Lecture Notes for Linear Models and Extensions

Fall 2025

These lecture notes are based on the book Linear Models and Extensions *by Peng Ding (2025).*

Contents

Preface

The importance of linear models

A central goal in statistics is to use data to build models to make inferences about the underlying data-generating processes or make predictions of future observations. Although real problems are often complex, the linear model can often serve as a good approximation to the true data-generating process. Sometimes, although the true data-generating process is nonlinear, the linear model can be a useful approximation if we properly transform the data based on domain knowledge.

Moreover, linear models possesses elegant algebraic and geometric properties. They often admit explicit formulas that provide deep insights into various aspects of statistical modeling and learning. For more complicated models, such closed-form expressions may be impossible. Nonetheless, linear models offer intuition for more complicated models. For example, the double-descent phenomenon—originally observed empirically in deep learning (Belkin et al., 2019)—can be rigorously examined and proved within the framework of linear models (Hastie et al., 2022). In our experience, only in rare cases the insights gained from linear models do not apply to more complicated models.

From a pedagogical perspective, the linear model plays a foundational role and serves as a building block for broader statistical training. These lecture notes closely follow Ding (2024). The book is freely available at https://arxiv.org/pdf/2401.00649v2.

R and Python

`R` is widely used in the statistics community and offers a rich ecosystem of packages for statistical modeling. Some commonly used datasets are available only in `R` but not in `Python`. For example, the Galton's dataset can be found in the `R` package `HistData`, but it is not natively available in `Python`. However, you can access and use datasets from `R` packages within `Python` via the `statsmodels` library. Specifically, `statsmodels` provides the function `sm.datasets.get_rdataset()`, which allows you to directly load datasets from `R` packages such as `HistData`.

Here is an example of how to load the Galton dataset from the `HistData` package in `Python`:

```
import statsmodels.api as sm

# Load the ’Guerry’ dataset from the ’HistData’ R package
data = sm.datasets.get_rdataset("GaltonFamilies", "HistData").data

# Now you can work with the ’data’ DataFrame in Python
print(data.head())
```

Prerequisites and corequisites

These lecture notes assume that the reader has basic training in linear algebra, probability theory, and statistical inference.

1

Introduction and Motivations

This book is about the linear model and its extensions. Before delving into the mathematical details of specific models, we will briefly provide some motivations for studying statistical models.

1.1 Data and statistical models

A wide range of problems in statistics and machine learning have the following data structure:

Unit	outcome/response	covariates/features/predictors			
i	Y	X_1	X_2	$\cdots$	X_p
1	y_1	x_{11}	x_{12}	$\cdots$	x_{1p}
2	y_2	x_{21}	x_{22}	$\cdots$	x_{2p}
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$
n	y_n	x_{n1}	x_{n2}	$\cdots$	x_{np}

For each unit i, we observe the outcome of interest (also called the response), y_i, as well as p covariates (also called features or predictors), $x_{i1}, \ldots, x_{ip}$. We often use

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

to denote the n-dimensional outcome vector, and

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}$$

to denote the $n \times p$ covariate matrix, also called the *design matrix*. In most cases, the first column of X contains constants 1s.

Based on the data (X, Y), we can ask the following questions:

(Q1) Describe the relationship between X and Y, i.e., their association or correlation. For example, how is the patients' average height related to the children's average height? How is one's height related to one's weight? How are one's education and working experience related to one's income?

(Q2) Predict Y^* with new data X^*, based on the old data (X, Y). In particular, we want to use the current data (X, Y) to train a predictor, and then use it to predict future Y^* based on future X^*. This is called *supervised learning* in the field of machine learning. For example, how do we predict whether an email is spam or not based on the frequencies of the most commonly occurring words and punctuation marks in the email? How do we predict cancer patients' survival time based on their clinical measures?

(Q3) Estimate the causal effect of some components in X on Y. What if we change some components of X? How do we measure the impact of the hypothetical intervention of some components of X on Y? This is a much harder question because most statistical tools are designed to infer association, not causation. For example, the U.S. Food and Drug Administration (FDA) approves drugs based on randomized controlled trials (RCT) because RCTs are most credible to infer causal effects of drugs on health outcomes. Economists are interested in evaluating the effect of a job training program on employment and wages. However, this is a notoriously difficult problem because participation in the job training program is not randomized in observational data.

The above descriptions are about generic X and Y, which can have many different types. We often use different statistical models to capture the features of different types of data. Below we give a brief overview of models that will appear in later parts of this book.

(T1) X and Y are univariate and continuous. In Francis Galton's[1] classic example, X is the parents' average height and Y is the children's average height (Galton, 1886). Let $\widehat{y}_i$ denote the "fitted value" of the outcome for unit i with covariate value x_i. Galton derived the following formula:

$$\widehat{y}_i = \bar{y} + \widehat{\rho} \frac{\widehat{\sigma}_y}{\widehat{\sigma}_x}(x_i - \bar{x})$$

which is equivalent to

$$\frac{\widehat{y} - \bar{y}}{\widehat{\sigma}_y} = \widehat{\rho} \frac{x_i - \bar{x}}{\widehat{\sigma}_x}, \tag{1.1}$$

where

$$\bar{x} = n^{-1} \sum_{i=1}^{n} x_i, \qquad \bar{y} = n^{-1} \sum_{i=1}^{n} y_i$$

are the sample means,

$$\widehat{\sigma}_x^2 = (n-1)^{-1} \sum_{i=1}^{n} (x_i - \bar{x})^2, \qquad \widehat{\sigma}_y^2 = (n-1)^{-1} \sum_{i=1}^{n} (y_i - \bar{y})^2$$

are the sample variances, and $\widehat{\rho} = \widehat{\sigma}_{xy} / (\widehat{\sigma}_x \widehat{\sigma}_y)$ is the sample Pearson correlation coefficient with the sample covariance

$$\widehat{\sigma}_{xy} = (n-1)^{-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}).$$

The identity (1.1) is the famous formula of "regression towards mediocrity" or "regression towards the mean". Galton first introduced the terminology "regression."[2] Galton

[1] Who was Francis Galton? He was Charles Darwin's nephew and was famous for his pioneer work in statistics and for devising a method for classifying fingerprints that proved useful in forensic science. He also invented the term *eugenics*, a field that causes a lot of controversies nowadays.

[2] The name "regression" is widely used in statistics now. For instance, we sometimes use "linear regression" interchangeably with "linear model." We also extend the name to "logistic regression" or "Cox regression."

called regression because the relative deviation of the children's average height is smaller than that of the parents' average height if $|\widehat{\rho}| < 1$. We will derive (1.1) in Chapter 2.

(T2) Y is univariate and continuous, and X is multivariate of mixed types. In the `R` package `ElemStatLearn`, the dataset `prostate` has an outcome of interest as the log of the prostate-specific antigen `lpsa` and some potential predictors including the log cancer volume `lcavol`, the log prostate weight `lweight`, age `age`, etc. Linear regression with multidimensional covariates will be one of our main focus in this course.

(T3) Y is binary or indicator of two classes, and X is multivariate of mixed types. For example, in the `R` package `wooldridge`, the dataset `mroz` contains an outcome of interest being the binary indicator for whether a woman was in the labor force in the year 1975, and some useful covariates are in the table below:

covariate name	covariate meaning
kidslt6	number of kids younger than six years old
kidsge6	number of kids between six and eighteen years old
age	age
educ	years of education
husage	husband's age
huseduc	husband's years of education

A commonly used model is *logistic regression* for binary outcomes.

(T4) Y is categorical without ordering. For example, the choice of housing type, single-family house, townhouse, or condominium, is a categorical variable. A commonly used model is the *multinomial logistic regression* for categorical outcomes without ordering.

(T5) Y is categorical and ordered. For example, the final course evaluation at UC Berkeley can take value in $\{1, 2, 3, 4, 5, 6, 7\}$. These numbers have clear ordering but they are not the usual real numbers. A commonly used model is the *proportional odds regression* for ordered categorical outcomes.

(T6) Y represents counts. For example, the number of times one went to the gym last week is a non-negative integer representing counts. Commonly used models are Poisson/negative-binomial/zero-inflated regression.

(T7) Y is multivariate and correlated. In medical trials, the data are often longitudinal, meaning that the patient's outcomes are measured repeatedly over time. So each patient has a multivariate outcome. In field experiments of public health and development economics, the randomized interventions are often at the village level but the outcome data are collected at the household level. So within villages, the outcomes are correlated. A commonly used model is the the *generalized estimating equation* for correlated data.

(T8) Y represent time-to-event outcomes. For example, in medical trials, a major outcome of interest is the survival time; in labor economics, a major outcome of interest is the time to find the next job. The former is called *survival analysis* in biostatistics and the latter is called *duration analysis* in econometrics.

1.2 Why linear models?

Why do we study linear models if many real problems may have nonlinear structures? There are important reasons.

(R1) Linear models are simple but non-trivial starting points for learning.

(R2) Linear models can provide insights because we can derive explicit formulas based on elegant algebra and geometry.

(R3) Linear models can handle nonlinearity by incorporating nonlinear terms of covariates, for example, X can contain the polynomials or nonlinear transformations of the original covariates. In statistics, "linear" means *linear in parameters*, not *linear in covariates*.

(R4) Linear models can be good approximations of nonlinear data-generating processes.

(R5) Linear models are simpler than nonlinear models, but they do not necessarily perform worse than more complicated nonlinear models. We have finite data so we cannot fit arbitrarily complicated models.

(R6) Linear models offer insights for more complicated models. In our experience, only in rare cases the insights gained from linear models do not apply to more complicated models.

2

Simple Linear Regression

Simple linear regression refers to linear regression with a single covariate. This chapter discusses ordinary least squares (OLS) with a single covariate. It can provide insights into later chapters because it is the building block of OLS with multiple covariates.

2.1 Ordinary least squares with a univariate covariate

Figure 2.1 shows the scatterplot of Galton's dataset which can be found in the `R` package `HistData` as `GaltonFamilies`. In this dataset, `father` denotes the height of the father and `mother` denotes the height of the mother. The x-axis denotes the mid-parent height, calculated as (`father` + 1.08*`mother`)/2, and the y-axis denotes the height of a child.

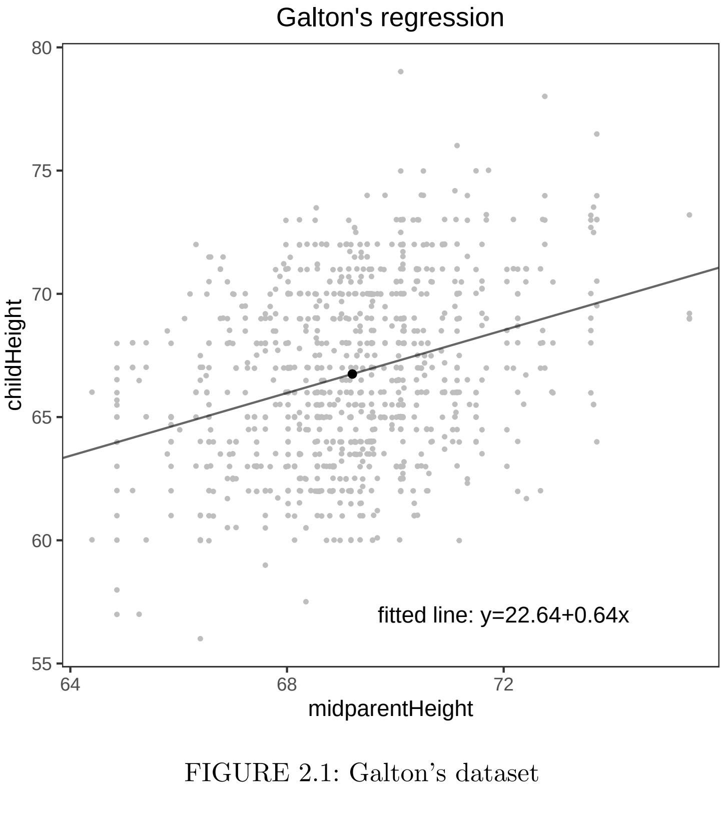

FIGURE 2.1: Galton's dataset

With n data points $(x_i,y_i)_{i=1}^n$, our goal is to find the best linear fit of the data

$$(x_i,\widehat{y}_i = \widehat{\alpha} + \widehat{\beta}x_i)_{i=1}^n,$$

where the coefficients $\widehat{\alpha}$ and $\widehat{\beta}$ are determined from the data. What do we mean by the "best" fit? Gauss proposed to use the following OLS criterion:[1]

$$(\widehat{\alpha}, \widehat{\beta}) = \arg\min_{a,b} n^{-1} \sum_{i=1}^n (y_i - a - bx_i)^2. \tag{2.1}$$

The OLS criterion is based on the squared "misfits" $y_i - a - bx_i$. Another intuitive criterion is based on the absolute values of those misfits, which is called the least absolute deviation (LAD). However, OLS is simpler because the objective function is smooth in (a, b), and we can obtain close-form solutions.

