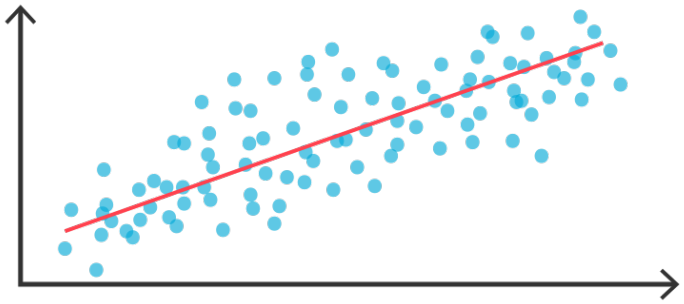
w203: Statistics for Data Science

w203 Instructors

2022-06-18

Contents

Cover



Part I

Probability Theory

Chapter 1

Probability Spaces

Chapter 2

Random Variables

Part II

Learning from Data

Chapter 3

Hypothesis Testing

Chapter 4

Regression

We write a k-vector (of scalars) as a row

$$x = \begin{bmatrix} x_1 & x_2 & ... & x_k \end{bmatrix}.$$

The transpose of x as

$$x^T = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_k \end{bmatrix}.$$

We use uppercase letters $X, Y, Z, ...$ to denote random variables. Random vectors are denoted by bold uppercase letters $X, Y, Z, ...$, and written as a row vector. For example,

$$X = \begin{bmatrix} X_{[1]} & X_{[2]} & ... & X_{[k]} \end{bmatrix}.$$

In order to distinguish random matrices from vectors, a random matrix is denoted by $\mathbb{X}$.

The expectation of X is defined as

$$\mathbb{E}[X] = \begin{bmatrix} \mathbb{E}\left[X_{[1]}\right] & \mathbb{E}\left[X_{[2]}\right] & ... & \mathbb{E}\left[X_{[k]}\right] \end{bmatrix}.$$

The $k \times k$ covariance matrix of X is defined as

$$\begin{aligned} \mathbb{V}[X] &= \mathbb{E}\left[(X - \mathbb{E}[X])^T (X - \mathbb{E}[X])\right] \\ &= \begin{bmatrix} \sigma_1^2 & \sigma_{12} & ... & \sigma_{1k} \\ \sigma_{21} & \sigma_2^2 & ... & \sigma_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{k1} & \sigma_{k2}^2 & ... & \sigma_k^2 \end{bmatrix}_{k \times k} \end{aligned}$$

where $\sigma_j = \mathbb{V}\left[X_{[j]}\right]$ and $\sigma_{ij} = \text{Cov}\left[X_{[i]}, X_{[j]}\right]$ for $i, j = 1, 2, ..., k$ and $i \neq j$.

Theorem 4.1 (Linearity of Exectation). Let $\mathbb{A}_{l\times k}, \mathbb{B}_{m\times l}$ be fixed matrices and c a fixed vector of size l. If X and Y are random vectors of size k and m, respectively, such that $\mathbb{E}[X] < \infty, \mathbb{E}[Y] < \infty$, then

$$\mathbb{E}[\mathbb{A}X + Y\mathbb{B} + c] = \mathbb{A}\mathbb{E}[X] + \mathbb{E}[Y]\mathbb{B} + c.$$

4.1 Conditional Expectation Function

Theorem 4.2 (Characterization of CEF). If $\mathbb{E}[Y^2] < \infty$ and X is a random vector such that $Y = m(X) + e$, then the following statements are equivalent:
1. $m(X) = \mathbb{E}[Y|X]$, the CEF of Y given X
2. $\mathbb{E}[e|X] = 0$

4.2 Best Linear Predictor

Let Y be a random variable and X be a random vector of k variables. We denote the best linear predictor of Y given X by $\mathcal{P}[Y|X]$. It's also called the linear projection of Y on X.

Theorem 4.3 (Best Linear Predictor). Under the following assumptions

1. $\mathbb{E}[Y^2] < \infty$
2. $\mathbb{E}[||\mathrm{X}||^2] < \infty$
3. $\mathbb{Q}_{\mathrm{XX}} \overset{\text{def}}{=} \mathbb{E}\left[X^T X\right]$ is positive-definite

the best linear predictor exists uniquely, and has the form

$$\mathcal{P}[Y|X] = X\beta,$$

where $\beta = \left(\mathbb{E}\left[X^T X\right]\right)^{-1} \mathbb{E}[X^T Y]$ is a column vector.

