2.1.1 Linear regression

We have already seen linear regression in the What is Data Science? Introduction but we will consider a more general setting here.
We define

- the input space $\mathcal{X} := \mathbb{R}^n$ and the output space $\mathcal{Y} := \mathbb{R}^m$
- and the hypothesis class

  $$\mathcal{H} := \{\hat{f} : \mathbb{R}^n \to \mathbb{R}^m | \exists W \in \mathbb{R}^{m \times n}, \hat{f}(x) = Wx\}$$

  To determine $\hat{f}$ from $N \in \mathbb{N}$ data points (x_i, y_i) for $i = 1, \ldots, N$ we consider the problem

  $$\min_{W \in \mathbb{R}^{m \times n}} \frac{1}{2} \sum_{i=1}^{N} ||Wx_i - y_i||^2$$

  Let us define the data matrix $X \in \mathbb{R}^{N \times n}$ with $X_{ij} := (x_i)_j$, the matrix $Y \in \mathbb{R}^{N \times m}$ with $Y_{ij} := (y_i)_j$, and $\beta := W^T$, then we can rewrite this problem as

2.1.1

$$\min_{\beta \in \mathbb{R}^{n \times m}} \frac{1}{2} ||X\beta - Y||^2_{\mathrm{Fro}}$$

where the Frobenius norm of a matrix $A \in \mathbb{R}^{N \times m}$ is defined as

$$||A||^2_{\mathrm{Fro}} := \sum_{i=1}^{N} \sum_{j=1}^{m} |A_{ij}|^2 = \mathrm{Tr}(AA^T)$$

(The sum of squares of all elements in the matrix)
We will first examine the solvability of the optimization problem

Proposition 2.1.1

If the matrix $X^T X \in \mathbb{R}^{n \times n}$ is invertible, then the unique solution of 2.1.1 is given by

2.1.2

$$\hat{\beta} := (X^T X)^{-1} X^T Y$$

Otherwise, there are infinitely many solutions.

Proof. We define the objective function $f(\beta) := \frac{1}{2}\|X\beta - Y\|_{\text{Fro}}^2$. Since f is a convex function of β, β solves the minimization problem (2.1.1) if and only if $\nabla f(\beta) = 0$. From Exercise 1.3.2, we know that the gradient of f is given by

$$\nabla f(\beta) = X^T(X\beta - Y).$$

Thus, $\nabla f(\beta) = 0$ is equivalent to $X^T X\beta = X^T Y$. This linear equation system has the unique solution $\hat{\beta} := (X^T X)^{-1} X^T Y$ if $X^T X$ is invertible, and infinitely many solutions otherwise. □

Remark 2.1.1 (Invertibility of $X^T X$)

Let us denote the columns of the matrix $X \in \mathbb{R}^{N\times n}$ by $a_i \in \mathbb{R}^N$ for $i = 1, \dots, n$. A fact from linear algebra states that $X^T X$ is **invertible if and only if the vectors** $(a_i)_{i=1}^n$ are linearly **independent** in $\mathbb{R}^N$. If $n >> N$, this is a "very likely" event and is referred to in data science as "independent features".
It is also important to note that the size of the matrix $X^T X \in \mathbb{R}^{n\times n}$ is independent of the number of data points. Thus, the difficulty of inversion does not increase with more data.

Next we will try to understand the case where $X^T X$ is not invertible and to construct a "meaningful" solution for 2.1.1. To do this we recall

Definition 2.1.1 (Singular Value Decomposition)