How do we solve the OLS minimization problem in (2.1)? The objective function is quadratic, and as a and b diverge, it diverges to infinity. So it must have a minimizer $(\widehat{\alpha}, \widehat{\beta})$, which satisfies the first-order condition:

$$-\frac{2}{n}\sum_{i=1}^n (y_i - \widehat{\alpha} - \widehat{\beta}x_i) = 0, \tag{2.2}$$

$$-\frac{2}{n}\sum_{i=1}^n x_i(y_i - \widehat{\alpha} - \widehat{\beta}x_i) = 0. \tag{2.3}$$

The equations (2.2) and (2.3) are called the Normal Equations of OLS. The first equation (2.2) implies

$$\bar{y} = \widehat{\alpha} + \widehat{\beta}\bar{x}, \tag{2.4}$$

that is, the OLS line must go through the sample mean of the data $(\bar{x}, \bar{y})$. The second equation (2.3) implies

$$\overline{xy} = \widehat{\alpha}\bar{x} + \widehat{\beta}\overline{x^2}, \tag{2.5}$$

where $\overline{xy}$ is the sample mean of the x_iy_i's, and $\overline{x^2}$ is the sample mean of the x_i^2's. Subtracting (2.4)$\times\bar{x}$ from (2.5), we have

$$\overline{xy} - \bar{x}\bar{y} = \widehat{\beta}(\overline{x^2} - \bar{x}^2),$$

which is

$$\widehat{\sigma}_{xy} = \widehat{\beta}\widehat{\sigma}_x^2,$$

and implies

$$\widehat{\beta} = \frac{\widehat{\sigma}_{xy}}{\widehat{\sigma}_x^2}. \tag{2.6}$$

So the OLS coefficient of x equals the sample covariance between x and y divided by the sample variance of x. From (2.4), we obtain that

$$\widehat{\alpha} = \bar{y} - \widehat{\beta}\bar{x}. \tag{2.7}$$

By (2.7), the fitted line $\widehat{y}_i = \widehat{\alpha} + \widehat{\beta}x_i$ simplifies to $\widehat{y}_i = \bar{y} - \widehat{\beta}\bar{x} + \widehat{\beta}x_i$, and more symmetrically, $\widehat{y}_i - \bar{y} = \widehat{\beta}(x_i - \bar{x})$. With (2.6), we can further simplify the fitted line as

$$\widehat{y}_i - \bar{y} = \frac{\widehat{\sigma}_{xy}}{\widehat{\sigma}_x^2}(x_i - \bar{x}) = \frac{\widehat{\rho}_{xy}\widehat{\sigma}_x\widehat{\sigma}_y}{\widehat{\sigma}_x^2}(x_i - \bar{x}),$$

[1] The idea of OLS is often attributed to Gauss and Legendre. Gauss used it in the process of discovering Ceres, and his work was published in 1809. Legendre's work appeared in 1805 but Gauss claimed that he had been using it since 1794 or 1795. Stigler (1981) reviews the history of OLS.

which implies

$$\frac{\widehat{y}_i - \bar{y}}{\widehat{\sigma}_y} = \widehat{\rho}_{xy} \frac{x_i - \bar{x}}{\widehat{\sigma}_x},$$

the Galtonian formula mentioned in Chapter 1.

We can obtain the fitted line based on Galton's data using the R code below.

```
> library("HistData")
> xx = GaltonFamilies$midparentHeight
> yy = GaltonFamilies$childHeight
>
> center_x = mean(xx)
> center_y = mean(yy)
> sd_x     = sd(xx)
> sd_y     = sd(yy)
> rho_xy   = cor(xx, yy)
>
> beta_fit  = rho_xy*sd_y/sd_x
> alpha_fit = center_y - beta_fit*center_x
> alpha_fit
[1] 22.63624
> beta_fit
[1] 0.6373609
```

Since the dataset is not natively available in Python, we can use the statsmodels package to load the dataset from the R package HistData, and then carry out the rest calculations in Python. The code below shows how to do this.

```
import numpy as np
import statsmodels.api as sm

# Load the Galton Families dataset from statsmodels
galton = sm.datasets.get_rdataset("GaltonFamilies", "HistData").data

# Extract midparentHeight and childHeight
xx = galton["midparentHeight"].values
yy = galton["childHeight"].values

# Compute statistics
center_x = np.mean(xx)
center_y = np.mean(yy)
sd_x     = np.std(xx, ddof=1)  # sample standard deviation (same as R)
sd_y     = np.std(yy, ddof=1)
rho_xy   = np.corrcoef(xx, yy)[0, 1]

# Regression slope and intercept (beta, alpha)
beta_fit  = rho_xy * sd_y / sd_x
alpha_fit = center_y - beta_fit * center_x


print("alpha_fit:", f"{alpha_fit:␣.6f}")
print("beta_fit:", f"{beta_fit:␣.6f}")
```

The outputs are:

```
alpha_fit: 22.636241
beta_fit: 0.637361
```

We then generate Figure 2.1 based on the original data and the OLS coefficients.

2.2 Final comments

I make two final comments on OLS.

(C1) We can write the sample mean as the solution to the OLS with only the intercept:

$$\bar{y} = \arg\min_{\mu} n^{-1} \sum_{i=1}^{n} (y_i - \mu)^2. \tag{2.8}$$

(C2) We can fit OLS of y_i on x_i without the intercept:

$$\widehat{\beta} = \arg\min_{b} n^{-1} \sum_{i=1}^{n} (y_i - b x_i)^2$$

which equals

$$\widehat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2} = \frac{\langle x, y \rangle}{\langle x, x \rangle}, \tag{2.9}$$

where $x = (x_1, \ldots, x_n)^{\mathrm{T}}$ and $y = (y_1, \ldots, y_n)^{\mathrm{T}}$ are the n-dimensional vectors containing all observations, and $\langle x, y \rangle = \sum_{i=1}^{n} x_i y_i$ denotes the inner product between x and y.

Although it is rare to fit the above OLS in practical data analysis, the formulas in (2.8) and (2.9) will be the building blocks for many discussions in later parts of the book. I leave the proof of (2.8) and (2.9) as Problem 2.1.

2.3 Homework problems

2.1 Univariate OLS

Prove (2.8) and (2.9).

2.2 Pairwise slopes

Prove Theorem 2.1 below.

Theorem 2.1 *Given $(x_i, y_i)_{i=1}^{n}$ with univariate x_i and y_i, show that $\widehat{\beta}$ in (2.6) equals*

$$\widehat{\beta} = \sum_{(i,j)} w_{ij} b_{ij},$$

where the summation is over all pairs of observations (i, j),

$$b_{ij} = \frac{y_i - y_j}{x_i - x_j}$$

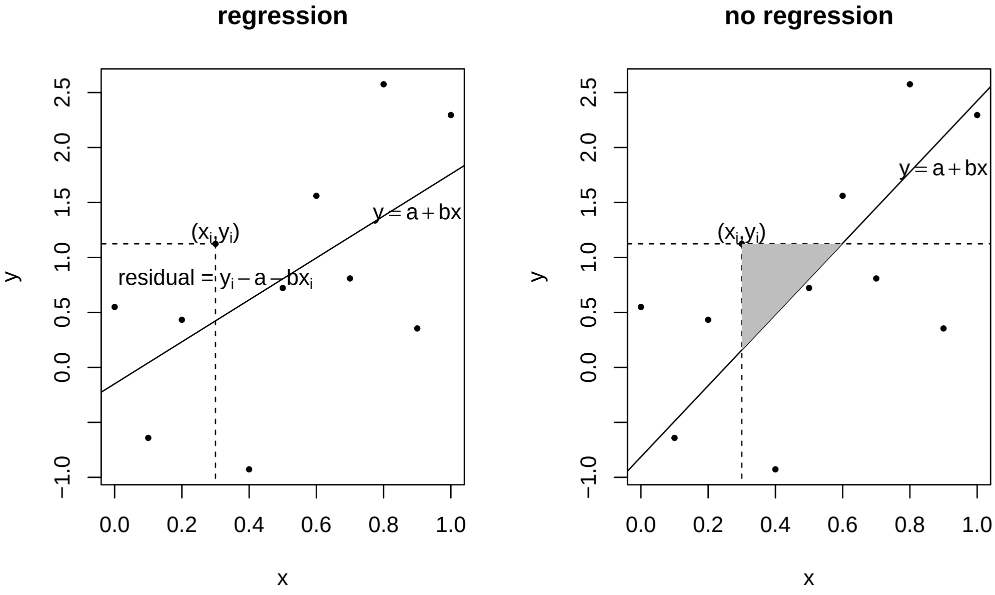

FIGURE 2.2: Regression (left) and no regression (right)

is the slope determined by two points (x_i, y_i) *and* (x_j, y_j)*, and*

$$w_{ij} = \frac{x_i - x_j)^2}{\sum_{(i',j')}(x_{i'} - x_{j'})^2}$$

is the weight proportional to the squared distance between x_i *and* x_j*. In the above formulas, if* $x_i = x_j$*, then we define* $b_{ij} = 0$*, and the corresponding weight* w_{ij} *equals* 0.

Remark: Wu (1986) and Gelman and Park (2009) used Theorem 2.1. Problem 3.10 gives a more general result.

2.3 No regression

Woolley (1941) proposed a method to minimize the sum of the areas formed between the data points and fitted line. The right panel of Figure 2.2 illustrates the area formed between data point (x_i, y_i) and the line $y = a + bx$.

Prove that if $\widehat{\rho}_{xy} > 0$, then the minimizer $(\widehat{\alpha}', \widehat{\beta}')$ satisfies

$$\widehat{\beta}' = \frac{\widehat{\sigma}_y}{\widehat{\sigma}_x}$$

and

$$\widehat{\alpha}' = \bar{y} - \widehat{\beta}\bar{x}.$$

Remark: The fitted line is

$$\begin{aligned}\widehat{y}_i &= \widehat{\alpha}' + \widehat{\beta}' x_i \\ &= \bar{y} - \widehat{\beta}\bar{x} + \widehat{\beta}' x_i,\end{aligned}$$

which is equivalent to

$$\frac{\widehat{y}_i - \bar{y}}{\widehat{\sigma}_y} = \frac{x_i - \bar{x}}{\widehat{\sigma}_x}.$$

It does not have the regression factor $\widehat{\rho}_{xy}$, compared with the Galtonian formula, which is derived by minimizing the residual sum of squares as illustrated by the left panel of Figure 2.2.

3

Multiple Linear Regression

Multiple linear regression refers to linear regression with multiple covariates. This chapter provides algebraic results about ordinary least squares (OLS). The results in this chapter do not rely on any stochastic assumptions.

3.1 The OLS formula

Recall that we have the outcome vector and covariate matrix:

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \qquad X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}.$$

Depending on the purpose, it is convenient to view X as a collection of row or column vectors:

$$X = \begin{pmatrix} x_1^{\mathrm{T}} \\ x_2^{\mathrm{T}} \\ \vdots \\ x_n^{\mathrm{T}} \end{pmatrix} = (X_1, \ldots, X_p)$$

where $x_i^{\mathrm{T}} = (x_{i1}, \ldots, x_{ip})$ is the row vector consisting of the covariates of unit i, and $X_j = (x_{1j}, \ldots, x_{nj})^{\mathrm{T}}$ is the column vector of the j-th covariate for all units.