In the following theorem, we show that the BLP error is uncorrelated to the explanatory variables.

Theorem 4.4 (Best Linear Predictor Error). If the BLP exists, the linear projection error $\varepsilon = Y - \mathcal{P}[Y|X]$ follows the following properties:

1. $\mathbb{E}[X^T \varepsilon] = 0$
2. moreover, $\mathbb{E}[\varepsilon] = 0$ if $X = \begin{bmatrix} 1 & X_{[1]} & \dots & X_{[k]} \end{bmatrix}$ contains a constant.

Chapter 5

Ordinary Least Squares

Let Y be our outcome random variable and

$$X = \begin{bmatrix} 1 & X_{[1]} & X_{[2]} & ... & X_{[k]} \end{bmatrix}$$

be our predictor (or explanatory) vector containing k predictors and a constant. We denote the joint distribution of (Y, X) by $F(y, x)$, i.e.,

$$F(y, x) = \mathbb{P}\left(Y \leq y, X \leq x\right) = \mathbb{P}\left(Y \leq y, X_1 \leq x_1, ..., X_k \leq x_k\right).$$

The dataset or sample is a collection of observations $\{(Y_i, X_i) : i = 1, 2, ..., n\}$. We assume that each observation (Y_i, X_i) is a random (row) vector drawn from the common distribution, sometimes referred to as the population, F.

For a given vector of (unknown) coefficients $\beta = \begin{bmatrix} \beta_0 & \beta_1 & ... & \beta_k \end{bmatrix}^T \in \mathbb{R}^{k+1}$, we define the following cost function:

$$\widehat{S}(\beta) = \frac{1}{n} \sum_{i=1}^{n} (Y_i - X_i \beta)^2.$$

The cost function $\widehat{S}(\beta)$ can also be thought of as the average sum of residuals. In fact, $\widehat{S}(\beta)$ is the moment (plug-in) estimator of the mean squared error,

$$S(\beta) = \mathbb{E}\left[(Y - X\beta)^2\right].$$

We now minimize $\widehat{S}(\beta)$ over all possible choices of $\beta \in \mathbb{R}^{k+1}$. When the minimizer exists and is unique, we call it the least squares estimator, denoted $\widehat{\beta}$.

Definition 5.1 ((Ordinary) Least Squares Estimator). The least square estimator is

$$\widehat{\beta} = \underset{\beta \in \mathbb{R}^{k+1}}{\arg\min}\ \widehat{S}(\beta),$$

provided it exists uniquely.

5.1 Solution of OLS

We rewrite the cost function as

$$\widehat{S}(\beta) = \frac{1}{n} SSE(\beta),$$

where $SSE(\beta) \stackrel{\text{def}}{=} \sum_{i=1}^{n} (Y_i - X_i\beta)^2$.

We now express $SSE(\beta)$ as a quadratic function of β.

$$\begin{aligned}
SSE &= \sum_{i=1}^{n} (Y_i - X_i\beta)^2 \\
&= \sum_{i=1}^{n} Y_i^2 - 2\sum_{i=1}^{n} Y_i(X_i\beta) + \sum_{i=1}^{n} (X_i\beta)^2 \\
&= \sum_{i=1}^{n} Y_i^2 - 2\sum_{i=1}^{n} Y_i(\beta^T {X_i}^T) + \sum_{i=1}^{n} (X_i\beta)(X_i\beta) \\
&= \sum_{i=1}^{n} Y_i^2 - 2\sum_{i=1}^{n} \beta^T(Y_i {X_i}^T) + \sum_{i=1}^{n} (\beta^T {X_i}^T)(X_i\beta) \\
&= \left(\sum_{i=1}^{n} Y_i^2\right) - 2\beta^T \left(\sum_{i=1}^{n} {X_i}^T Y_i\right) + \beta^T \left(\sum_{i=1}^{n} {X_i}^T X_i\right) \beta
\end{aligned}$$

Taking partial derivative w.r.t. β_j, we get

$$\frac{\partial}{\partial \beta_j} SSE(\beta) = -2\left[\sum_{i=1}^{n} {X_i}^T Y_i\right]_j + 2\left[\left(\sum_{i=1}^{n} {X_i}^T X_i\right)\beta\right]_j.$$

Therefore,

$$\frac{\partial}{\partial \beta} SSE(\beta) = -2\left(\sum_{i=1}^{n} {X_i}^T Y_i\right) + 2\left(\sum_{i=1}^{n} {X_i}^T X_i\right)\beta.$$

