Pre-requisites

- $\text{span}(A) = \{Ax: x \in \mathbb{R}^d\}$
- $\ker(A) = \{x \in \mathbb{R}^d: Ax = 0\}$
  - $\ker(A) = 0 \Leftrightarrow A$ is invertible
  - $A \in \mathbb{R}^{n\times p}, n > p, \text{rank}(A) = p$ (A is full rank) then A is injective: $\ker(A) = \{0\}$
- Linearity of $\mathbb{E}$: $\mathbb{E}(AX) = A\mathbb{E}(X)$, $\mathbb{E}(XA) = \mathbb{E}(X)A$, $\mathbb{E}(X + A) = \mathbb{E}(X) + A$
- Covariance: $\text{cov}(X) = \mathbb{E}((X - \mathbb{E}(X))(X - \mathbb{E}(X))^T) = (\text{cov}(x_i, x_j))_{i,j}$
- $\text{var}(aX + b) = a^2\text{var}(X)$, $\text{cov}(AX + B) = A\text{cov}(X)A^T$
- Transposition: $(A^T)^T = A$, $(AB)^T = B^TA^T$, $(A + B)^T = A^T + B^T$
  - Symmetric invertible matrix $A \Leftrightarrow A^{-1}$ is symmetric.
  - X^TX is positive symmetric (symmetric with positive eigenvalues).
- Dot product: $(a|b) = a^Tb$, $\| a \|^2 = a^Ta$, $\| (a|b) \| \leq \| a \| \| b \|_2$, $\| a \| = 0 \Rightarrow a = 0$
- Gradient: $\nabla_x(a^Tx) = a$, $\nabla_x(x^TAx) = (A^T + A)x$ in general, $\nabla_x(x^TAx) = 2Ax$ if A is symmetric.
- Trace of a matrix $A \in \mathbb{R}^{n\times n}$ is defined by $\text{tr}(A) = \sum_{i=1}^n A_{i,i}$.
  - $\text{tr}(A) = \text{tr}(A^T)$
  - Linearity: $\text{tr}(\alpha A + B) = \alpha\text{tr}(A) + \text{tr}(B)$
  - $\text{tr}(A^TA) = \sum_{i=1}^n \sum_{j=1}^n A_{ij}^2 = \| A \|_F^2$
  - $\text{tr}(AB) = \text{tr}(BA)$
  - $\text{tr}(PAP^{-1}) = \text{tr}(A)$. Hence, if A is diagonalizable, the trace is the sum of the eigenvalues.
  - If H is an orthogonal projector, $\text{tr}(H) = \text{rank}(H)$.
  - $\text{tr}(u^Tu) = u^Tu$
- Normal distribution: $x \sim \mathcal{N}(0,1) \Rightarrow \sigma x + \mu \sim \mathcal{N}(\mu, \sigma^2)$
  - $x \sim \mathcal{N}(\mu, \sigma^2) \Rightarrow \frac{(x-\mu)}{\sigma} \sim \mathcal{N}(0,1)$
  - $X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$, $X_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$ independent $\Rightarrow X_1 + X_2 \sim \mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$
  - Confidence interval for μ with known variance: $X \sim \mathcal{N}(\mu, \sigma^2) \Rightarrow Z = \frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \sim \mathcal{N}(0,1)$
- Chi-squared distribution: $X_n \sim \mathcal{N}(0,1)$, $Z = \sum_{i=1}^n X_i^2 \sim \mathcal{X}_n^2$
  - $\mathbb{E}(Z) = n$, $\text{var}(Z) = 2n$
- T-Student distribution: $U \sim \mathcal{N}(0,1)$, $Z \sim \mathcal{X}_n$, $\frac{U}{\sqrt{\frac{Z}{n}}} \sim T_n$
  - $\mathbb{E}(T) = 0, n > 0$, $\text{var}(T) = \frac{n}{n-2}, n > 2$