We want to find the "best" linear fit of the data $(x_i, \widehat{y}_i)_{i=1}^n$ with

$$\widehat{y}_i = x_i^{\mathrm{T}} \widehat{\beta} = \widehat{\beta}_1 x_{i1} + \cdots + \widehat{\beta}_p x_{ip}$$

in the sense that

$$\begin{aligned} \widehat{\beta} &= \arg\min_{b \in \mathbb{R}^p} n^{-1} \sum_{i=1}^n (y_i - x_i^{\mathrm{T}} b)^2 & (3.1) \\ &= \arg\min_{b \in \mathbb{R}^p} n^{-1} \|Y - Xb\|^2, & (3.2) \end{aligned}$$

where $\widehat{\beta} = (\widehat{\beta}_1, \ldots, \widehat{\beta}_p)^{\mathrm{T}}$ is called the OLS coefficient, the $\widehat{y}_i$'s are called the fitted values, and the $\widehat{\varepsilon}_i = y_i - \widehat{y}_i$'s are called the residuals.

The objective function in (3.1) is quadratic in b, which diverges to infinity when b diverges to infinity. So it must have a minimizer $\widehat{\beta}$ satisfying the first-order condition

$$-\frac{2}{n} \sum_{i=1}^n x_i (y_i - x_i^{\mathrm{T}} \widehat{\beta}) = 0,$$

which simplifies to

$$\sum_{i=1}^{n} x_i(y_i - x_i^{\mathrm{T}}\widehat{\beta}) = 0, \tag{3.3}$$

or, equivalently, in matrix form:

$$X^{\mathrm{T}}(Y - X\widehat{\beta}) = 0. \tag{3.4}$$

The above equations (3.3) and (3.4) are called the *Normal equation* of the OLS, which implies the main theorem:

Theorem 3.1 *The OLS coefficient in* (3.1) *and* (3.2) *equals*

$$\begin{aligned} \widehat{\beta} &= \left(\sum_{i=1}^{n} x_i x_i^{\mathrm{T}}\right)^{-1} \left(\sum_{i=1}^{n} x_i y_i\right) \\ &= (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}Y \end{aligned}$$

if $X^{\mathrm{T}}X = \sum_{i=1}^{n} x_i x_i^{\mathrm{T}}$ *is non-degenerate.*

Comment on the two equivalent forms in Theorem 3.1. The equivalence of the two forms of the OLS coefficient follows from

$$X^{\mathrm{T}}X = (x_1, \ldots, x_n) \begin{pmatrix} x_1^{\mathrm{T}} \\ x_2^{\mathrm{T}} \\ \vdots \\ x_n^{\mathrm{T}} \end{pmatrix} = \sum_{i=1}^{n} x_i x_i^{\mathrm{T}},$$

and

$$X^{\mathrm{T}}Y = (x_1, \ldots, x_n) \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \sum_{i=1}^{n} x_i y_i.$$

Depending on the purpose, both forms can be useful in later discussions.

Comment on the condition in Theorem 3.1. The non-degeneracy of $X^{\mathrm{T}}X$ in Theorem 3.1 requires that for any non-zero vector $\alpha = (\alpha_1, \ldots, \alpha_p)^{\mathrm{T}} \in \mathbb{R}^p$, we must have

$$\alpha^{\mathrm{T}}X^{\mathrm{T}}X\alpha = \|X\alpha\|^2 \neq 0$$

which is equivalent to

$$X\alpha = \alpha_1 X_1 + \cdots + \alpha_p X_p \neq 0,$$

i.e., the columns of X are *linearly independent.*[1] This effectively rules out redundant columns in the design matrix X. If X_1 can be represented by other columns $X_1 = c_2X_2 + \cdots + c_pX_p$ for some $(c_2, \ldots, c_p)$, then $X^{\mathrm{T}}X$ is degenerate.

Throughout the book, we invoke the following condition unless stated otherwise.

Condition 3.1 *The column vectors of X are linearly independent.*

[1] This book uses different notions of “independence,” which can be confusing sometimes. In linear algebra, a set of vectors is linearly *independent* if any nonzero linear combination of them is not zero; see Appendix A. In probability theory, two random variables are *independent* if their joint density factorizes into the product of the marginal distributions; see Appendix B.

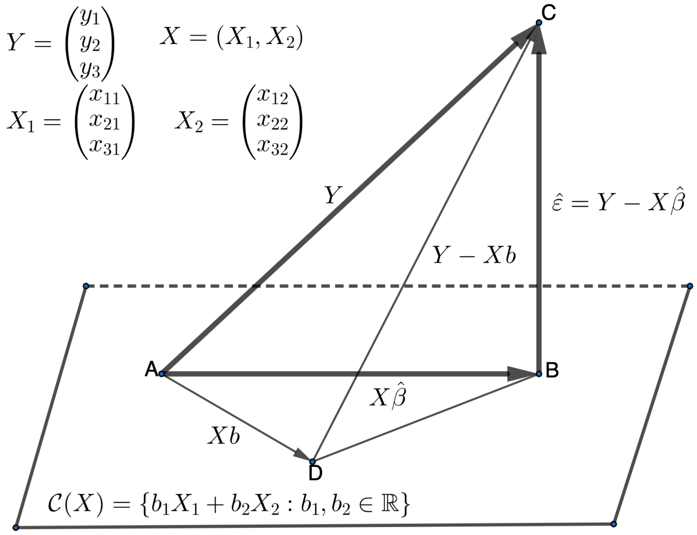
$Y = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix}$
$X = (X_1, X_2)$
$X_1 = \begin{pmatrix} x_{11} \\ x_{21} \\ x_{31} \end{pmatrix}$
$X_2 = \begin{pmatrix} x_{12} \\ x_{22} \\ x_{32} \end{pmatrix}$
C
Y
$Y - Xb$
$\hat{\varepsilon} = Y - X\hat{\beta}$
A
B
$X\hat{\beta}$
Xb
D
$\mathcal{C}(X) = \{b_1X_1 + b_2X_2 : b_1, b_2 \in \mathbb{R}\}$

FIGURE 3.1: The geometry of OLS

3.2 The geometry of OLS

The OLS has clear geometric interpretations. Figure 3.1 illustrate its geometry with $n = 3$ and $p = 2$. For any $b = (b_1, \dots, b_p)^{\mathrm{T}} \in \mathbb{R}^p$ and $X = (X_1, \dots, X_p) \in \mathbb{R}^{n \times p}$,

$$Xb = (X_1, \dots, X_p) \begin{pmatrix} b_1 \\ \vdots \\ b_p \end{pmatrix} = b_1X_1 + \cdots + b_pX_p$$

represents a linear combination of the column vectors of the design matrix X. So the OLS problem is to find the best linear combination of the column vectors of X to approximate the response vector Y. Recall that all linear combinations of the column vectors of X constitute the column space of X, denoted by[2]

$$\mathcal{C}(X) = \{b_1X_1 + \cdots + b_pX_p : b_1, \dots, b_p \in \mathbb{R}\}.$$

So the OLS problem is to find the vector in $\mathcal{C}(X)$ that is the closest to Y. Geometrically, the vector must be the projection of Y onto $\mathcal{C}(X)$. By projection, the residual vector $\widehat{\varepsilon} = Y - X\widehat{\beta}$ must be orthogonal to $\mathcal{C}(X)$, or, equivalently, the residual vector is orthogonal to $X_1, \dots, X_p$. This geometric intuition implies that

$$X_1^{\mathrm{T}}\widehat{\varepsilon} = 0, \dots, X_p^{\mathrm{T}}\widehat{\varepsilon} = 0.$$

In matrix form, we have

$$X^{\mathrm{T}}\widehat{\varepsilon} = \begin{pmatrix} X_1^{\mathrm{T}}\widehat{\varepsilon} \\ \vdots \\ X_p^{\mathrm{T}}\widehat{\varepsilon} \end{pmatrix} = 0,$$

[2] Please review Appendix A for some basic linear algebra background.

which is equivalent to

$$X^{\mathrm{T}}(Y - X\widehat{\beta}) = 0,$$

the Normal equation in (3.4). The above argument gives a geometric derivation of the OLS formula in Theorem 3.1.

In Figure 3.1, since the triangle ABC is rectangular, the fitted vector $\widehat{Y} = X\widehat{\beta}$ is orthogonal to the residual vector $\widehat{\varepsilon}$, and moreover, the Pythagorean Theorem implies that

$$\|Y\|^2 = \|X\widehat{\beta}\|^2 + \|\widehat{\varepsilon}\|^2.$$

The following theorem states an algebraic fact that gives an alternative proof of the OLS formula. It is essentially the Pythagorean Theorem for the rectangular triangle BCD in Figure 3.1.

Theorem 3.2 *For any $b \in \mathbb{R}^p$, we have the following decomposition*

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

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

Proof of Theorem 3.2: We have the following decomposition:

$$\begin{aligned}
\|Y - Xb\|^2 &= (Y - Xb)^{\mathrm{T}}(Y - Xb) \\
&= (Y - X\widehat{\beta} + X\widehat{\beta} - Xb)^{\mathrm{T}}(Y - X\widehat{\beta} + X\widehat{\beta} - Xb) \\
&= (Y - X\widehat{\beta})^{\mathrm{T}}(Y - X\widehat{\beta}) + (X\widehat{\beta} - Xb)^{\mathrm{T}}(X\widehat{\beta} - Xb) \\
&\quad + (Y - X\widehat{\beta})^{\mathrm{T}}(X\widehat{\beta} - Xb) + (X\widehat{\beta} - Xb)^{\mathrm{T}}(Y - X\widehat{\beta}).
\end{aligned}$$

The first term equals $\|Y - X\widehat{\beta}\|^2$ and the second term equals $\|X(\widehat{\beta} - b)\|^2$. We need to show the last two terms are zero. By symmetry of these two terms, we only need to show that the last term is zero. This is true by the Normal equation (3.4) of the OLS:

$$(X\widehat{\beta} - Xb)^{\mathrm{T}}(Y - X\widehat{\beta}) = (\widehat{\beta} - b)^{\mathrm{T}}X^{\mathrm{T}}(Y - X\widehat{\beta}) = 0.$$

□

I end this section by commenting on the role of the intercept in OLS.

The role of the intercept in OLS. In most applications, X contains a column of $1_n = (1, \ldots, 1)^{\mathrm{T}}$. In those cases, we have

$$1_n^{\mathrm{T}}\widehat{\varepsilon} = 0,$$

and therefore,

$$n^{-1}\sum_{i=1}^{n}\widehat{\varepsilon}_i = 0.$$

That is, with the intercept in OLS, the residuals are automatically centered to have mean 0.

3.3 The projection matrix from OLS

The geometry in Section 3.2 also shows that $\widehat{Y} = X\widehat{\beta}$ is the solution to the following problem

$$\widehat{Y} = \arg\min_{v \in \mathcal{C}(X)} \|Y - v\|^2.$$

Using Theorem 3.1, we have $\widehat{Y} = X\widehat{\beta} = HY$, where

$$H = X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}$$

is an $n \times n$ matrix. It is called the *hat matrix* because it puts a hat on Y when multiplying Y. Algebraically, we can show that H is a projection matrix[3] because

$$\begin{aligned} H^2 &= X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}} \\ &= X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}} \\ &= H, \end{aligned}$$

and

$$\begin{aligned} H^{\mathrm{T}} &= \{X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}\}^{\mathrm{T}} \\ &= X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}} \\ &= H. \end{aligned}$$

Its rank equals its trace, so

$$\begin{aligned} \text{rank}(H) = \text{trace}(H) &= \text{trace}\{X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}\} \\ &= \text{trace}\{(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}X\} \\ &= \text{trace}(I_p) \\ &= p. \end{aligned}$$

The projection matrix H has the following geometric interpretations.

Proposition 3.1 *The projection matrix $H = X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}$ satisfies*

(G1) $Hv = v$ if and only if $v \in \mathcal{C}(X)$;

(G2) $Hw = 0$ if and only if $w \perp \mathcal{C}(X)$.

Recall that $\mathcal{C}(X)$ is the column space of X. (G1) states that projecting any vector in $\mathcal{C}(X)$ onto $\mathcal{C}(X)$ does not change the vector. (G2) states that projecting any vector orthogonal to $\mathcal{C}(X)$ onto $\mathcal{C}(X)$ results in a zero vector.