In order to miniminize $SSE(\beta)$, a necessary condition for $\widehat{\beta}$ is

$$\left.\frac{\partial}{\partial \beta} SSE(\beta)\right|_{\beta=\widehat{\beta}} = 0,$$

i.e.,

$$-2\left(\sum_{i=1}^{n} {X_i}^T Y_i\right) + 2\left(\sum_{i=1}^{n} {X_i}^T X_i\right)\widehat{\beta} = 0$$

So,

$$\left(\sum_{i=1}^{n} {X_i}^T Y_i\right) = \left(\sum_{i=1}^{n} {X_i}^T X_i\right)\widehat{\beta} \tag{5.1}$$

Both the left and right hand side of the above equation are $k+1$ vectors. So, we have a system of $(k+1)$ linear equations with $(k+1)$ unknowns—the elements of β.

Let us define

$$\widehat{\mathbb{Q}}_{XX} = \frac{1}{n}\left(\sum_{i=1}^{n} {X_i}^T X_i\right) \text{ and } \widehat{\mathbb{Q}}_{XY} = \frac{1}{n}\left(\sum_{i=1}^{n} {X_i}^T Y_i\right).$$

Rewriting (5.1), we get

$$\widehat{\mathbb{Q}}_{XY} = \widehat{\mathbb{Q}}_{XX}\hat{\beta}. \tag{5.2}$$

Equation (5.2) is sometimes referred to as the first-order moment condition. For the uniqueness of solution, we require that $\widehat{\mathbb{Q}}_{XX}$ is non-singular. In that case, we can solve for $\hat{\beta}$ to get,

$$\hat{\beta} = \left[\widehat{\mathbb{Q}}_{XX}\right]^{-1}\widehat{\mathbb{Q}}_{XY}.$$

To verify that the above choice minimizes $SSE(\beta)$, one can consider the second-order moment conditions.

$$\frac{\partial^2}{\partial\beta\partial\beta^T}SSE(\beta) = 2\widehat{\mathbb{Q}}_{XX}.$$

If $\widehat{\mathbb{Q}}_{XX}$ is non-singular, it is also positive-definite. So, we have actually proved the following theorem.

Theorem 5.1. If $\widehat{\mathbb{Q}}_{XX}$ is non-singular, then the least squares estimator is unique, and is given by

$$\hat{\beta} = \left[\widehat{\mathbb{Q}}_{XX}\right]^{-1}\widehat{\mathbb{Q}}_{XY}.$$

5.2 Errors and Residuals

Recall that β denotes the coefficients of the best linear predictor 4.3. We first define the fitted value as

$$\widehat{Y_i} = X_i\hat{\beta} \text{ for } i = 1, 2, \ldots, n.$$

For the least squares estimators, we define the errors and residuals in the following way:

$$\varepsilon_i = Y_i - X_i\beta, \text{ and } e_i = Y_i - \widehat{Y_i}.$$

Theorem 5.2 (Least Squares Error). If $\widehat{\mathbb{Q}}_{XX}$ is non-singular, then

1. $\sum_{i=1}^{n} X_i^T e_i = 0$
2. $\sum_{i=1}^{n} e_i = 0$

Proof.

$$\begin{aligned}
\sum_{i=1}^{n} X_i^T e_i &= \sum_{i=1}^{n} X_i^T (Y_i - \widehat{Y}_i) \\
&= \sum_{i=1}^{n} X_i^T Y_i - \sum_{i=1}^{n} X_i^T \widehat{Y}_i \\
&= \sum_{i=1}^{n} X_i^T Y_i - \sum_{i=1}^{n} X_i^T X_i \hat{\beta} \\
&= \widehat{\mathbb{Q}}_{XY} - \widehat{\mathbb{Q}}_{XX} \hat{\beta} \\
&= \widehat{\mathbb{Q}}_{XY} - \widehat{\mathbb{Q}}_{XX} \left(\widehat{\mathbb{Q}}_{XX}^{-1} \widehat{\mathbb{Q}}_{XY} \right) \\
&= 0
\end{aligned}$$

From the first row of (1) we get

$$\sum_{i=1}^{n} e_i = 0.$$

Hence the result. □

5.3 Matrix Notation

Taking the definition of errors from the last section, we can write down a system of n linear equations:

$$\begin{aligned}
Y_1 &= X_1 \beta + \varepsilon_1 \\
Y_2 &= X_2 \beta + \varepsilon_2 \\
&\vdots \\
Y_n &= X_n \beta + \varepsilon_n
\end{aligned}$$

