Lecture 3: Multiple Linear Regression

Ailin Zhang

2023-05-16

Recap: Linear Regression Models

$$Y = \beta_0 + \beta_1 X1 + \beta_2 X2 + \cdots + \beta_p X_p + \epsilon$$

More generally, linear regression can be categorized as a model is linear in its parameters $\beta_0, \beta_1, \ldots, \beta_p$. For example, the following are linear regression models:

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \epsilon$$

$$Y = \beta_0 + \beta_1 \log(X) + \epsilon$$

even though the relationship between Y and X is not linear.

- Linear in parameters, not linear in predictors
- less restrictive than you might think

Recap: Multiple Linear Regression Model

	SLR		MLR				
	X	Y	X_1	X_2	$\dots$	X_p	Y
case 1:	x_1	y_1	x_{11}	x_{12}	$\dots$	x_{1p}	y_1
case 2:	x_2	y_2	x_{21}	x_{22}	$\dots$	x_{2p}	y_2
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
case n:	x_n	y_n	x_{n1}	x_{n2}	$\dots$	x_{np}	y_n

- In simple linear regression (SLR), we observe one predictor X.
- In multiple linear regression (MLR), we observe p predictors (explanatory variables, covariates)
- Each row is called a case, a record, or a data point
- y_i is the response (or dependent variable) of the ith case
- x_{ik} is the value of the explanatory variable X_k of the ith case

Recap: Multiple Linear Regression Models in Matrix Notation

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n1} & \dots & x_{np} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

Or

$$Y = X\beta + \epsilon$$

Recap: Errors and Residuals

- Error (ϵ_i) can not be directly computed, $\epsilon_i = y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_p x_{ip}$
- The errors ϵ_i can be estimatedby residuals e_i

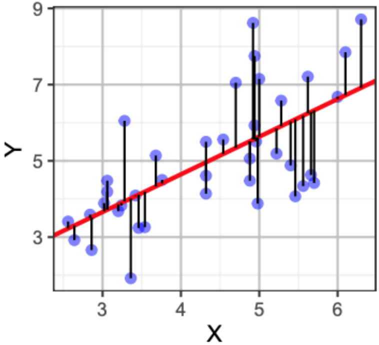
  residual e_i = observed y_i - predicted y_i

$$e_i = y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_p x_{ip})$$

- To estimate parameters, try to get $\hat{y}_i$ to be as close to y_i, so we want each $y_i - \hat{y}_i$ to be close to 0
- To make all of these residuals on average close to zero, consider minimizing the **sum of squares**

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Least Squares Method



In SLR, the least squares estimate is the intercept and slope of the straight line with the minimum sum of squared vertical distances to the data points:

$$\text{argmin}_{\beta_0,\beta_1} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

In MLR, the least squares estimate $(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p)$ is the intercept and slopes of the (hyper)plane with the minimum sum of squared vertical distance to the data points

$$\text{argmin}_{\beta_0,\beta_1,\ldots,\beta_p} \sum^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_p x_{ip})^2$$

Least Squares Solution to SLR

- To find $(\hat{\beta}_0, \hat{\beta}_1)$ that minimize:

$$L(\hat{\beta}_0, \hat{\beta}_1) = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

- One can set the derivatives of L with respect to $\hat{\beta}_0$ and $\hat{\beta}_1$ to 0

$$\frac{\partial L}{\partial \hat{\beta}_0} = -2 \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

$$\frac{\partial L}{\partial \hat{\beta}_1} = -2 \sum_{i=1}^{n} x_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

Least Squares Solution to SLR

- This results in the 2 equations with 2 unknowns:

$$n\hat{\beta}_0 + \hat{\beta}_1 \underbrace{\sum_{i=1}^{n} x_i}_{n\bar{x}} = \underbrace{\sum_{i=1}^{n} y_i}_{n\bar{y}} \xrightarrow{\text{divide by n}} \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\hat{\beta}_0 \underbrace{\sum_{i=1}^{n} x_i}_{n\bar{x}} + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i \xrightarrow{\text{replace } \hat{\beta}_0} (\bar{y} - \hat{\beta}_1 \bar{x}) n\bar{x} + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i$$

$$\iff \hat{\beta}_1 \left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right) = \left(\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}\right)$$

$$\iff \hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}$$

Least Squares Solution to MLR

- To find $(\hat{\beta}_0, \hat{\beta}_1, \ldots, \beta_p)$ that minimize:

$$L(\hat{\beta}_0, \hat{\beta}_1, \ldots, \beta_p) = \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_p x_{ip})^2$$

- One can set the derivatives of L with respect to $\hat{\beta}_j$ to 0

$$\frac{\partial L}{\partial \hat{\beta}_0} = -2\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_p x_{ip}) = 0$$

$$\frac{\partial L}{\partial \hat{\beta}_k} = -2\sum_{i=1}^{n} x_{ik}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_p x_{ip}) = 0, \ k = 1, 2, \ldots, p$$

This results in a linear system of $(p+1)$ equations in $(p+1)$ unknowns.