Proof of Proposition 3.1: I first prove (G1). If $v \in \mathcal{C}(X)$, then $v = Xb$ for some b, which implies

$$Hv = X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}Xb = Xb = v.$$

Conversely, if $v = Hv$, then $v = X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}v = Xb$ with $b = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}v$, which ensures $v \in \mathcal{C}(X)$.

I then prove (G2). If $w \perp \mathcal{C}(X)$, then w is orthogonal to all column vectors of X, that is, $X_j^{\mathrm{T}}w = 0 \quad (j = 1, \ldots, p)$. In matrix form, we have $X^{\mathrm{T}}w = 0$, which implies

$$Hw = X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}w = 0.$$

Conversely, if $Hw = X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}w = 0$, then $w^{\mathrm{T}}X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}w = 0$. Because $(X^{\mathrm{T}}X)^{-1}$ is positive definite under Condition 3.1, we have $X^{\mathrm{T}}w = 0$, which implies $w \perp \mathcal{C}(X)$. □

Writing $H = (h_{ij})_{1 \leq i,j \leq n}$ and $\widehat{y} = (\widehat{y}_1, \ldots, \widehat{y}_n)^{\mathrm{T}}$, we have another basic identity

$$\begin{aligned} \widehat{y}_i &= \sum_{j=1}^{n} h_{ij}y_j \\ &= h_{ii}y_i + \sum_{j \neq i} h_{ij}y_j. \end{aligned}$$

[3] Review the definition and properties of projection matrices in Appendix A.

It shows that the predicted value $\widehat{y}_i$ is a linear combination of the outcomes of all units and the coefficients depend on H. Moreover, if X contains a column of intercepts $1_n = (1, \ldots, 1)^{\mathrm{T}}$, then

$$H 1_n = 1_n, \tag{3.5}$$

which implies

$$\sum_{j=1}^{n} h_{ij} = 1 \quad (i = 1, \ldots, n) \tag{3.6}$$

and therefore, $\widehat{y}_i$ is a weighted average of the outcomes of all units. Although the sum of the weights is 1, some of them can be negative. Readers, make sure the claims of (3.5) and (3.6) make sense to you. See Problem 3.6.

In general, the hat matrix has complex forms, but when the covariates are dummy variables for group indicators, it has more explicit forms. I give two examples below.

Example 3.1 *In a treatment-control experiment with n_1 treated and n_0 control units, the matrix X contains* 1 *and a dummy variable for the treatment:*

$$X = \begin{pmatrix} 1_{n_1} & 1_{n_1} \\ 1_{n_0} & 0_{n_0} \end{pmatrix}.$$

We can show that

$$H = \mathrm{diag}\{ n_1^{-1} 1_{n_1} 1_{n_1}^{\mathrm{T}}, n_0^{-1} 1_{n_0} 1_{n_0}^{\mathrm{T}} \}.$$

Example 3.2 *In an experiment with n_j units receiving treatment level j $(j = 1, \ldots, J)$, the covariate matrix X contains J dummy variables for the treatment levels:*

$$X = \mathrm{diag}\{ 1_{n_1}, \ldots, 1_{n_J} \}.$$

We can show that

$$H = \mathrm{diag}\{ n_1^{-1} 1_{n_1} 1_{n_1}^{\mathrm{T}}, \ldots, n_J^{-1} 1_{n_J} 1_{n_J}^{\mathrm{T}} \}.$$

I leave the proofs of Examples 3.1 and 3.2 as Problem 3.7.

3.4 Homework problems

3.1 Univariate and multivariate OLS

Derive the univariate OLS based on the multivariate OLS formula with

$$X = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix},$$

where the x_i's are scalars.

3.2 OLS via vector and matrix calculus

Use vector and matrix calculus to prove that the OLS coefficient $\widehat{\beta}$ minimizes $(Y - Xb)^{\mathrm{T}}(Y - Xb)$.

3.3 OLS based on pseudo inverse

Prove that $\widehat{\beta} = X^{+}Y$.

Remark: Recall the definition of the pseudo inverse in Appendix A. Under Condition 3.1, we have $X^{+} = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}$.

3.4 Invariance of OLS

Theorem 3.3 below states the invariance properties of OLS. Prove Theorem 3.3.

Theorem 3.3 *Assume that $X^{\mathrm{T}}X$ is non-degenerate and Γ is a $p\times p$ non-degenerate matrix. Define $\widetilde{X} = X\Gamma$. From the OLS fit of Y on X, we obtain the coefficient $\widehat{\beta}$, the fitted value $\widehat{Y}$, and the residual $\widehat{\varepsilon}$; from the OLS fit of Y on $\widetilde{X}$, we obtain the coefficient $\widetilde{\beta}$, the fitted value $\widetilde{Y}$, and the residual $\widetilde{\varepsilon}$.*

We have

$$\widehat{\beta} = \Gamma\widetilde{\beta}, \quad \widehat{Y} = \widetilde{Y}, \quad \widehat{\varepsilon} = \widetilde{\varepsilon}.$$

Remark: From a linear algebra perspective, X and $X\Gamma$ have the same column space if Γ is a non-degenerate matrix:

$$\{Xb : b \in \mathbb{R}^p\} = \{X\Gamma c : c \in \mathbb{R}^p\}.$$

Consequently, there must be a unique projection of Y onto the common column space.

3.5 Invariance of the hat matrix

This problem extends Theorem 3.3 in Problem 3.4.

Prove that H does not change if we change X to $X\Gamma$ where $\Gamma \in \mathbb{R}^{p\times p}$ is a non-degenerate matrix.

3.6 Hat matrix with the intercept

Prove (3.5) and (3.6).

3.7 Special hat matrices

Verify the formulas of the hat matrices in Examples 3.1 and 3.2.

3.8 OLS with multiple responses

For each unit $i = 1, \ldots, n$, we have multiple responses $y_i = (y_{i1}, \ldots, y_{iq})^{\mathrm{T}} \in \mathbb{R}^q$ and multiple covariates $x_i = (x_{i1}, \ldots, x_{ip})^{\mathrm{T}} \in \mathbb{R}^p$. Define

$$Y = \begin{pmatrix} y_{11} & \cdots & y_{1q} \\ \vdots & & \vdots \\ y_{n1} & \cdots & y_{nq} \end{pmatrix} = \begin{pmatrix} y_1^{\mathrm{T}} \\ \vdots \\ y_n^{\mathrm{T}} \end{pmatrix} = (Y_1, \ldots, Y_q) \in \mathbb{R}^{n\times q}$$

and

$$X = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix} = \begin{pmatrix} x_1^{\mathrm{T}} \\ \vdots \\ x_n^{\mathrm{T}} \end{pmatrix} = (X_1, \ldots, X_p) \in \mathbb{R}^{n\times p}$$

as the response and covariate matrices, respectively. Define the multiple OLS coefficient matrix as

$$\widehat{B} = \arg\min_{B\in\mathbb{R}^{p\times q}} \sum_{i=1}^{n} \|y_i - B^{\mathrm{T}}x_i\|^2$$

Show that $\widehat{B} = (\widehat{B}_1, \ldots, \widehat{B}_q)$ has column vectors

$$\begin{aligned} \widehat{B}_1 &= (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}Y_1, \\ &\vdots \\ \widehat{B}_q &= (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}Y_q. \end{aligned}$$

Remark: This result tells us that the OLS fit with a vector outcome reduces to multiple separate OLS fits, or, the OLS fit of a matrix Y on a matrix X reduces to the column-wise OLS fits of Y on X.

3.9 Full sample and subsample OLS coefficients

Partition the full sample into K subsamples:

$$X = \begin{pmatrix} X_{(1)} \\ \vdots \\ X_{(K)} \end{pmatrix}, \quad Y = \begin{pmatrix} Y_{(1)} \\ \vdots \\ Y_{(K)} \end{pmatrix},$$

where the kth sample consists of $(X_{(k)}, Y_{(k)})$ with $X_{(k)} \in \mathbb{R}^{n_k \times p}$ and $Y_{(k)} \in \mathbb{R}^{n_k}$ being the covariate matrix and outcome vector. The sample sizes satisfy $n = \sum_{k=1}^{K} n_k$. Let $\widehat{\beta}$ be the OLS coefficient based on the full sample, and $\widehat{\beta}_{(k)}$ be the OLS coefficient based on the kth sample.

Prove that

$$\widehat{\beta} = \sum_{k=1}^{K} W_{(k)} \widehat{\beta}_{(k)},$$

where the weight matrix equals

$$W_{(k)} = (X^{\mathrm{T}}X)^{-1}X_{(k)}^{\mathrm{T}}X_{(k)}.$$

Remark: In the special case of a univariate y_i and x_i, the OLS of y_i on x_i without the intercept gives the coefficient

$$\widehat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}.$$

Partition the units into K disjoint parts: $\{1, \ldots, n\} = I_1 \cup \cdots \cup I_K$. Run OLS of y_i on x_i without the intercept using units in I_k to obtain the coefficient $\widehat{\beta}_{(k)}$. The above formula implies that

$$\widehat{\beta} = \sum_{k=1}^{K} W_{(k)} \widehat{\beta}_{(k)}$$

where

$$W_{(k)} = \frac{\sum_{i \in I_k} x_i^2}{\sum_{i=1}^{n} x_i^2}$$

is proportional to the sum of squares of the regressor x_i's in I_k.

3.10 Jacobi's theorem

Prove Theorem 3.4 below.

Theorem 3.4 (Jacobi's Theorem) *The set $\{1, \ldots, n\}$ has $\binom{n}{p}$ size-p subsets. Each subset S defines a linear equation for $b \in \mathbb{R}^p$:*

$$Y_S = X_S b$$

where $Y_S \in \mathbb{R}^p$ is the subvector of Y and $X_S \in \mathbb{R}^{p \times p}$ is the submatrix of X, corresponding to the units in S. Define the subset coefficient

$$\widehat{\beta}_S = X_S^{-1} Y_S$$

if X_S is invertible and $\widehat{\beta}_S = 0$ otherwise.

The OLS coefficient equals a weighted average of these subset coefficients:

$$\widehat{\beta} = \sum_S w_S \widehat{\beta}_S$$

where the summation is over all subsets, and the weights are

$$w_S = \frac{|\det(X_S)|^2}{\sum_{S'} |\det(X_{S'})|^2}.$$

Remark: Theorem 3.4 extends Problem 2.2. Subrahmanyam (1972) reported Theorem 3.4 although Berman (1988) attributed it to Jacobi. Wu (1986) used it in analyzing the statistical properties of OLS.

To prove Theorem 3.4, you can use Cramer's rule to express the OLS coefficient and use the Cauchy–Binet formula to expand the determinant of $X^{\mathrm{T}} X$.

A

Linear Algebra

Linear algebra is crucial for understanding the theory of the linear model. This Appendix reviews the basics of linear algebra that are closely related to this book.

All vectors are column vectors in this book. This is coherent with `R`.

A.1 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:

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

where $x = (x_1, \ldots, x_n)^{\mathrm{T}}$ and $y = (y_1, \ldots, y_n)^{\mathrm{T}}$ are two n-dimensional vectors. The length of a vector x is defined as

$$\|x\| = \sqrt{\langle x, x \rangle} = \sqrt{x^{\mathrm{T}} x}.$$

The Cauchy–Schwarz inequality states that the inner product of x and y is bounded from above by the product of their length.

Proposition A.1 *For two n-dimensional vectors x and y, we have*

$$|\langle x, y \rangle| \leq \|x\| \cdot \|y\|,$$

or, more transparently,

$$\left(\sum_{i=1}^{n} x_i y_i \right)^2 \leq \left(\sum_{i=1}^{n} x_i^2 \right) \left(\sum_{i=1}^{n} y_i^2 \right).$$

The equality holds if and only if $a y_i = b x_i$ for some a and b, for all $i = 1, \ldots, n$.

We can use the Cauchy–Schwarz inequality to prove the triangle inequality that the length of the summation of x and y is bounded from above by the summation of their length.

Proposition A.2 *For two n-dimensional vectors x and y, we have*

$$\|x + y\| \leq \|x\| + \|y\|.$$

We say that x and y are *orthogonal*, denoted by $x \perp y$, if $\langle x, y \rangle = 0$. We call a set of vectors $v_1, \ldots, v_m \in \mathbb{R}^n$ *orthonormal* if they all have unit length and are mutually orthogonal.