Define

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}_{n \times 1}, \; \mathbb{X} = \begin{bmatrix} 1 & X_{[1]1} & X_{[2]1} & \dots & X_{[k]1} \\ 1 & X_{[1]2} & X_{[2]2} & \dots & X_{[k]2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{[1]n} & X_{[2]n} & \dots & X_{[k]n} \end{bmatrix}, \text{ and } \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}_{n \times 1}.$$

We can now rewrite the system as the following:

$$Y = \mathbb{X} \beta + \varepsilon.$$

We also note that

$$\widehat{\mathbb{Q}}_{XX} = \sum_{i=1}^{n} X_i^T X_i = \mathbb{X}^T \mathbb{X},$$

and

$$\widehat{\mathbb{Q}}_{XY} = \sum_{i=1}^{n} X_i^T Y_i = \mathbb{X}^T Y.$$

So, we have write the least squares estimator as

$$\widehat{\beta} = \left[\mathbb{X}^T\mathbb{X}\right]^{-1}\mathbb{X}^T Y.$$

Similarly, the residual vector is

$$e = Y - \mathbb{X}\widehat{\beta}.$$

As a consequence of 5.2, we can write

$$\mathbb{X}^T e = 0.$$

Chapter 6

Linear Conditional Expectation Function

6.1 Variance of Error

We first compute the (unconditional) variance of the error vector $\boldsymbol{e}$. The covariance matrix

$$\mathbb{V}[\boldsymbol{e}] = \mathbb{E}[\boldsymbol{e}\boldsymbol{e}'] - \mathbb{E}[\boldsymbol{e}]\mathbb{E}[\boldsymbol{e}'] = \mathbb{E}[\boldsymbol{e}\boldsymbol{e}'] \stackrel{\text{def}}{=} \mathbb{D}.$$

For $i \neq j$, the errors e_i,e_j are independent. As a result, $\mathbb{E}\left[e_i e_j\right] = \mathbb{E}[e_i]\mathbb{E}\left[e_j\right] = 0$. So, $\mathbb{D}$ is a diagonal matrix with the i-th diagonal element σ_i^2:

$$\mathbb{D} = \begin{bmatrix} \sigma_1^2 & 0 & ... & 0 \\ 0 & \sigma_2^2 & ... & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & ... & \sigma_n^2 \end{bmatrix}.$$

6.2 Variance of OLS Estimators

Chapter 7

Large-Sample Regression

We assume that the best linear predictor, $\mathcal{P}[Y|X]$, of Y given X is $X\beta$.

$$Y = X\beta + \varepsilon.$$

We have from Theorem 4.4

$$\mathbb{E}[\varepsilon] = 0, \text{ and } \mathbb{E}\left[X^T\varepsilon\right] = 0.$$

We also assume that the dataset $\{(Y_i, X_i)\}$ is taken i.i.d. from the joint distribution of (Y, X). For each i, we can write

$$Y_i = X_i\beta + \varepsilon_i.$$

In matrix notation, we can write

$$Y = \mathbb{X}\beta + \varepsilon.$$

Then

$$\mathbb{E}[\varepsilon] = 0, \text{ and } \mathbb{E}[\varepsilon] = 0$$

7.1 Consistency of OLS Estimators

7.2 Asymptotic Normality

We start by revealing an alternative expression for the OLS estimators $\hat{\beta}$ using matrix notation.

$$\begin{aligned}\widehat{\beta} &= \left[\mathbb{X}^T\mathbb{X}\right]^{-1}\mathbb{X}^T Y \\ &= \left[\mathbb{X}^T\mathbb{X}\right]^{-1}\mathbb{X}^T(\mathbb{X}\beta+\varepsilon) \\ &= \left[\mathbb{X}^T\mathbb{X}\right]^{-1}(\mathbb{X}^T\mathbb{X})\beta + \left[\mathbb{X}^T\mathbb{X}\right]^{-1}\mathbb{X}^T\varepsilon \\ &= \beta + \left[\mathbb{X}^T\mathbb{X}\right]^{-1}\mathbb{X}^T\varepsilon\end{aligned}$$

So,

$$\widehat{\beta} - \beta = \left[\mathbb{X}^T\mathbb{X}\right]^{-1}\mathbb{X}^T\varepsilon \tag{7.1}$$

We can then multiply by $\sqrt{n}$ both sides of Equation (7.1) to get

$$\begin{aligned}\sqrt{n}\left(\widehat{\beta}-\beta\right) &= \left(\frac{1}{n}\sum_{i=1}^{n} X_i^T X_i\right)^{-1}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_i^T\varepsilon_i\right) \\ &= \widehat{\mathbb{Q}}_{XX}^{-1}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_i^T\varepsilon_i\right)\end{aligned}$$

From the consistency of OLS estimators, we already have

$$\widehat{\mathbb{Q}}_{XX} \xrightarrow[p]{} \mathbb{Q}_{XX}$$

Our aim now is to understand the distribution of the stochastic term (the second term) in the above expression.