Least Squares Solution to MLR

- Normal equations: a system of equations whose solution is the Ordinary Least Squares (OLS) estimator of the regression coefficients

$$
\begin{aligned}
\hat{\beta}_0 \cdot n + \hat{\beta}_1 \sum_{i=1}^{n} x_{i1} \quad & + \cdots \quad & + \hat{\beta}_p \sum_{i=1}^{n} x_{ip} &= \sum_{i=1}^{n} y_i \\
\hat{\beta}_0 \sum_{i=1}^{n} x_{i1} + \hat{\beta}_1 \sum_{i=1}^{n} x_{i1} x_{i1} \quad & + \cdots \quad & + \hat{\beta}_p \sum_{i=1}^{n} x_{i1} x_{ip} &= \sum_{i=1}^{n} x_{i1} y_i \\
& \vdots & & \\
\hat{\beta}_0 \sum_{i=1}^{n} x_{ik} + \hat{\beta}_1 \sum_{i=1}^{n} x_{ik} x_{i1} \quad & + \cdots \quad & + \hat{\beta}_p \sum_{i=1}^{n} x_{ik} x_{ip} &= \sum_{i=1}^{n} x_{ik} y_i \\
& \vdots & & \\
\hat{\beta}_0 \sum_{i=1}^{n} x_{ip} + \hat{\beta}_1 \sum_{i=1}^{n} x_{ip} x_{i1} \quad & + \cdots \quad & + \hat{\beta}_p \sum_{i=1}^{n} x_{ip} x_{ip} &= \sum_{i=1}^{n} x_{ip} y_i
\end{aligned}
$$

Matrix Notation

$$
\begin{bmatrix}
n & \sum_{i=1}^{n} x_{i1} & \sum_{i=1}^{n} x_{i2} & \cdots & \sum_{i=1}^{n} x_{ip} \\
\sum_{i=1}^{n} x_{i1} & \sum_{i=1}^{n} x_{i1}x_{i1} & \sum_{i=1}^{n} x_{i1}x_{i2} & \cdots & \sum_{i=1}^{n} x_{i1}x_{ip} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
\sum_{i=1}^{n} x_{ik} & \sum_{i=1}^{n} x_{ik}x_{i1} & \sum_{i=1}^{n} x_{ik}x_{i2} & \cdots & \sum_{i=1}^{n} x_{ik}x_{ip} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
\sum_{i=1}^{n} x_{ip} & \sum_{i=1}^{n} x_{ip}x_{i1} & \sum_{i=1}^{n} x_{ip}x_{i2} & \cdots & \sum_{i=1}^{n} x_{ip}x_{ip}
\end{bmatrix}
\begin{bmatrix}
\hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_k \\ \vdots \\ \hat{\beta}_p
\end{bmatrix}
=
\begin{bmatrix}
\sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_{i1}y_i \\ \vdots \\ \sum_{i=1}^{n} x_{ik}y_i \\ \vdots \\ \sum_{i=1}^{n} x_{ip}y_i
\end{bmatrix}
$$

- In matrix notation, the normal equation is $(X^T X)\hat{\beta} = X^T Y$
- And the least squares estimate is $\hat{\beta} = (X^T X)^{-1} X^T Y$

Introducing Some Linear Algebra

Recall

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, X = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n1} & \dots & x_{np} \end{bmatrix} = \begin{bmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_n^T \end{bmatrix} = (X_1, X_2, \dots, X_p)$$

$$RSS(\beta) = \sum_{i=1}^{n} (y_i - x_i^T \beta)^2 = (Y - X\beta)^T (Y - X\beta)$$