Geometrically, we can define the cosine of the angle between two vectors $x, y \in \mathbb{R}^n$ as

$$\cos \angle(x, y) = \frac{\langle x, y \rangle}{\|x\| \|y\|} = \frac{\sum_{i=1}^n x_i y_i}{\sqrt{\sum_{i=1}^n x_i^2 \sum_{i=1}^n y_i^2}}.$$

For unit vectors, it reduces to the inner product. When both x and y are orthogonal to 1_n, that is, $\bar{x} = n^{-1} \sum_{i=1}^n x_i = 0$ and $\bar{y} = n^{-1} \sum_{i=1}^n y_i = 0$, the formula of the cosine of the angle is identical to the sample Pearson correlation coefficient

$$\widehat{\rho}_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}.$$

Sometimes, we simply say that the cosine of the angle of two vectors measures their correlation even when they are not orthogonal to 1_n.

Column space of a matrix

Given an $n \times m$ matrix A, we can view it in terms of all elements

$$A = (a_{ij}) = \begin{pmatrix} a_{11} & \cdots & a_{1m} \\ \vdots & & \vdots \\ a_{n1} & \cdots & a_{nm} \end{pmatrix},$$

or row vectors

$$A = \begin{pmatrix} a_1^{\mathrm{T}} \\ \vdots \\ a_n^{\mathrm{T}} \end{pmatrix},$$

where $a_i \in \mathbb{R}^m$ $(i = 1, \ldots, n)$, or column vectors

$$A = (A_1, \ldots, A_m),$$

where $A_j \in \mathbb{R}^n$ $j = 1, \ldots, m$. In statistics, the rows correspond to the units, so the ith row vector is the vector observations for unit i. Moreover, viewing A in terms of its column vectors can give more insights.

Definition A.1 (column space) *For an $n \times m$ matrix $A = (A_1, \ldots, A_m)$, define the column space of A as*

$$\mathcal{C}(A) = \{\alpha_1 A_1 + \cdots + \alpha_m A_m : \alpha_1, \ldots, \alpha_m \in \mathbb{R}\}.$$

The column space of A is the set of all linear combinations of the column vectors $A_1, \ldots, A_m$. The column space is important because we can write $A\alpha$, with $\alpha = (\alpha_1, \ldots, \alpha_m)^{\mathrm{T}}$, as

$$A\alpha = (A_1, \ldots, A_m) \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_m \end{pmatrix} = \alpha_1 A_1 + \cdots + \alpha_m A_m,$$

which is in $\mathcal{C}(A)$.

We define the row space of A as the column space of A^{T}.

A.1 Matrix product

Given an $n \times m$ matrix $A = (a_{ij})$ and an $m \times r$ matrix $B = (b_{ij})$, we can define their product as $C = AB$ where the $n \times r$ matrix $C = (c_{ij})$ has the (i, j)th element

$$c_{ij} = \sum_{k=1}^{m} a_{ik} b_{kj}.$$

In terms of the row vectors of A or column vectors of B, we have

$$c_{ij} = a_i^{\mathrm{T}} B_j,$$

that is, c_{ij} equals the inner product of the ith row vector of A and the jth column vector of B. Moreover, the matrix product satisfies

$$AB = A(B_1, \ldots, B_r) = (AB_1, \ldots, AB_r) \tag{A.1}$$

so the column vectors of AB belongs to the column space of A; it also satisfies

$$AB = \begin{pmatrix} a_1^{\mathrm{T}} \\ \vdots \\ a_n^{\mathrm{T}} \end{pmatrix} B = \begin{pmatrix} a_1^{\mathrm{T}} B \\ \vdots \\ a_n^{\mathrm{T}} B \end{pmatrix} \tag{A.2}$$

so the row vectors of AB belong to the column space of B^{T}, or equivalently, the row space of B.

A.2 Linearly independent vectors and rank

We call a set of vectors $A_1, \ldots, A_m \in \mathbb{R}^n$ *linearly independent* if

$$x_1 A_1 + \cdots + x_m A_m = 0$$

must imply $x_1 = \cdots = x_m = 0$. We call $A_{j_1}, \ldots, A_{j_k}$ maximally linearly independent if adding another vector makes them linearly dependent. Define k as the rank of $\{A_1, \ldots, A_m\}$ and also define k as the rank of the matrix $A = (A_1, \ldots, A_m)$.

A set of vectors may have different subsets of vectors that are maximally linearly independent. But the rank k is unique. We can also define the rank of a matrix in terms of its row vectors. A remarkable theorem in linear algebra is that it does not matter whether we define the rank of a matrix in terms of its column vectors or row vectors.

From the matrix product formulas (A.1) and (A.2), we have the following result.

Proposition A.3 $\text{rank}(AB) \leq \min\{\text{rank}(A), \text{rank}(B)\}$.

The rank decomposition of a matrix decomposes A into the product of two matrices of full ranks.

Proposition A.4 *If an $n \times m$ matrix A has rank k, then $A = BC$ for some $n \times k$ matrix B and $k \times m$ matrix C.*

Proof of Proposition A.4: Let $A_{j_1}, \ldots, A_{j_k}$ be the maximally linearly independent column vectors of A. Stack them into an $n \times k$ matrix $B = (A_{j_1}, \ldots, A_{j_k})$. They can linearly represent all column vectors of A:

$$\begin{aligned} A &= (c_{11} A_{j_1} + \cdots + c_{k1} A_{j_k}, \ldots, c_{1m} A_{j_1} + \cdots + c_{km} A_{j_k}) \\ &= (BC_1, \ldots, BC_m) \\ &= BC, \end{aligned}$$

where $C = (C_1, \ldots, C_m)$ is an $k \times m$ matrix with column vectors

$$C_1 = \begin{pmatrix} c_{11} \\ \vdots \\ c_{k1} \end{pmatrix}, \cdots C_m = \begin{pmatrix} c_{1m} \\ \vdots \\ c_{km} \end{pmatrix}.$$

□

Proposition A.3 ensures that the B and C in Proposition A.4 must satisfy $\text{rank}(B) \geq k$ and $\text{rank}(C) \geq k$, so they must both have rank k. The decomposition in Proposition A.4 is not unique since the choice of the maximally linearly independent column vectors of A is not unique.

Some special matrices

An $n \times n$ matrix A is symmetric if $A^{\mathrm{T}} = A$.

An $n \times n$ diagonal matrix A has zero off-diagonal elements, denoted by $A = \text{diag}\{a_{11}, \ldots, a_{nn}\}$. Diagonal matrices are symmetric.

An $n \times n$ matrix is orthogonal if $A^{\mathrm{T}} A = AA^{\mathrm{T}} = I_n$. The column vectors of an orthogonal matrix are orthonormal; so are its row vectors. If A is orthogonal, then

$$\|Ax\| = \|x\|$$

for any vector $x \in \mathbb{R}^n$. That is, multiplying a vector by an orthogonal matrix does not change the length of the vector. Geometrically, an orthogonal matrix corresponds to rotation.

An $n \times n$ matrix A is upper triangular if $a_{ij} = 0$ for $i > j$ and lower triangular if $a_{ij} = 0$ for $i < j$.

Determinant

The original definition of the determinant of a matrix $A = (a_{ij}) \in \mathbb{R}^{n \times n}$, due to Leibniz, is quite complicated, which relies on the notation of permutation. A permutation σ on $\{1, \ldots, n\}$ is a one-to-one mapping from $\{1, \ldots, n\}$ to $\{1, \ldots, n\}$. Let $\text{sgn}(\sigma)$ denote the sign of the permutation σ, which equals 1 if σ can be obtained via an even number of transpositions and 0 if σ can be obtained via an odd number of transpositions. Define

$$\det(A) = \sum_{\sigma} \text{sgn}(\sigma) \prod_{i=1}^{n} a_{i,\sigma(i)},$$

where the summation is over all possible permutations, and $a_{i,\sigma(i)}$ is the $(i, \sigma(i))$-th element of A.

The determinant of a 2×2 matrix has a simple form:

$$\det \begin{pmatrix} a & b \\ c & d \end{pmatrix} = ad - bc. \tag{A.3}$$

The determinant of the Vandermonde matrix has the following formula:

$$\det \begin{pmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^{n-1} \\ 1 & x_2 & x_2^2 & \cdots & x_2^{n-1} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_n & x_n^2 & \cdots & x_n^{n-1} \end{pmatrix} = \prod_{1 \leq i,j \leq n} (x_j - x_i). \tag{A.4}$$

This book will not use the above definition of the determinant. The properties of the determinant are more useful. I will review two.

Proposition A.5 *For two square matrices A and B, we have*

$$\det(AB) = \det(A)\det(B) = \det(BA).$$

Proposition A.6 *For two square matrices $A \in \mathbb{R}^{m\times m}$ and $B \in \mathbb{R}^{n\times n}$, we have*

$$\det\begin{pmatrix} A & 0 \\ C & B \end{pmatrix} = \det\begin{pmatrix} A & D \\ 0 & B \end{pmatrix} = \det(A)\det(B).$$

Inverse of a matrix

Let I_n be the $n \times n$ identity matrix. An $n \times n$ matrix A is invertible or 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}. If A is an orthogonal matrix, then $A^{\mathrm{T}} = A^{-1}$.

A square matrix is invertible if and only if $\det(A) \neq 0$.

The inverse of a 2×2 matrix is

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}^{-1} = \frac{1}{ad - bc}\begin{pmatrix} d & -b \\ -c & a \end{pmatrix}. \tag{A.5}$$

The inverse of a 3×3 lower triangular matrix is

$$\begin{pmatrix} a & 0 & 0 \\ b & c & 0 \\ d & e & f \end{pmatrix}^{-1} = \frac{1}{acf}\begin{pmatrix} cf & 0 & 0 \\ -bf & af & 0 \\ be - cd & -ae & ac \end{pmatrix}. \tag{A.6}$$

A useful identity is

$$(AB)^{-1} = B^{-1}A^{-1},$$

if both A and B are invertible.

Eigenvalues and eigenvectors

For an $n \times n$ matrix A, if there exists a pair of n-dimensional, non-zero vector x and a scalar λ such that

$$Ax = \lambda x,$$

then we call λ an eigenvalue and x the associated eigenvector of A. From the definition, eigenvalue and eigenvector always come in pairs. The following eigen-decomposition theorem is fundamental for real symmetric matrices.

Theorem A.1 *If A is an $n \times n$ symmetric matrix, then there exists an orthogonal matrix P such that*

$$P^{\mathrm{T}}AP = \mathrm{diag}\{\lambda_1, \dots, \lambda_n\}, \tag{A.7}$$

where the λ's are the n eigenvalues of A, and the column vectors of $P = (\gamma_1, \cdots, \gamma_n)$ are the corresponding eigenvectors.

If we multiply (A.7) by P from the left, then we can write the eigendecomposition as

$$AP = P\mathrm{diag}\{\lambda_1, \dots, \lambda_n\}$$

or, equivalently,

$$A(\gamma_1, \cdots, \gamma_n) = (\lambda_1\gamma_1, \cdots, \lambda_n\gamma_n),$$

then (λ_i, γ_i) must be a pair of eigenvalue and eigenvector. Moreover, the eigendecomposition in Theorem A.1 is unique up to the permutation of the columns of P and the corresponding λ_i's.

Corollary A.1 *If a real symmetric matrix A has eigen-decomposition $P^{\mathrm{T}}AP = \operatorname{diag}\{\lambda_1, \ldots, \lambda_n\}$, then*

$$A = P\operatorname{diag}\{\lambda_1, \ldots, \lambda_n\}P^{\mathrm{T}},$$

and therefore,

$$A^k = AA\cdots A = P\operatorname{diag}\{\lambda_1^k, \ldots, \lambda_n^k\}P^{\mathrm{T}}.$$

If the eigenvalues of A are nonzero, then

$$A^{-1} = P\operatorname{diag}\{1/\lambda_1, \ldots, 1/\lambda_n\}P^{\mathrm{T}}.$$

The eigen-decomposition is also useful for defining the square root of an $n \times n$ symmetric matrix. In particular, if the eigenvalues of A are nonnegative, then we can define

$$A^{1/2} = P\operatorname{diag}\{\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_n}\}P^{\mathrm{T}}.$$

By definition, $A^{1/2}$ is a symmetric matrix satisfying $A^{1/2}A^{1/2} = A$. There are other definitions of the square root of a symmetric matrix, but we adopt this form in this book.