We first note (from i.i.d. and Theorem 4.4) that

$$\mathbb{E}\left[X_i^T\varepsilon_i\right] = \mathbb{E}\left[X^T\varepsilon\right] = 0.$$

Let us compute the covariance matrix of $X_i\varepsilon_i$. Since the expectation vector is zero, we have

$$\mathbb{V}[X_i^T\varepsilon_i] = \mathbb{E}\left[X_i^T\varepsilon_i\left(X_i^T\varepsilon_i\right)^T\right] = \mathbb{E}\left[X^T X\varepsilon^2\right] \overset{\text{def}}{=} \mathbb{A}.$$

As any function of $\{(Y_i, X_i)\}$'s are independent, $\{X_i\varepsilon_i\}$'s are independent. By the (multivariate) Central Limit Theorem, as $n \to \infty$

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_i^T\varepsilon_i \xrightarrow[d]{} \mathcal{N}(0, \mathbb{A}).$$

There is a small technicality here, we must have $\mathbb{A} < \infty$. This can be imposed by a stronger regularity condition on the moments, e.g., $\mathbb{E}\left[Y^4\right], \mathbb{E}\left[||X||^4\right] < \infty$. Putting everything together, we conclude

$$\sqrt{n}(\widehat{\beta}-\beta) \xrightarrow[d]{} \mathbb{Q}_{XX}^{-1}\mathcal{N}(0,\mathbb{A}) = \mathcal{N}\left(0, \left[\mathbb{Q}_{XX}^{-1}\right]^T\mathbb{A}\mathbb{Q}_{XX}^{-1}\right) = \mathcal{N}\left(0, \mathbb{Q}_{XX}^{-1}\mathbb{A}\mathbb{Q}_{XX}^{-1}\right)$$

Theorem 7.1 (Asymptotic Distribution of OLS Estimators). We assume the following:
1. The observations $\{(Y_i, X_i)\}_{i=1}^n$ are i.i.d from the joint distribution of (Y, X)
2. $\mathbb{E}\left[Y^4\right] < \infty$
3. $\mathbb{E}\left[||X||^4\right] < \infty$
4. $\mathbb{Q}_{XX} = \mathbb{E}\left[XX'\right]$ is positive-definite. Under these assumptions, as $n \to \infty$

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow[d]{} \mathcal{N}\left(0, \mathbb{V}_\beta\right),$$

where

$$\mathbb{V}_\beta \overset{\text{def}}{=} \mathbb{Q}_{XX}^{-1}\mathbb{A}\mathbb{Q}_{XX}^{-1}$$

and $\mathbb{Q}_{XX} = \mathbb{E}\left[X^T X\right]$, $\mathbb{A} = \mathbb{E}\left[X^T X \varepsilon^2\right]$.

The covariance matrix $\mathbb{V}_\beta$ is called the asymptotic variance matrix of $\hat{\beta}$. The matrix is sometimes referred to as the sandwich form.

7.3 Covariance Matrix Estimation

We now turn our attention to the estimation of the sandwich matrix using a finite sample.

7.3.1 Heteroskedastic Variance

Theorem 7.1 presented the asymptotic covariance matrix of $\sqrt{n}(\hat{\beta} - \beta)$ is

$$\mathbb{V}_\beta = \mathbb{Q}_{XX}^{-1}\mathbb{A}\mathbb{Q}_{XX}^{-1}.$$

Without imposing any homoskedasticity condition, we estimate $\mathbb{V}_\beta$ using a plug-in estimator.

We have already seen that $\widehat{\mathbb{Q}}_{XX} = \frac{1}{n}\sum_{i=1}^n X_i^T X_i$ is a natural estimator for $\mathbb{Q}_{XX}$. For $\mathbb{A}$, we use the moment estimator

$$\hat{\mathbb{A}} = \frac{1}{n}\sum_{i=1}^n X_i^T X_i e_i^2,$$

where $e_i = (Y_i - X_i\hat{\beta})$ is the i-th residual. As it turns out, $\hat{\mathbb{A}}$ is a consistent estimator for $\mathbb{A}$.