Introducing Some Linear Algebra

The minimizer is achieved when $\nabla_\beta RSS(\beta) = 0$

$$
\begin{aligned}
\nabla_\beta RSS(\beta) &= \nabla_\beta \sum_{i=1}^{n} (y_i - x_i^T \beta)^2 \\
&= \sum_{i=1}^{n} \nabla_\beta (y_i - x_i^T \beta)^2 \\
&= \sum_{i=1}^{n} 2 \cdot (y_i - x_i^T \beta) \underbrace{\nabla_\beta (y_i - x_i^T \beta)}_{= -x_i \text{ dims:}(p+1)\times 1} \\
&= -2 \sum_{i=1}^{n} (y_i - x_i^T \beta) x_i
\end{aligned}
$$

We can also acquire normal equations: $\sum_{i=1}^{n} (y_i - x_i^T \hat{\beta}) x_i = 0$

Introducing Some Linear Algebra

Using the matrix notation, we have

$$LHS = \sum_{i=1}^{n}(y_i - x_i^T\hat{\beta})x_i = \begin{bmatrix} x_1, x_2, \ldots, x_n \end{bmatrix} \begin{bmatrix} y_1 - x_1^T\hat{\beta} \\ y_2 - x_2^T\hat{\beta} \\ \vdots \\ y_n - x_n^T\hat{\beta} \end{bmatrix}$$

$$= X^T(Y - X\hat{\beta}) = 0$$

$$\Rightarrow X^TY - X^TX\hat{\beta} = 0$$

$$(X^TX)\hat{\beta} = X^TY$$

Existence of $\hat{\beta}$

Theorem

The OLS coefficient equals

$$\hat{\beta} = (\sum_{i=1}^{n} x_i x_i^T)^{-1} (\sum_{i=1}^{n} x_i y_i) = (X^T X)^{-1} (X^T Y)$$

if $X^T X$ is non-degenerate.

The non-degeneracy of $X^T X$ in the theorem requires that for any non-zero vector $\alpha \in \mathbb{R}^p$, $\alpha^T X^T X \alpha = ||X\alpha||^2 \neq 0 \Leftrightarrow X\alpha \neq 0$

i.e., the columns of X are **linearly independent.**

If X_1 can be represented by other columns $X_1 = c_2 X_2 + \cdots + c_p X_p$ for some $(c_2, ..., c_p)$, then $X^T X$ is degenerate.

Geometric Solution to OLS

Theorem

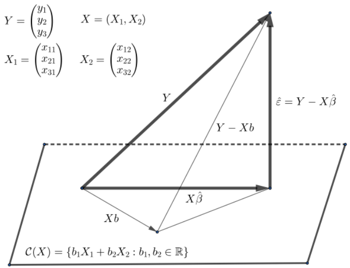
For any $b \in \mathbb{R}^p$, we have the following decomposition

$$||Y - Xb||^2 = ||Y - X\hat{\beta}||^2 + ||X(\hat{\beta} - b)||^2$$

where implies that $||Y - Xb||^2 \geq ||Y - X\hat{\beta}||^2$ with equality holding ***if and only if*** $b = \hat{\beta}$

- Proof (on board)

Geometric Solution to OLS



- The OLS problem is to find the best linear combination of the column vectors of X to approximate the response vector Y
- By projection, the residual vector $\hat{\epsilon} = Y - X\hat{\beta}$ must be orthogonal to $C(X)$, or, equivalently, the residual vector is orthogonal to $X_1, \ldots, X_p$

- This geometric intuition in turn implies that $X_1^T(Y - X\hat{\beta}) = 0, X_2^T(Y - X\hat{\beta}) = 0, \ldots, X_p^T(Y - X\hat{\beta}) = 0$

  which is essentially the normal equation

Revisit Assumptions about ϵ

1. The error ϵ is a random variable with mean of zero.

   If X contains a column of intercepts $1_n = (1, 1, \ldots, 1)^T$, then

$$1_n^T \hat{\epsilon} = 0 \Rightarrow n^{-1} \sum_{i=1}^{n} \hat{\epsilon}_i = 0$$

   So the residuals are automatically centered.

Review: Basics of vectors and matrices

- Euclidean space: The n-dimensional Euclidean space $\mathbb{R}^n$ is a set of all n-dimensional vectors equipped with an inner product:

  $\langle x, y \rangle = x^T y = \sum_{i=1}^{n} x_i y_i$

  where $x = (x_1, \ldots, x_n)^T$ and $y = (y_1, \ldots, y_n)^T$ are two n-dimensional vectors.