From (A.7), we can write A as

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

For an $n \times n$ symmetric matrix A, 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, we can define the quadratic form as

$$x^{\mathrm{T}}Ax = \langle x, Ax \rangle = \sum_{i=1}^n \sum_{j=1}^n a_{ij}x_i x_j.$$

We always consider a symmetric matrix in the quadratic form without loss of generality. Otherwise, we can symmetrize A as $\widetilde{A} = (A + A^{\mathrm{T}})/2$ without changing the value of the quadratic form because

$$x^{\mathrm{T}}Ax = x^{\mathrm{T}}\widetilde{A}x.$$

We call A positive semi-definite, denoted by $A \succeq 0$, if $x^{\mathrm{T}}Ax \geq 0$ for all x; we call A positive definite, denoted by $A \succ 0$, if $x^{\mathrm{T}}Ax > 0$ for all nonzero x.

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 variances or covariance matrices. Given two unbiased estimators $\widehat{\theta}_1$ and $\widehat{\theta}_2$ for a scalar parameter θ, we say that $\widehat{\theta}_1$ is at least as efficient as $\widehat{\theta}_2$ if $\operatorname{var}(\widehat{\theta}_2) \geq \operatorname{var}(\widehat{\theta}_1)$. In the vector case, we say that $\widehat{\theta}_1$ is at least as efficient as $\widehat{\theta}_2$ if $\text{cov}(\widehat{\theta}_2) \succeq \text{cov}(\widehat{\theta}_1)$, which is equivalent to $\text{var}(\ell^{\mathrm{T}}\widehat{\theta}_2) \geq \text{var}(\ell^{\mathrm{T}}\widehat{\theta}_1)$ for any linear combination of the estimators.

The eigenvalues of a symmetric matrix determine whether it is positive semi-definite or positive definite.

Theorem A.2 *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.*

An important result is the relationship between the eigenvalues and the extreme values of the quadratic form. Assume that the eigenvalues are rearranged in decreasing order such that $\lambda_1 \geq \cdots \geq \lambda_n$. For a unit vector x with length $\|x\| = 1$, we have that

$$x^{\mathrm{T}}Ax = x^{\mathrm{T}}\sum_{i=1}^{n}\lambda_i\gamma_i\gamma_i^{\mathrm{T}}x = \sum_{i=1}^{n}\lambda_i\alpha_i^2$$

where

$$\alpha = \begin{pmatrix}\alpha_1\\ \vdots\\ \alpha_n\end{pmatrix} = \begin{pmatrix}\gamma_1^{\mathrm{T}}x\\ \vdots\\ \gamma_n^{\mathrm{T}}x\end{pmatrix} = P^{\mathrm{T}}x$$

has length $\|\alpha\|^2 = \|x\|^2 = 1$. Then the maximum value of $x^{\mathrm{T}}Ax$ is λ_1, which is achieved at $\alpha_1 = 1$ and $\alpha_2 = \cdots = \alpha_n = 0$ (for example, if $x = \gamma_1$, then $\alpha_1 = 1$ and $\alpha_2 = \cdots = \alpha_n = 0$). For a unit vector x that is orthogonal to γ_1, we have that

$$x^{\mathrm{T}}Ax = \sum_{i=2}^{n}\lambda_i\alpha_i^2$$

where $\alpha = P^{\mathrm{T}}x$ has unit length with $\alpha_1 = 0$. The maximum value of $x^{\mathrm{T}}Ax$ is λ_2, which is achieved at $\alpha_2 = 1$ and $\alpha_1 = \alpha_3 = \cdots = \alpha_n = 0$, for example, $x = \gamma_2$. By induction, we have the following theorem.

Theorem A.3 *Suppose that an $n \times n$ symmetric matrix has eigen-decomposition $\sum_{i=1}^{n}\lambda_i\gamma_i\gamma_i^{\mathrm{T}}$ where $\lambda_1 \geq \cdots \geq \lambda_n$.*

1. *The optimization problem*

$$\max_{x\in\mathbb{R}^n} x^{\mathrm{T}}Ax \textit{ such that } \|x\| = 1$$

has maximum λ_1, which can be achieved by γ_1.

2. *The optimization problem*

$$\max_{x\in\mathbb{R}^n} x^{\mathrm{T}}Ax \textit{ such that } \|x\| = 1, x \perp \gamma_1$$

has maximum λ_2, which can be achieved by γ_2.

3. *The optimization problem*

$$\max_{x\in\mathbb{R}^n} x^{\mathrm{T}}Ax \textit{ such that } \|x\| = 1, x \perp \gamma_1, \ldots, x \perp \gamma_k$$

has maximum λ_{k+1}, which can be achieved by γ_{k+1} $(k = 1, \ldots, n-1)$.

Theorem A.3 implies the following theorem on the Rayleigh quotient

$$r(x) = x^{\mathrm{T}}Ax/x^{\mathrm{T}}x \qquad (x \in \mathbb{R}^n).$$

Theorem A.4 *(Rayleigh quotient and eigenvalues) The maximum and minimum eigenvalues of an $n \times n$ symmetric matrix A equals*

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

with the maximizer and minimizer being the eigenvectors corresponding to the maximum and minimum eigenvalues, respectively.

An immediate consequence of Theorem A.4 is that the diagonal elements of A are bounded by the smallest and largest eigenvalues of A. This follows by taking $x = (0, \dots, 1, \dots, 0)^{\mathrm{T}}$, where only the ith element equals 1.

Trace

The trace of an $n \times n$ 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 A.7 $\mathrm{trace}(AB) = \mathrm{trace}(BA)$ *as long as AB and BA are both square matrices.*

We can verify Proposition A.7 by definition. It states that AB and BA have the same trace although AB differs from BA in general. In fact, it is particularly useful if the dimension of BA is much lower than the dimension of AB. For example, if both $A = (a_1, \dots, a_n)^{\mathrm{T}}$ and $B = (b_1, \dots, b_n)$ are vectors, then $\mathrm{trace}(AB) = \mathrm{trace}(BA) = \langle B^{\mathrm{T}}, A \rangle = \sum_{i=1}^{n} a_i b_i$.

Proposition A.8 *The trace of an $n \times n$ symmetric matrix A equals the sum of its eigenvalues:* $\mathrm{trace}(A) = \sum_{i=1}^{n} \lambda_i$.

Proof of Proposition A.8: It follows from the eigen-decomposition and Proposition A.7. Let $\Lambda = \mathrm{diag}\{\lambda_1, \dots, \lambda_n\}$. Then we have

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

□

Projection matrix

An $n \times n$ matrix H is a projection matrix, if it is symmetric and $H^2 = H$. The eigenvalues of H must be either 1 or 0. To see this, we assume that $Hx = \lambda x$ for some nonzero vector x, and use two ways to calculate $H^2 x$:

$$\begin{aligned} H^2 x &= Hx = \lambda x, \\ H^2 x &= H(Hx) = H(\lambda x) = \lambda Hx = \lambda^2 x. \end{aligned}$$

So $(\lambda - \lambda^2)x = 0$ which implies that $\lambda - \lambda^2 = 0$, i.e., $\lambda = 0$ or 1. So the trace of H equals its rank:

$$\text{trace}(H) = \text{rank}(H).$$

Why is this a reasonable definition of a "projection matrix"? Or, why must a projection matrix satisfy $H^{\mathrm{T}} = H$ and $H^2 = H$? First, it is reasonable to require that $Hx_1 = x_1$ for any $x_1 \in \mathcal{C}(H)$, the column space of H. Since $x_1 = H\alpha$ for some α, we indeed have $Hx_1 = H(H\alpha) = H^2\alpha = H\alpha = x_1$ because of the property $H^2 = H$. Second, it is reasonable to require that $x_1 \perp x_2$ for any vector $x_1 = H\alpha \in \mathcal{C}(H)$ and x_2 such that $Hx_2 = 0$. So we need $\alpha^{\mathrm{T}} H^{\mathrm{T}} x_2 = 0$ which is true if $H = H^{\mathrm{T}}$. Therefore, the two conditions are natural for the definition of a projection matrix.

More interestingly, a project matrix has a more explicit form as stated below.

Theorem A.5 *If an $n \times p$ matrix X has p linearly independent columns, then $H = X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}$ is a projection matrix. Conversely, if an $n \times n$ matrix H is a projection matrix with rank p, then $H = X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}$ for some $n \times p$ matrix X with linearly independent columns.*

It is relatively easy to verify the first part of Theorem A.5; see Chapter 3. The second part of Theorem A.5 follows from the eigen-decomposition of H, with the first p eigen-vectors being the column vectors of X.

Cholesky decomposition

An $n \times n$ positive semi-definite matrix A can be decomposed as $A = LL^{\mathrm{T}}$ where L is an $n \times n$ lower triangular matrix with non-negative diagonal elements. If A is positive definite, the decomposition is unique. In general, it is not. Take an arbitrary orthogonal matrix Q, we have $A = LQQ^{\mathrm{T}}L^{\mathrm{T}} = CC^{\mathrm{T}}$ where $C = LQ$. So we can decompose a positive semi-definite matrix A as $A = CC^{\mathrm{T}}$, but this decomposition is not unique.

Singular value decomposition (SVD)

Any $n \times m$ matrix A can be decomposed as

$$A = UDV^{\mathrm{T}}$$

where U is $n \times n$ orthogonal matrix, V is $m \times m$ orthogonal matrix, and D is $n \times m$ matrix with all zeros for the non-diagonal elements. For a tall matrix with $n \geq m$, the diagonal matrix D has many zeros, so we can also write

$$A = UDV^{\mathrm{T}}$$

where U is $n \times m$ matrix with orthonormal columns ($U^{\mathrm{T}}U = I_m$), V is $m \times m$ orthogonal matrix, and D is $m \times m$ diagonal matrix. Similarly, for a wide matrix with $n \leq m$, we can write

$$A = UDV^{\mathrm{T}}$$

where U is $n \times n$ orthogonal matrix, V is $m \times n$ matrix with orthonormal columns ($V^{\mathrm{T}}V = I_n$), and D is $n \times n$ diagonal matrix.

If D has only $r \leq \min(m, n)$ nonzero elements, then we can further simplify the decomposition as

$$A = UDV^{\mathrm{T}}$$

where U is $n \times r$ matrix with orthonormal columns ($U^{\mathrm{T}}U = I_r$), V is $m \times r$ matrix with orthonormal columns ($V^{\mathrm{T}}V = I_r$), and D is $r \times r$ diagonal matrix. With more explicit forms of

$$U = (U_1, \dots, U_r), \quad D = \mathrm{diag}(d_1, \dots, d_r), \quad V = (V_1, \dots, V_r),$$

we can write A as

$$A = (U_1, \dots, U_r) \begin{pmatrix} d_1 & & \\ & \ddots & \\ & & d_r \end{pmatrix} \begin{pmatrix} V_1^{\mathrm{T}} \\ \vdots \\ V_r^{\mathrm{T}} \end{pmatrix} = \sum_{k=1}^{r} d_k U_k V_k^{\mathrm{T}}.$$

The SVD implies that

$$AA^{\mathrm{T}} = UDD^{\mathrm{T}}U^{\mathrm{T}}, \quad A^{\mathrm{T}}A = VD^{\mathrm{T}}DV^{\mathrm{T}},$$

which are the eigen decompositions of AA^{T} and $A^{\mathrm{T}}A$. This ensures that AA^{T} and $A^{\mathrm{T}}A$ have the same non-zero eigenvalues.