As a result, we get the following plug-in estimator for $\mathbb{V}_\beta$:

$$\hat{\mathbb{V}}_\beta = \widehat{\mathbb{Q}}_{XX}^{-1}\hat{\mathbb{A}}\widehat{\mathbb{Q}}_{XX}^{-1}$$

The estimator is also consistent. For a proof, see Hensen 2013.

As a consequence, we can get the following estimator for the variance, $\mathbb{V}_{\widehat{\beta}}$, of $\widehat{\beta}$ in the heteroskedastic case.

$$\begin{aligned}
\widehat{\mathbb{V}}\left[\widehat{\beta}\right] &= \frac{1}{n}\widehat{\mathbb{V}}_{\beta}^{\mathrm{HC0}} \\
&= \frac{1}{n}\widehat{\mathbb{Q}}_{XX}^{-1}\widehat{\mathbb{A}}\widehat{\mathbb{Q}}_{XX}^{-1} \\
&= \frac{1}{n}\left(\frac{1}{n}\sum_{i=1}^{n} X_i^T X_i\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} e_i^2 X_i^T X_i\right)\left(\frac{1}{n}\sum_{i=1}^{n} X_i^T X_i\right)^{-1} \\
&= \left(\mathbb{X}^T\mathbb{X}\right)^{-1}\mathbb{X}^T\mathbb{D}\mathbb{X}\left(\mathbb{X}^T\mathbb{X}\right)^{-1}
\end{aligned}$$

where $\mathbb{D}$ is an $n \times n$ diagonal matrix with diagonal entries $e_1^2, e_2^2, \ldots, e_n^2$. The estimator, $\widehat{\mathbb{V}}\left[\widehat{\beta}\right]$, is referred to as the robust error variance estimator for the OLS coefficients $\widehat{\beta}$.

7.3.2 Homeskedastic Variance

Appendix A

Matrix Algebra

In this book, we reserve boldface letter to denote vectors (of scalars and random variables), and "blackboard bold" typeface to denote matrices.

We always write a vector as a column

$$\boldsymbol{v} = \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_k \end{bmatrix}_{k\times 1}$$

Definition A.1 (Transpose of a Matrix). Let $\mathbb{A}_{k\times l}$ be a matrix, it's transpose, denoted $\mathbb{A}^T$, is an $l \times k$ matrix such that the (i, j)-th entry of $\mathbb{A}$ becomes the (j, i)-th entry of $\mathbb{A}^T$.

Definition A.2 (Sum of Matrices). Let $\mathbb{A}, \mathbb{B}$ are matrices both of size $k \times l$, then the sum $\mathbb{A}+\mathbb{B}$ is defined as the another matrix $\mathbb{C}$ size $k \times l$ such that the (i, j)-th entry is the sum of the (i, j)-th entries of $\mathbb{A}$ and $\mathbb{B}$.

$$\mathbb{C} = \begin{bmatrix} a_{11}+b_{11} & a_{12}+b_{12} & ... & a_{1l}+b_{1l} \\ a_{21}+b_{21} & a_{22}+b_{22} & ... & a_{2l}+b_{2l} \\ \vdots & \vdots & \ddots & \vdots \\ a_{k1}+b_{k1} & a_{k2}+b_{k2} & ... & a_{kl}+b_{kl} \end{bmatrix}_{k\times l}$$

Definition A.3 (Product of Matrices). Let $\mathbb{A}, \mathbb{B}$ are matrices both of size $k \times l$, then the sum $\mathbb{A} + \mathbb{B}$ is defined as the another matrix $\mathbb{C}$ size $k \times l$ such that the (i, j)-th entry is the sum of the (i, j)-th entries of $\mathbb{A}$ and $\mathbb{B}$.

$$\mathbb{C} = \begin{bmatrix} a_{11}+b_{11} & a_{12}+b_{12} & ... & a_{1l}+b_{1l} \\ a_{21}+b_{21} & a_{22}+b_{22} & ... & a_{2l}+b_{2l} \\ \vdots & \vdots & \ddots & \vdots \\ a_{k1}+b_{k1} & a_{k2}+b_{k2} & ... & a_{kl}+b_{kl} \end{bmatrix}_{k\times l}$$

Appendix B

Matrix Calculus