  - Confidence interval for μ with unknown variance: $X \sim \mathcal{N}(\mu, \sigma^2)$, $S^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \hat{X}_i)^2 \Rightarrow T = \frac{\overline{X}-\mu}{\frac{S}{\sqrt{n-1}}} \sim T_{n-1}$
  - Confidence interval for the regression coefficients θ_j^*: $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$, $\hat{\sigma}^2 = \frac{1}{n-p-1}\sum_{i=1}^n (Y_i - \hat{Y}_i)^2 \Rightarrow T_j = \frac{\hat{\theta}_j - \theta_j^*}{\hat{\sigma}\sqrt{(X^TX)_{ii}^{-1}}} \sim T_{n-p-1}$
  - Confidence interval for the predicted values $y^* = x^T\theta^*$: $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$, $\hat{\sigma}^2 = \frac{1}{n-p-1}\sum_{i=1}^n (Y_i - \hat{Y}_i)^2 \Rightarrow T_j = \frac{x^T\hat{\theta}_j - x^T\theta_j^*}{\hat{\sigma}\sqrt{x^T(X^TX)_{ii}^{-1}x}} \sim T_{n-p-1}$
  - Confidence interval for the predicted values $y = y^* + \varepsilon$: $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$, $\hat{\sigma}^2 = \frac{1}{n-p-1}\sum_{i=1}^n (Y_i - \hat{Y}_i)^2 \Rightarrow T_j = \frac{x^T\hat{\theta}_j - x^T\theta_j^*}{\hat{\sigma}\sqrt{1 + x^T(X^TX)_{ii}^{-1}x}} \sim T_{n-p-1}$
- Eigenvalues: A is invertible if and only if its eigenvalues are nonzero.
  - If $\text{vp}(A)$ denotes the set of eigenvalues of A, then $\text{vp}(A + \lambda I) = \lambda + \text{vp}(A)$
- Singular Value Decomposition (SVD): $A \in \mathbb{R}^{n\times p} \Rightarrow \exists U \in \mathbb{R}^{n\times n}$, $\exists V \in \mathbb{R}^{p\times p}$ orthogonal, and $\exists \Sigma \in \mathbb{R}^{n\times p}$ diagonal such that $A = U\Sigma V^T$.
  - The eigenvectors of A^TA are the columns of V.
  - The eigenvectors of AA^T are the columns of U.
  - Singular values in S are on the diagonal component and are the square roots of eigenvalues, arranged in descending order.
- Convexity: $f: \mathbb{R}^p \to \mathbb{R}^n$ and $\nabla^2 f \in \mathbb{R}^{p\times p}$ symmetric positive $\Rightarrow f$ is convex.
- An orthogonal projector P on E, a subspace of $\mathbb{R}^n$: $P^2 = P$, $P^T = P$, $\ker(P) = E^\perp$.
  - Hat matrix: $H = X(X^TX)^{-1}X^T$ is an orthogonal projector onto the column space of X
- λ eigenvalue of $A \Leftrightarrow \exists v$ eigenvector: $Av = \lambda v$
  - The eigenvalues of an idempotent matrix ($A^2 = A$) are either 0 or 1
    - Number of eigenvalues equal to 1 is then $\text{tr}(A)$
- Orthogonal matrix: $P^T = P^{-1}$
- Similar matrices A and B: there exists an orthogonal matrix P such that $B = P^{-1}AP$, they share the same eigenvalues
- Diagonalizable matrix A: there exists an orthogonal matrix P, such that $D := PAP^T$ is diagonal, and its elements being are the eigen values of A
- Quantile function: $Q(p) = F_X^{-1}(p)$, $F_X^{-1}(x) = \mathbb{P}(X \leq x) = p$