An application of the SVD is to define the pseudoinverse of any matrix. Define D^{+} as the pseudoinverse of D with the non-zero elements inverted but the zero elements intact at zero. Define

$$A^{+} = VD^{+}U^{\mathrm{T}} = \sum_{k=1}^{r} d_k^{-1} V_k U_k^{\mathrm{T}}$$

as the pseudoinverse of A. The definition holds even if A is not a square matrix. We can verify that

$$AA^{+}A = A, \quad A^{+}AA^{+} = A^{+}.$$

If A is a square nondegenerate matrix, then $A^{+} = A^{-1}$ equals the standard definition of the inverse. In the special case with a symmetric A, its SVD is identical to its eigen decomposition $A = P\mathrm{diag}(\lambda_1, \dots, \lambda_n)P^{\mathrm{T}}$. If $A = P\mathrm{diag}(\lambda_1, \dots, \lambda_k, 0, \dots, 0)P^{\mathrm{T}}$ is not invertible, its pseudoinverse equals

$$A^{+} = P\mathrm{diag}(\lambda_1^{-1}, \dots, \lambda_k^{-1}, 0, \dots, 0)P^{\mathrm{T}}$$

if $\mathrm{rank}(A) = k < n$ and $\lambda_1, \lambda_1, \dots, \lambda_k$ are the nonzero eigen-values.

Another application of the SVD is the *polar decomposition* for any square matrix A. Since $A = UDV^{\mathrm{T}} = UDU^{\mathrm{T}}UV^{\mathrm{T}}$ with orthogonal U and V, we have

$$A = (AA^{\mathrm{T}})^{1/2}\Gamma, \tag{A.8}$$

where $(AA^{\mathrm{T}})^{1/2} = UDU^{\mathrm{T}}$ and $\Gamma = UV^{\mathrm{T}}$ is an orthogonal matrix.

A.2 Vector calculus

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

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

for the component-wise partial derivative, which must have the same dimension as x. It is often called the *gradient* of f. For example, for a linear function $f(x) = x^{\mathrm{T}} a = a^{\mathrm{T}} x$ with $a, x \in \mathbb{R}^p$, we have

$$\frac{\partial a^{\mathrm{T}} x}{\partial x} = \begin{pmatrix} \frac{\partial a^{\mathrm{T}} x}{\partial x_1} \\ \vdots \\ \frac{\partial a^{\mathrm{T}} x}{\partial x_p} \end{pmatrix} = \begin{pmatrix} \frac{\partial \sum_{j=1}^p a_j x_j}{\partial x_1} \\ \vdots \\ \frac{\partial \sum_{j=1}^p a_j x_j}{\partial x_p} \end{pmatrix} = \begin{pmatrix} a_1 \\ \vdots \\ a_p \end{pmatrix} = a; \tag{A.9}$$

for a quadratic function $f(x) = x^{\mathrm{T}} A x$ with a symmetric $A \in \mathbb{R}^{p \times p}$ and $x \in \mathbb{R}^p$, we have

$$\frac{\partial x^{\mathrm{T}} A x}{\partial x} = \begin{pmatrix} \frac{\partial x^{\mathrm{T}} A x}{\partial x_1} \\ \vdots \\ \frac{\partial x^{\mathrm{T}} A x}{\partial x_p} \end{pmatrix} = \begin{pmatrix} \frac{\partial \sum_{i=1}^p \sum_{j=1}^p a_{ij} x_i x_j}{\partial x_1} \\ \vdots \\ \frac{\partial \sum_{i=1}^p \sum_{j=1}^p a_{ij} x_i x_j}{\partial x_p} \end{pmatrix} = \begin{pmatrix} 2a_{11} x_1 + \cdots + 2a_{1p} x_p \\ \vdots \\ 2a_{p1} x_1 + \cdots + 2a_{pp} x_p \end{pmatrix} = 2Ax.$$

These are two important rules of vector calculus used in this book, summarized below.

Proposition A.9 *We have*

$$\begin{aligned} \frac{\partial a^{\mathrm{T}} x}{\partial x} &= a, \\ \frac{\partial x^{\mathrm{T}} A x}{\partial x} &= 2Ax. \end{aligned}$$

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

$$\frac{\partial f(x)}{\partial x} \equiv \left(\frac{\partial f_1(x)}{\partial x}, \cdots, \frac{\partial f_q(x)}{\partial x} \right) = \begin{pmatrix} \frac{\partial f_1(x)}{\partial x_1} & \cdots & \frac{\partial f_q(x)}{\partial x_1} \\ \vdots & & \vdots \\ \frac{\partial f_1(x)}{\partial x_p} & \cdots & \frac{\partial f_q(x)}{\partial x_p} \end{pmatrix}, \tag{A.10}$$

which is a $p \times q$ matrix with rows corresponding to the elements of x and the columns corresponding to the elements of $f(x)$. We can easily extend the first result of Proposition A.9.

Proposition A.10 *For $B \in \mathbb{R}^{p \times q}$ and $x \in \mathbb{R}^p$, we have*

$$\frac{\partial B^{\mathrm{T}} x}{\partial x} = B.$$

Proof of Proposition A.10: Partition $B = (B_1, \ldots, B_q)$ in terms of its column vectors. The jth element of $B^{\mathrm{T}} x$ is $B_j^{\mathrm{T}} x$ so the j-th column of $\partial B^{\mathrm{T}} x / \partial x$ is B_j based on Proposition A.9. This verifies that $\partial B^{\mathrm{T}} x / \partial x$ equals B. □

Some authors define $\partial f(x)/\partial x$ as the transpose of (A.10). I adopt this form for its natural connection with (A.9) when $q = 1$. Sometimes, it is indeed more convenient to work with the transpose of $\partial f(x)/\partial x$. Then I will use the notation

$$\frac{\partial f(x)}{\partial x^{\mathrm{T}}} = \left(\frac{\partial f(x)}{\partial x} \right)^{\mathrm{T}} = \left(\frac{\partial f(x)}{\partial x_1}, \cdots, \frac{\partial f(x)}{\partial x_p} \right),$$

which puts the transpose notation on x.

The above formulas become more powerful in conjunction with the chain rule. For example, for any differentiable function $h(z)$ mapping from $\mathbb{R}$ to $\mathbb{R}$ with derivative $h'(z)$, we have

$$\begin{aligned} \frac{\partial h(a^{\mathrm{T}} x)}{\partial x} &= h'(a^{\mathrm{T}} x) a, \\ \frac{\partial h(x^{\mathrm{T}} A x)}{\partial x} &= 2h'(x^{\mathrm{T}} A x) A x. \end{aligned}$$

For any differentiable function $h(z)$ mapping from $\mathbb{R}^q$ to $\mathbb{R}$ with gradient $\partial h(z)/\partial z$, we have

$$\begin{aligned} \frac{\partial h(B^{\mathrm{T}} x)}{\partial x} &= \frac{\partial h(B_1^{\mathrm{T}} x, \ldots, B_q^{\mathrm{T}} x)}{\partial x} \\ &= \sum_{j=1}^{q} \frac{\partial h(B_1^{\mathrm{T}} x, \ldots, B_q^{\mathrm{T}} x)}{\partial z_j} B_j \\ &= B \frac{\partial h(B^{\mathrm{T}} x)}{\partial z}. \end{aligned}$$

Moreover, we can also define the Hessian matrix of a function $f(x)$ mapping from $\mathbb{R}^p$ to $\mathbb{R}$:

$$\frac{\partial^2 f(x)}{\partial x \partial x^{\mathrm{T}}} = \left(\frac{\partial^2 f(x)}{\partial x_i \partial x_j} \right)_{1 \le i,j \le p} = \frac{\partial}{\partial x^{\mathrm{T}}} \left(\frac{\partial f(x)}{\partial x} \right).$$

A.3 Homework problems

A.1 Triangle inequality of the inner product

With three unit vectors $u, v, w \in \mathbb{R}^n$, prove that

$$\sqrt{1 - \langle u, w \rangle} \le \sqrt{1 - \langle u, v \rangle} + \sqrt{1 - \langle v, w \rangle}.$$

Remark: The result is a direct consequence of the standard triangle inequality but it has an interesting implication. If $\langle u, v \rangle \ge 1 - \epsilon$ and $\langle v, w \rangle \ge 1 - \epsilon$, then $\langle u, w \rangle \ge 1 - 4\epsilon$. This implied inequality is mostly interesting when ϵ is small. It states that when u and v are highly correlated and v and w are highly correlated, then u and w must also be highly correlated. Note that we can find counterexamples for the following relationship:

$$\langle u, v \rangle > 0, \quad \langle v, w \rangle > 0 \quad \text{but} \quad \langle u, w \rangle = 0.$$

A.2 Van der Corput inequality

Assume that $v, u_1, \ldots, u_m \in \mathbb{R}^n$ have unit length. Prove that

$$\left(\sum_{i=1}^{m} \langle v, u_i \rangle \right)^2 \le \sum_{i=1}^{m} \sum_{j=1}^{m} \langle u_i, u_j \rangle.$$

Remark: This result is not too difficult to prove, but it says something fundamentally interesting. If v is correlated with many vectors $u_1, \ldots, u_m$, then at least some vectors in $u_1, \ldots, u_m$ must be correlated.

A.3 Inverse of a block matrix

Prove that

$$\begin{aligned}
&\begin{pmatrix} A & B \\ C & D \end{pmatrix}^{-1} \\
&= \begin{pmatrix} A^{-1} + A^{-1}B(D - CA^{-1}B)^{-1}CA^{-1} & -A^{-1}B(D - CA^{-1}B)^{-1} \\ -(D - CA^{-1}B)^{-1}CA^{-1} & (D - CA^{-1}B)^{-1} \end{pmatrix} \\
&= \begin{pmatrix} (A - BD^{-1}C)^{-1} & -(A - BD^{-1}C)^{-1}BD^{-1} \\ -D^{-1}C(A - BD^{-1}C)^{-1} & D^{-1} + D^{-1}C(A - BD^{-1}C)^{-1}BD^{-1} \end{pmatrix},
\end{aligned}$$

provided that all the inverses of the matrices exist. The two forms of the inverse imply the Woodbury formula:

$$(A - BD^{-1}C)^{-1} = A^{-1} + A^{-1}B(D - CA^{-1}B)^{-1}CA^{-1},$$

which further implies the Sherman–Morrison formula:

$$(A + uv^{\mathrm{T}})^{-1} = A^{-1} - (1 + v^{\mathrm{T}}A^{-1}u)^{-1}A^{-1}uv^{\mathrm{T}}A^{-1},$$

where A is an invertible square matrix, and u and v are two column vectors.

A.4 Matrix determinant lemma

Prove that given the identity matrix I_n and two n-vectors u and v, we have

$$\det(I_n + uv^{\mathrm{T}}) = 1 + v^{\mathrm{T}}u.$$

Further prove that if I_n is replaced by an $n \times n$ invertible matrix A, we have

$$\det(A + uv^{\mathrm{T}}) = (1 + v^{\mathrm{T}}A^{-1}u) \cdot \det(A).$$

A.5 Symmetric rank one update of the identity matrix

Given a real number c and a vector $x \in \mathbb{R}^n$, consider the matrix $I_n + cxx^{\mathrm{T}}$, which is a symmetric $n \times n$ matrix. Assume $1 + c\|x\|^2 > 0$.

1. Find the eigenvalues and eigenvectors of $I_n + cxx^{\mathrm{T}}$.
2. Use the eigendecomposition to prove that

   $$\det(I_n + cxx^{\mathrm{T}}) = 1 + c\|x\|^2.$$

   Remark: You can compare this result with Problem A.4.
3. Use the eigendecomposition to prove that

   $$(I_n + cxx^{\mathrm{T}})^{-1} = I_n - \frac{c}{1 + c\|x\|^2}xx^{\mathrm{T}}.$$

   Remark: You can compare this result with the Sherman–Morrison formula in Problem A.3.
4. Use the eigendecomposition to prove that

   $$(I_n + cxx^{\mathrm{T}})^{1/2} = I_n + \frac{c}{1 + \sqrt{1 + c\|x\|^2}}xx^{\mathrm{T}}.$$
5. Use the eigendecomposition to prove that

   $$(I_n + cxx^{\mathrm{T}})^{-1/2} = I_n - \frac{c}{\sqrt{1 + c\|x\|^2}(1 + \sqrt{1 + c\|x\|^2})}xx^{\mathrm{T}}.$$

A.6 Rank one update and positive definiteness

Assume A is an $n \times n$ positive definite matrix. Assume b is an n-dimensional vector.