- Orthogonality: $x \perp y \Leftrightarrow \langle x, y \rangle = 0$
- Length of a vector x: $||x|| = \sqrt{\langle x, x \rangle} = \sqrt{x^T x}$
- Cauchy–Schwarz inequality: $|\langle x, y \rangle| \leq ||x|| \cdot ||y||$

Review: Basics of vectors and matrices

- Column space of a matrix: Given an $n \times m$ matrix $A = (A_1, A_2, \ldots, A_m)$, we define its column space as $\mathcal{C}(A) = \{\alpha_1 A_1 + \cdots + \alpha_m A_m : \alpha_1, \ldots, \alpha_m \in \mathbb{R}\}$

  which is the set of all linear combinations of the column vectors of A.

- Inverse of a matrix: Let I_n be the $n \times n$ identity matrix. An $n \times n$ matrix A is nonsingular if there exists an $n \times n$ matrix B such that $AB = BA = I_n$. We call B the inverse of A, denoted by A^{-1}
- Symmetric matirx: A is symmetric if $A^T = A$
- Orthogonal matrix: An $n \times n$ matrix is orthogonal if $A^T A = AA^T = I_n$, that is $A^T = A^{-1}$.
- Diagonal matrix: An $n \times n$ diagonal matrix A has zero off-diagonal elements, denoted by $A = \text{diag}\{a_{11}, \ldots, a_{nn}\}$

Review: Eigenvalues and eigenvectors

- Eigenvalue and Eigenvectors: For an $n \times n$ matrix A, if there exists a pair of n-dimensional vector x and a scalar λ such that $Ax = \lambda x$, then we call λ an eigenvalue and x an eigenvector of A. From the definition, eigenvalue and eigenvector come always in pair.

- Eigen-decomposition Theorem:

Theorem

If A is an $n \times n$ symmetric matrix, then there exists an orthogonal matrix P such that $P^T AP = diag\{\lambda_1, \ldots, \lambda_n\}$ Where the λ's are the n eigenvalues of A, and the column vectors of $P = (\gamma_1, \ldots, \gamma_n)$ are the corresponding eigenvectors.

$$AP = P\text{diag}\{\lambda_1, \ldots, \lambda_n\} \Longleftrightarrow A(\gamma_1, \ldots, \gamma_n) = (\lambda_1\gamma_1, \ldots, \lambda_n\gamma_n).$$

Moreover, the eigen-decomposition in the theorem is unique up to the permutation of the columns of P and the corresponding λ_i's.

Review: Eigenvalues and eigenvectors

Corollary

If $P^T AP = diag\{\lambda_1, \ldots, \lambda_n\}$, then

$$A = Pdiag\{\lambda_1, \ldots, \lambda_n\}P^T,\ A^k = A \cdot A \ldots A = Pdiag\{\lambda_1^k, \ldots, \lambda_n^k\}P^T$$

If the eigenvalues of A are nonzero, then $A^{-1} = Pdiag\{\lambda_1^{-1}, \ldots, \lambda_n^{-1}\}P^T$

- If the eigenvalues of A are nonnegative, $A^{1/2} =$?
- From eigen-decomposition theorem, we can write A as:

$$\begin{aligned}
A &= P\text{diag}\{\lambda_1, \ldots, \lambda_n\}P^T \\
&= (\gamma_1, \ldots, \gamma_n)\text{diag}\{\lambda_1, \ldots, \lambda_n\} \begin{pmatrix} \gamma_1^T \\ \vdots \\ \gamma_n^T \end{pmatrix} \\
&= \sum_{i=1}^{n} \lambda_i \gamma_i \gamma_i^T
\end{aligned}$$

Review: Rayleigh quotient and eigenvalues

Theorem

For an $n \times n$ symmetric matrix A, let $r(x) = x^T A x / x^T x$ be the Rayleigh quotient of x. The maximum and minimum eigenvalues of A are

$$\lambda_{max}(A) = \max_{x \neq 0} r(x), \qquad \lambda_{min}(A) = \min_{x \neq 0} r(x)$$

Review: Rank and Quadratic form

- Rank and determinant: For an $n \times n$ symmetric matrix, its rank equals the number of non-zero eigenvalues and its determinant equals the product of all eigenvalues. The matrix A is of full rank if all its eigenvalues are non-zero, which implies that its rank equals n and its determinant is non-zero.
- Quadratic form: For an $n \times n$ symmetric matrix $A = (a_{ij})$ and an n-dimensional vector x: $x^T Ax =< x, Ax >= \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} x_i x_j$
- Semi-definite and definite:

  We call A positive semi-definite, denoted by $A \succeq 0$, if $x^T Ax \geq 0$ for all nonzero x.

