{Y}|\mathbb{X}\right] - \widehat{\boldsymbol{Y}} \\ &= \mathbb{X}'\boldsymbol{\beta} - \mathbb{X}'\hat{\boldsymbol{\beta}} \\ &= \mathbb{X}'(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) \\ &= \mathbb{X}'\boldsymbol{\delta}\end{aligned}$$

which is often used to calculate the prediction norm:

$$\|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\|_{2,n} = \|\boldsymbol{\delta}\|_{2,n} = \sqrt{\frac{1}{n}\boldsymbol{\delta}'\mathbb{X}\mathbb{X}'\boldsymbol{\delta}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{X}_i'\boldsymbol{\delta})^2} = \sqrt{\frac{1}{n}\boldsymbol{e}'\boldsymbol{e}} = \text{RMSE}\left(\mathbb{X}\hat{\boldsymbol{\beta}}\right)$$

Definition 1. *Given a vector of coefficients $\boldsymbol{\beta} \in \mathbb{R}^p$, the sets T and $T^{\complement}$ are defined as:*

$$\begin{aligned}T &= \{j \in \{1, \cdots, p\} : \boldsymbol{\beta}_j \neq 0, \forall\} \\ T^{\complement} &= \{1, 2, \cdots, p\} \setminus T\end{aligned}$$

where T denotes the set of indices of the non-zero coefficients in $\boldsymbol{\beta}$, and $T^{\complement}$ is the set of indices of the zero coefficients.

the cardinality of T is $S = |T|$ and the cardinality of $T^{\complement}$ is $|T^{\complement}| = p - S$, where S is the Sparsity Index. *By definition, $|T \cup T^{\complement}| = p$.*

Definition 2. *Being T and $T^{\complement}$ the sets of indices of the non-zero and zero coefficients in $\boldsymbol{\beta}$, respectively, and a vector $\boldsymbol{v} \in \mathbb{R}^p$,*

$$\boldsymbol{v}_T = \begin{cases} v_j, & j \in T \\ 0, & j \notin T \end{cases}$$

$$\boldsymbol{v}_{T^{\complement}} = \begin{cases} v_j, & j \in T^{\complement} \\ 0, & j \notin T^{\complement} \end{cases}$$

Are the projections of $\boldsymbol{v}$ onto the sets T and $T^{\complement}$, respectively.

By definition:

$$\boldsymbol{v} = \boldsymbol{v}_T + \boldsymbol{v}_{T^{\complement}}$$

Definition 3 (Compatibility Constant)**.** $\forall c > 0$,

$$k_c = \inf_{\boldsymbol{\delta} \in \mathbb{R}^p} \frac{\sqrt{S}\|\boldsymbol{\delta}\|_{2,n}}{\|\boldsymbol{\delta}\|_1}$$
$$s.t.$$
$$\|\boldsymbol{\delta}_{T^\complement}\|_1 \leq c\|\boldsymbol{\delta}_T\|_1$$

where $\boldsymbol{\delta} = \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}$ *is the diference between the estimated and true coefficients, and* $\|\boldsymbol{\delta}\|_{2,n} = \sqrt{\frac{1}{n}\boldsymbol{\delta}'\mathbb{X}\mathbb{X}\boldsymbol{\delta}}$ *where* $\mathbb{X} = (\boldsymbol{X}_1, \cdots, \boldsymbol{X}_n)'$ *is the design matrix.*

The restriction set is often denoted as: $\mathcal{R}_c = \{\boldsymbol{\delta} \in \mathbb{R}^p : \|\boldsymbol{\delta}_{T^\complement}\| \leq c\|\boldsymbol{\delta}_T\|_1\|\}$

Definition 4 (Compatibility Constant (Alternative))**.** $\forall \bar{c} > 0$,

$$k_{\bar{c}} = \inf_{\boldsymbol{\delta} \in \mathbb{R}^p} \frac{\sqrt{S}\|\boldsymbol{\delta}\|_{2,n}}{\|\boldsymbol{\delta}\|_1}$$
$$s.t.$$
$$\|\boldsymbol{\delta}_{T^\complement}\|_1 \leq \bar{c}\|\boldsymbol{\delta}_T\|_1$$

where $\boldsymbol{\delta} = \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}$ *is the diference between the estimated and true coefficients, and* $\|\boldsymbol{\delta}\|_{2,n} = \sqrt{\frac{1}{n}\boldsymbol{\delta}'\mathbb{X}\mathbb{X}\boldsymbol{\delta}}$ *where* $\mathbb{X} = (\boldsymbol{X}_1, \cdots, \boldsymbol{X}_n)'$ *is the design matrix.*

The restriction set is often denoted as $\mathcal{R}_{\bar{c}} = \{\boldsymbol{\delta} \in \mathbb{R}^p : \|\boldsymbol{\delta}_{T^\complement}\| \leq \bar{c}\|\boldsymbol{\delta}_T\|_1\|\}$, *and* $\bar{c} = \dfrac{c+1}{c-1}$.