Synthèse

Ordinary Least Square

- $\min_\theta \| Y - X\theta \|_2^2$
- $\hat{\theta}_n \in \arg\min_{\theta\in\mathbb{R}^{p+1}} \| Y - X\theta \|_2^2$
- Gram matrix: $\hat{G}_n = \frac{X^TX}{n}$
- Orthogonal projector on $\text{span}(X)$: $\hat{H}_{n,X} \in \mathbb{R}^{n\times n}$
- The OLS estimator always exists, and the associated prediction is given by $\hat{Y} = \hat{H}_{n,X}Y$. It is either:
  - *uniquely defined* $\Leftrightarrow$ the Gram matrix is invertible, which is equivalent to $\ker(X) = \ker(X^TX) = \{0\}$
    - $\hat{\theta} = (X^TX)^{-1}X^TY$
      - $\text{b}(\hat{\theta}_n, \theta^*) = 0$
      - $\text{cov}(\hat{\theta}_n) = \sigma^2(X^TX)^{-1}$
      - $R_{\text{pred}}(\hat{\theta}_n, \theta^*) = (p + 1)\frac{\sigma^2}{n}$
      - $R_{\text{quad}}(\hat{\theta}_n, \theta^*) = \text{tr}((X^TX)^{-1})\sigma^2$
  - *non-unique*, with an infinite number of solutions. This happens if and only if $\ker(X) \neq \{0\}$
    - $\hat{\theta} + \ker(X)$, where $\hat{\theta}$ is a particular solution
    - The traditionally considered solution is $\hat{\theta} = (X^TX)^+X^TY$
      - Moore-Penrose inverse: For a positive semi-definite symmetric matrix A with eigenvectors u_i and corresponding eigenvalues $\lambda_i \geq 0$, $A^+ = \sum_i \lambda_i^{-1} u_i u_i^T \mathbb{1}_{\{\lambda_i > 0\}}$
- $\min_{\tilde{\theta}\in\mathbb{R}^p} \| Y_c - \tilde{X}_c\tilde{\theta} \| = \min_{\theta\in\mathbb{R}^{p+1}} \| Y - X\theta \|$
  - $X = (1_n, \tilde{X})$, $Y_c = Y - 1_n(1_n^TY)$ and $\tilde{X}_c = \tilde{X} - 1_n(1_n^T\tilde{X})$
- Determination coefficient $R^2 = \frac{\|\hat{Y} - \bar{y}_n 1_n\|_2^2}{\|Y - \bar{y}_n 1_n\|_2^2} = 1 - \frac{\|\hat{Y} - Y\|_2^2}{\|Y - \bar{y}_n 1_n\|_2^2}$, because of the orthogonality between $\hat{Y} - Y$ and $\hat{Y}$, and between $\hat{Y} - Y$ and $\bar{y}_n 1_n$

- $R^2 = 0 \Leftrightarrow \hat{Y} = \hat{H}_{1_n} Y$, implying that $\hat{\theta}_n = (\bar{y}_n, 0, \dots, 0)$ is one OLS estimator.

Statistical Model

Fixed-design model

- $Y = X\theta^* + \varepsilon$, $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ iid
- Matrix notations X, Y: each row corresponds to a sample x_i or y_i.
  - We handle the intercept by either centering the vectors or by fixing the first coordinate of each sample $x_{i,1} = 1$.
- $\hat{\theta}_n - \theta^* = (X^T X)^{-1} X^T \varepsilon$
- Bias: $\mathrm{b}(\hat{\theta}_n, \theta^*) = \mathbb{E}(\hat{\theta}) - \theta^*$
  - Unbiased if $\mathrm{b}(\hat{\theta}_n, \theta^*) = 0$
- Quadratic risk: $R_{\text{quad}}(\hat{\theta}_n, \theta^*) = \mathbb{E}(\| \hat{\theta}_n - \theta^* \|^2) = \mathrm{b}(\hat{\theta}_n, \theta^*) - \mathrm{var}(\hat{\theta})$
- Prediction risk: $R_{\text{pred}}(\hat{\theta}_n, \theta^*) = \frac{\mathbb{E}(\|Y^* - \hat{Y}\|^2)}{n}$
- Linear estimator: AY, $A \in \mathbb{R}^{(p+1)\times n}$, A depends only on X
- Under the fixed design model: $\mathrm{cov}(\hat{\theta}_n) \leq \mathrm{cov}(AY)$
- Empirical variance: $\tilde{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$
  - $\mathbb{E}(\tilde{\sigma})_n^2 = \sigma^2 \frac{n-p-1}{n}$
  - Unbiased: $\hat{\sigma}_n^2 = \frac{1}{n-p-1}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$