  We call A positive definite,denoted by $A \succ 0$, if $x^T Ax > 0$ for all nonzero x.

Review: Definite matrix and eigenvalues

Theorem

For a symmetric matrix A, it is positive semi-definite if and only if all its eigenvalues are nonnegative, and it is positive definite if and only if all its eigenvalues are positive.

- We can also define the partial order between matrices. We call $A \succeq B$ if and only if $A - B \succeq 0$, and we call $A \succ B$ if and only if $A - B \succ 0$.
- This is important in statistics because we often compare the efficiency of estimators based on their covariance matrices.

Review: Trace

- Trace: The trace of an $n \times n$ symmetric matrix $A = (a_{ij})$ is the sum of all its diagonal elements, denoted by

$$\mathrm{trace}(A) = \sum_{i=1}^{n} a_{ii}$$

The trace operator has two important properties that can sometimes help to simplify calculations.

- Proposition 1: $\mathrm{trace}(AB) = \mathrm{trace}(BA)$ as long as AB and BA are both square matrices.
- Proposition 2: The trace of an $n \times n$ symmetric matrix A equals the sum of its eigenvalues (Proof on board):

$$\mathrm{trace}(A) = \sum_{i=1}^{n} \lambda_i.$$

Review: Projection matrix

In MLR: $\hat{Y} = HY$

An $n \times n$ symmetric matrix H is a projection matrix if $H^2 = H$. Based on the eigen-decomposition $H = \sum_{i=1}^{n} \lambda_i \gamma_i \gamma_i^T$, we have

$$
\begin{aligned}
H^2 = H &\Rightarrow \sum_{i=1}^{n} \lambda_i^2 \gamma_i \gamma_i^T = \sum_{i=1}^{n} \lambda_i \gamma_i \gamma_i^T \\
&\Rightarrow \sum_{i=1}^{n} (\lambda_i^2 - \lambda_i) \gamma_i \gamma_i^T = 0 \\
&\Rightarrow \lambda_i^2 - \lambda_i = 0, \qquad (i = 1, \ldots, n)
\end{aligned}
$$

which implies that the eigenvalues of H are either 1 or 0. So the trace of H equals its rank: trace(H) = rank(H)

- Question: What is the projection matrix in MLR?

Review: Matrix decomposition

- Cholesky decomposition: An $n \times n$ positive semi-definite matrix A can be decomposed as $A = LL^T$ where L is an $n \times n$ lower triangular matrix with non-negative diagonal elements.

  Take an arbitrary orthogonal matrix Q, we have $A = LQQ^TL^T = CC^T$ where $C = LQ$. So we can decompose a positive semi-definite matrix A as $A = CC^T$, but this decomposition is not unique.

Review: Vector calculus

If $f(x)$ is a function from $\mathbb{R}^p$ to $\mathbb{R}$, then we use the notation:

$$\frac{\partial f(x)}{\partial x} \equiv \begin{pmatrix} \frac{\partial f(x)}{\partial x_1} \\ \vdots \\ \frac{\partial f(x)}{\partial x_p} \end{pmatrix}$$

1. If $f(x) = x^T a = a^T x$, with $a, x \in \mathbb{R}^p$, what is $\frac{\partial f(x)}{\partial x}$? $\frac{\partial f(x)}{\partial x} = a$
2. If $f(x) = x^T A x$, with $x \in \mathbb{R}^p$, a symmetric $A \in \mathbb{R}^{n \times n}$, what is $\frac{\partial f(x)}{\partial x}$?

Review: Vector calculus

We can also exent the definition to vector functions. If $f(x) = (f_1(x), \ldots, f_q(x))^T$ is a function from $\mathbb{R}^p$ to $\mathbb{R}^q$, then we use the notation

$$\frac{\partial f(x)}{\partial x} \equiv (\frac{\partial f_1(x)}{\partial x}, \ldots, \frac{\partial f_q(x)}{\partial x}) = \begin{pmatrix} \frac{\partial f_1(x)}{\partial x_1} & \cdots & \frac{\partial f_q(x)}{\partial x_1} \\ \vdots & & \vdots \\ \frac{\partial f_1(x)}{\partial x_p} & \cdots & \frac{\partial f_q(x)}{\partial x_p} \end{pmatrix}$$

- For $B \in \mathbb{R}^{q \times p}$ and $x \in \mathbb{R}^p$, we have $\frac{\partial Bx}{\partial x} = B$