D Spectral Theory

Spectral theory refers to the study of the eigenvalues and eigenvectors of a matrix. In this section, we will discuss some basic results from spectral theory that will be useful in our discussion of Lasso Theory.

Definition 5 (Eigenvalues and Eigenvectors)**.** *Given a symmetric square matrix* $\mathbb{A} \in \mathbb{R}^{p \times p}$

$$\mathbb{A}\boldsymbol{x} = \lambda\boldsymbol{x}$$

for some $\lambda \in \mathbb{R}$ *and* $\boldsymbol{x} \in \mathbb{R}^p$, *where* $\boldsymbol{x} \neq \boldsymbol{0}$, *then* λ *is called an eigenvalue of* $\mathbb{A}$ *and* $\boldsymbol{x}$ *is called an eigenvector of* $\mathbb{A}$.

Without loss of generality, we can assume that the eigenvectors are normalized, i.e., $\boldsymbol{x}'\boldsymbol{x} = \|\boldsymbol{x}\|_2 = 1$.

Theorem 5 (Spectral Decomposition)**.** *For all symmetric matrices* $\mathbb{A} \in \mathbb{R}^{p \times p}$, *there exists* p *pairs* $(\lambda_1, \boldsymbol{x}_1), \ldots, (\lambda_p, \boldsymbol{x}_p)$ *such that:*

$$\mathbb{A}\boldsymbol{x}_j = \lambda_j\boldsymbol{x}_j, \quad \forall j = 1, \ldots, p$$

Without loss of generality, we can assume that the eigenvalues are ordered in an increasing order, i.e., $\lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_p$.

Proof. Consider the following optimization problem:

$$\boldsymbol{x}^* = \arg\min_{\boldsymbol{x} \in \mathbb{R}^p} \boldsymbol{x}'\mathbb{A}\boldsymbol{x} \quad \text{s.t.} \quad \boldsymbol{x}'\boldsymbol{x} = 1$$

Applying the Lagrange multiplier method, we know that:

$$\exists \lambda^* \in \mathbb{R} \wedge \boldsymbol{x}^* \in \mathbb{R}^p : \nabla f(\boldsymbol{x}^*) = \lambda^* \nabla g(\boldsymbol{x}^*) \implies \boldsymbol{x}^* = \underset{\boldsymbol{x} \in \mathbb{R}^p}{\arg\min}\, \boldsymbol{x}' \mathbb{A} \boldsymbol{x} \quad \text{s.t.} \quad \boldsymbol{x}'\boldsymbol{x} = 1$$

where $f(\boldsymbol{x}) = \boldsymbol{x}' \mathbb{A} \boldsymbol{x}$ and $g(\boldsymbol{x}) = \boldsymbol{x}'\boldsymbol{x} - 1$. Then, by the FOC, we have:

$$\nabla f(\boldsymbol{x}^*) = \lambda^* \nabla g(\boldsymbol{x}^*) \implies 2\mathbb{A}\boldsymbol{x}^* = 2\lambda^* \boldsymbol{x}^* \implies \mathbb{A}\boldsymbol{x}^* = \lambda^* \boldsymbol{x}^*$$

Thus, λ^* is an eigenvalue of $\mathbb{A}$ and $\boldsymbol{x}^*$ is an eigenvector of $\mathbb{A}$.

Also, substituting $\boldsymbol{x}^*$ and λ^* into $f(\boldsymbol{x}) = \boldsymbol{x}' \mathbb{A} \boldsymbol{x}$, and knowing that $\lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_p$, we have:

$$\lambda_1 = \boldsymbol{x}_1' \mathbb{A} \boldsymbol{x}_1 \geq \boldsymbol{x}'^* \mathbb{A} \boldsymbol{x}^* = \lambda^* \boldsymbol{x}'^* \boldsymbol{x}^* = \lambda^* \implies \lambda_1 = \lambda^*$$

Proving that λ^* is the smallest eigenvalue of $\mathbb{A}$. By repeating the same process for the remaining $p - 1$ eigenvalues, we can prove the theorem.

In general, the pairs $(\lambda_j, \boldsymbol{x}_j)$ in $j = 1, \ldots, p$, can be found by solving the following optimization problem:

$$\boldsymbol{x}_j = \underset{\boldsymbol{x}_j \in \mathbb{R}^p}{\arg\min}\, \boldsymbol{x}' \mathbb{A} \boldsymbol{x} \quad \text{s.t.} \quad \boldsymbol{x}'\boldsymbol{x} = 1, \quad \boldsymbol{x}'\boldsymbol{x}_i = 0, \quad \forall i = 1, \ldots, j-1$$

□