A singular value decomposition of a matrix $X \in \mathbb{R}^{N\times n}$ is a decomposition of the form $X = U\Sigma V^T$ with orthogonal matrices $U \in \mathbb{R}^{N\times N}$ and $V \in \mathbb{R}^{n\times n}$ and a matrix $\Sigma \in \mathbb{R}^{N\times n}$ of the form

$$\Sigma = \begin{pmatrix} \sigma_1 & \dots & \dots & \dots & \dots & \dots \\ \dots & \ddots & \dots & \dots & \dots & \vdots \\ \dots & \dots & \sigma_k & \dots & \dots & 0 \\ \dots & \dots & \dots & 0 & \dots & \vdots \\ \dots & \dots & \dots & & \ddots & \vdots \\ \dots & \dots & 0 & \dots & \dots & 0 \end{pmatrix}$$

with $k \leq n$ singular values $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_k > 0$ (Singular values are ordered by largest to smallest along the diagonal).
We define the pseudo inverse of X as

$$X^\dagger = V\Sigma^\dagger U^T$$

where $\Sigma^\dagger \in \mathbb{R}^{n\times N}$ is given by

$$\Sigma^\dagger = \begin{pmatrix} \frac{1}{\sigma_1} & \dots & \dots & \dots & \dots & \dots \\ \dots & \ddots & \dots & \dots & \dots & \vdots \\ \dots & \dots & \frac{1}{\sigma_k} & \dots & \dots & 0 \\ \dots & \dots & \dots & 0 & \dots & \vdots \\ \dots & \dots & \dots & & \ddots & \vdots \\ \dots & \dots & 0 & \dots & \dots & 0 \end{pmatrix}$$

Remark 2.1.2 The existence of the SVD can be proven by applying the Spectral Theorem to the matrix X^TX

Proposition 2.1.2. The following holds:

$$XX^\dagger X = X, \quad X^\dagger XX^\dagger = X^\dagger$$

$$X^\dagger X \text{ and } XX^\dagger \text{ are symmetric}$$

We observe that for the SVD $X = U\Sigma V^T$, it holds that

$$X^TX = (U\Sigma V^T)^T U\Sigma V^T = V\Sigma^T U^T U\Sigma V^T = V\Sigma^T\Sigma V^T$$

Definition of an Orthogonal matrix

An orthogonal matrix U is a **square matrix** whose columns (and rows) form an orthonormal set. This means:

- Orthonormal columns: Each column vector has a unit length ($||u_i|| = 1$) and any two distinct columns are perpendicular ($u_i \cdot u_j = 0$ for $i \neq j$)
- Inverse property: The inverse of U is equal to its transpose: $U^T = U^{-1}$
  - if $i = j : u_i \cdot u_j = 1$ (unit vectors)
- Orthogonal matrices represent rotations or reflections in space

Back to the previous definition

Thus, we have found a diagonalization (and in particular, a singular value decomposition) of X^TX. The positive eigenvalues of X^TX are exactly given by σ_i^2 for $i = 1, \dots, k$.
In the case that $k = n$, i.e. there are n singular values, X^TX is invertible with $(X^TX)^{-1} = VDV^T$ where $D = \text{diag}(\sigma_1^{-2}, \dots, \sigma_n^{-2})$. Otherwise, we can use the pseudoinverse of X^TX to calculate the solution of 2.1.1 with minimal norm.

Proposition 2.1.3

If the matrix $X^TX \in \mathbb{R}^{n \times n}$ is not invertible, then the solution of 2.1.1 with minimal norm is given by

2.1.3

$$\hat{\beta} := (X^T X)^{\dagger} X^T Y$$

(similarly see 2.1.2)
where $(X^T X)^{\dagger} = V D V^T$ with (D := squares of the singular values of X)

$$D := \begin{pmatrix} \frac{1}{\sigma_1^2} & \dots & \dots & \dots \\ \dots & \ddots & \dots & \dots \\ \dots & \dots & \frac{1}{\sigma_k^2} & \dots \\ \dots & \dots & \dots & 0_{n-k,n-k} \end{pmatrix}$$

How do we know if $X^T X$ is invertible?

$$X = \begin{pmatrix} \dots & x_1^T & \dots \\ & \vdots & \\ \dots & x_N^T & \dots \end{pmatrix}$$

$C = X^T X$ is the covariance matrix. If you have data for which the variance of one direction is zero, you know that the eigenvalue for this direction will be 0 and C will not be invertible.
--> A (covariance?) Matrix with an eigenvalue of 0 is not invertible