Gaussian model

- $Y \overset{\text{iid}}{\sim} \mathcal{N}(X\theta^*, \sigma^2)$
- $\hat{\theta}_n \sim \mathcal{N}(\theta^*, \sigma^2 (X^T X)^{-1})$
  - $\mathrm{b}(\hat{\theta}_n, \theta^*) = 0$
  - $\mathrm{cov}(\hat{\theta}_n) = \sigma^2 (X^T X)^{-1}$
- Hat matrix
  - $H = X(X^T X)^{-1} X^T$
    - $H^T = H$
    - $H^2 = H$
    - $HX = X$
- Cochran lemma
  - $H\varepsilon$ and $(I - H)\varepsilon$ are independent
  - $\frac{1}{\sigma^2}\varepsilon^T H \varepsilon \sim \mathcal{X}_{p+1}^2$
  - $\frac{1}{\sigma^2}\varepsilon^T (I - H) \varepsilon \sim \mathcal{X}_{n-p-1}^2$
- $\hat{\theta}$ is independent of $\hat{\sigma}^2$
- Central Limit Theorem (CLT): X_n sequence of iid random variables with the same mean μ and the same standard deviation σ, by defining $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$: $\frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \xrightarrow{L} \mathcal{N}(0,1)$
  - Sufficiently large: $n > 30$

Hypothesis testing

$$\begin{cases} \text{Reject whenever } \hat{T}_n \in \mathcal{R} \\ \text{Do not reject whenever } \hat{T}_n \notin \mathcal{R} \end{cases}$$

- Level $1 - \alpha$
- Errors:
  - Type 1: to reject whereas $\mathcal{H}_0$ is true
  - Type 2: not to reject whereas $\mathcal{H}_0$ is false
- Test of no effect: $\mathcal{H}_0: \theta_k^* = 0$

Ridge Regression

- When X is not full rank, one can add L2 regularization to make the problem solvable: $\min_\theta \| X\theta - Y \|_2^2 + n\lambda \| \theta \|_2^2$
- $\hat{\theta}_n \in \arg\min_{\theta \in \mathbb{R}^{p+1}} \| Y - X\theta \|_2^2 + n\lambda \| \theta \|_2^2$
- $\hat{\theta}_n^{(Ridge)} = (X^T X + \lambda I)^{-1} X^T Y$
  - $\mathrm{b}(\hat{\theta}_n^{(Ridge)}, \theta^{(Ridge)*}) = -n\lambda(X^T X + n\lambda I_p)^{-1}\theta^*$
    - Reduce bias $\lambda \to 0$
    - Reduce variance $\lambda \to \infty$
  - $\mathrm{var}(\hat{\theta}_n^{(Ridge)}) = \sigma^2 (X^T X + n\lambda I_p)^{-1} X^T X (X^T X + n\lambda I_p)^{-1}$
- $\mathrm{var}(\hat{\theta}_n^{(Ridge)}) < \mathrm{var}(\hat{\theta}_n)$

Least Absolute Shrinkage and Selection Operator (LASSO) Regression

- If we know that only certain coordinates of the samples x_i are useful for predicting y_i, we can perform variable selection. One simple way is to use L1 regularization, which forces most coordinates of θ to be zero: $\min_\theta \frac{1}{2} \| Y - X\theta \|_2^2 + \lambda \| \theta \|_1$