Prove that $A - bb^{\mathrm{T}}$ is positive definite if and only if $b^{\mathrm{T}} A^{-1} b < 1$, and $A - bb^{\mathrm{T}}$ is positive semi-definite if and only if $b^{\mathrm{T}} A^{-1} b \leq 1$.

Remark: Farebrother (1976, Appendix) gave this result. It is not directly used in this book but is related to the leave-one-out formula in Theorem **??**.

A.7 Positive definiteness of the difference of inverses

With scalars $a \geq b > 0$, we know that $a^{-1} \leq b^{-1}$. A similar result hold for matrices.

Assume A and B are positive definite matrices. First, prove that if $A - I$ is positive definite, then $I - A^{-1}$ is positive definite; if $A - I$ is positive semi-definite, then $I - A^{-1}$ is positive semi-definite. Second, prove that if $A - B$ is positive definite, then $B^{-1} - A^{-1}$ is positive definite; if $A - B$ is positive semi-definite, then $B^{-1} - A^{-1}$ is positive semi-definite.

A.8 Decomposition of a positive semi-definite matrix

Prove that if A is positive semi-definite, then there exists a matrix C such that $A = CC^{\mathrm{T}}$.

A.9 Trace of the product of two matrices

Prove that if A and B are two $n \times n$ positive semi-definite matrices, then $\mathrm{trace}(AB) \geq 0$.

Remark: Use the eigen-decomposition of $A = \sum_{i=1}^{n} \lambda_i \gamma_i \gamma_i^{\mathrm{T}}$ to prove the result.

In fact, a stronger result holds. If two $n \times n$ symmetric matrices A and B have eigenvalues

$$\lambda_1 \geq \cdots \geq \lambda_n, \quad \mu_1 \geq \cdots \geq \mu_n$$

respectively, then

$$\sum_{i=1}^{n} \lambda_i \mu_{n+1-i} \leq \mathrm{trace}(AB) \leq \sum_{i=1}^{n} \lambda_i \mu_i.$$

The result is due to Von Neumann (1937) and Ruhe (1970). See also Chen and Li (2019, Lemma 4.12).

A.10 Trace of the product of two matrices and positive semi-definiteness

This problem gives the other direction of Problem A.9.

Assume A is a symmetric matrix. Prove that A is positive semi-definite if and only if $\mathrm{trace}(AB) \geq 0$ for all positive semi-definite matrices B.

Remark: One direction of this statement is in Problem A.9. We only need to prove the other direction. Theobald (1974) used it to analyze ridge regression.

A.11 Vector calculus

What is the formula for $\partial x^{\mathrm{T}} A x / \partial x$ if A is not symmetric in Proposition A.9?

B

Random Variables

This Appendix reviews the basics of random variables. Let "IID" denote "independent and identically distributed", "$\overset{\text{IID}}{\sim}$" denote a sequence of random variables that are IID with some common distribution, and "$\perp\!\!\!\perp$" denote independence between random variables.

Define Euler's Gamma function as

$$\Gamma(z) = \int_0^\infty x^{z-1} e^{-x} \mathrm{d}x, \qquad (z > 0),$$

which is a natural extension of the factorial since $\Gamma(n) = (n-1)!$. Further, define the digamma function as $\psi(z) = \mathrm{d}\log\Gamma(z)/\mathrm{d}z$ and the trigamma function $\psi'(z)$ as the derivative of $\psi(z)$. In R, we can use

```
gamma(z)
lgamma(z)
digamma(z)
trigamma(z)
```

to compute $\Gamma(z)$, $\log\Gamma(z)$, $\psi(z)$, and $\psi'(z)$.

B.1 Some important univariate random variables

B.1.1 Normal, chi-squared, t and F

The standard Normal random variable $Z \sim \mathrm{N}(0,1)$ has density

$$f(z) = (2\pi)^{-1/2} \exp\left(-z^2/2\right).$$

A Normal random variable X has mean μ and variance σ^2, denoted by $\mathrm{N}(\mu, \sigma^2)$, if $X = \mu + \sigma Z$. We can show that X has density

$$f(x) = (2\pi)^{1/2} \exp\left\{-(x-\mu)^2/(2\sigma^2)\right\}.$$

A chi-squared random variable with degrees of freedom n, denoted by $Q_n \sim \chi^2_n$, can be represented as

$$Q_n = \sum_{i=1}^n Z_i^2,$$

where $Z_i \overset{\text{IID}}{\sim} \mathrm{N}(0,1)$. We can show that Q_n has density

$$f_n(q) = q^{n/2-1} \exp(-q/2) \Big/ \left\{2^{n/2}\Gamma(n/2)\right\}, \qquad (q > 0). \tag{B.1}$$

We can verify that the above density (B.1) is well-defined even if we change the integer n to be an arbitrary positive real number ν, and call the corresponding random variable Q_ν a chi-squared random variable with degrees of freedom ν, denoted by $Q_\nu \sim \chi^2_\nu$.

A t random variable with degrees of freedom ν can be represented as

$$t_\nu = \frac{Z}{\sqrt{Q_\nu/\nu}}$$

where $Z \sim \mathrm{N}(0,1), Q_\nu \sim \chi^2_\nu$, and $Z \perp\!\!\!\perp Q_\nu$.

An F random variable with degrees of freedom (r,s) can be represented as

$$F = \frac{Q_r/r}{Q_s/s}$$

where $Q_r \sim \chi^2_r, Q_s \sim \chi^2_s$, and $Q_r \perp\!\!\!\perp Q_s$.

B.1.2 Beta–Gamma duality

The Gamma(α,β) random variable with parameters $\alpha,\beta > 0$ has density

$$f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}, \quad (x > 0). \tag{B.2}$$

The Beta(α,β) random variable with parameters $\alpha,\beta > 0$ has density

$$f(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}, \quad (0 < x < 1).$$

These two random variables are closely related as shown in Theorem B.1 below.

Theorem B.1 (Beta–Gamma duality) *If* $X \sim \mathrm{Gamma}(\alpha,\theta), Y \sim \mathrm{Gamma}(\beta,\theta)$ *and* $X \perp\!\!\!\perp Y$*, then*

1. $X + Y \sim \mathrm{Gamma}(\alpha+\beta,\theta)$,
2. $X/(X+Y) \sim \mathrm{Beta}(\alpha,\beta)$,
3. $X + Y \perp\!\!\!\perp X/(X+Y)$.

Another simple but useful fact is that χ^2 is a special Gamma random variable. Comparing the densities in (B.1) and (B.2), we obtain the following result.

Proposition B.1 $\chi^2_n \sim \mathrm{Gamma}(n/2, 1/2)$.

We can also calculate the moments of the Gamma and Beta distributions.

Proposition B.2 *If* $X \sim \mathrm{Gamma}(\alpha,\beta)$*, then*

$$\begin{aligned} E(X) &= \frac{\alpha}{\beta}, \\ \mathrm{var}(X) &= \frac{\alpha}{\beta^2}. \end{aligned}$$

Proposition B.3 *If* $X \sim \mathrm{Gamma}(\alpha,\beta)$*, then*

$$\begin{aligned} E(\log X) &= \psi(\alpha) - \log\beta, \\ \mathrm{var}(\log X) &= \psi'(\alpha). \end{aligned}$$

Proposition B.4 *If* $X \sim \text{Beta}(\alpha, \beta)$, *then*

$$\begin{aligned} E(X) &= \frac{\alpha}{\alpha+\beta}, \\ \text{var}(X) &= \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}. \end{aligned}$$

Proposition B.5 *If* $X \sim \text{Beta}(\alpha, \beta)$, *then*

$$\begin{aligned} E(\log X) &= \psi(\alpha) - \psi(\alpha+\beta), \\ \text{var}(\log X) &= \psi'(\alpha) - \psi'(\alpha+\beta). \end{aligned}$$

I leave the proofs of the above propositions to Problem B.3.

B.1.3 Exponential, Laplace, and Gumbel distributions

An Exponential(λ) random variable $X \geq 0$ has density $f(x) = \lambda e^{-\lambda x}$, mean $1/\lambda$, median $\log 2/\lambda$ and variance $1/\lambda^2$. The standard Exponential random variable X_0 has $\lambda = 1$, and X_0/λ generates Exponential(λ).

An important feature of Exponential(λ) is the memoryless property.

Proposition B.6 (memoryless property of Exponential) *If* $X \sim \text{Exponential}(\lambda)$, *then*

$$\text{pr}(X \geq x + c \mid X \geq c) = \text{pr}(X \geq x).$$

Proposition B.6 states that if X represents the survival time, then the probability of surviving another x time is always the same no matter how long the existing survival time is. I leave the proof of Proposition B.6 to Problem B.4.

The minimum of independent exponential random variables also follows an exponential distribution.

Proposition B.7 *Assume that* $X_i \sim \text{Exponential}(\lambda_i)$ *are independent* $(i = 1, \ldots, n)$. *Then*

$$\underline{X} = \min(X_1, \ldots, X_n) \sim \text{Exponential}(\lambda_1 + \cdots + \lambda_n)$$

and

$$\text{pr}(X_i = \underline{X}) = \frac{\lambda_i}{\lambda_1 + \cdots + \lambda_n}.$$

I leave the proof of Proposition B.7 to Problem B.5. Theorem **??** states a more general result without assuming the Exponential distribution.

The difference between two IID exponential random variables follows the Laplace distribution.

Proposition B.8 *If* y_1 *and* y_2 *are two IID Exponential*(λ), *then* $y = y_1 - y_2$ *has density*

$$\frac{\lambda}{2} \exp(-\lambda|c|), \quad -\infty < c < \infty$$

which is the density of a Laplace distribution with mean 0 *and variance* $2/\lambda^2$.

I leave the proof of Proposition B.8 to Problem B.6.

If X_0 is the standard exponential random variable, then we define the Gumbel(μ, β) random variable as

$$Y = \mu - \beta \log X_0.$$

The standard Gumbel distribution has $\mu = 0$ and $\beta = 1$, with cumulative distribution function (CDF)

$$F(y) = \exp(-e^{-y}), \quad y \in \mathbb{R}$$

and density

$$f(y) = \exp(-e^{-y})e^{-y}, \quad y \in \mathbb{R}.$$

By definition and Proposition B.7, we can verify that the maximum of IID Gumbels is also Gumbel.

Proposition B.9 *If $Y_1, \ldots, Y_n$ are IID* Gumbel(μ, β)*, then*

$$\max_{1 \leq i \leq n} Y_i \sim \text{Gumbel}(\mu + \beta \log n, \beta).$$

If $Y_1, \ldots, Y_n$ are independent Gumbel$(\mu_i, 1)$*, then*

$$\max_{1 \leq i \leq n} Y_i \sim \text{Gumbel}\left(\log \sum_{i=1}^{n} e^{\mu_i}, 1\right).$$

I leave the proof to Problem B.7.

B.2 Multivariate distributions

A random vector $(X_1, \ldots, X_n)^{\mathrm{T}}$ is a vector consisting of n random variables. If all components are continuous, we can define its joint density $f_{X_1 \cdots X_n}(x_1, \ldots, x_n)$.

For a random vector $\binom{X}{Y}$ with X and Y possibly being vectors, if it has joint density $f_{XY}(x, y)$, then we can obtain the marginal distribution of X

$$f_X(x) = \int f_{XY}(x, y) \mathrm{d}y$$

and define the conditional density

$$f_{Y|X}(y \mid x) = \frac{f_{XY}(x, y)}{f_X(x)} \qquad \text{if } f_X(x) \neq 0.$$

Based on the conditional density, we can define the conditional expectation of any function of Y as

$$E\{g(Y) \mid X = x\} = \int g(y) f_{Y|X}(y \mid x) \mathrm{d}y$$

and the conditional variance as

$$\text{var}\{g(Y) \mid X = x\} = E\left[\{g(Y)\}^2 \mid X = x\right] - [E\{g(Y) \mid X = x\}]^2.$$

In the above definitions, the conditional mean and variance are both deterministic functions of x. We can replace x by the random variable X to define $E\{g(Y) \mid X\}$ and var$\{g(Y) \mid X\}$, which are functions of the random variable X and are thus random variables.

Below are two important laws of conditional expectation and variance.

Theorem B.2 (Law of total expectation) *We have*

$$E(Y) = E\{E(Y \mid X)\}.$$

Theorem B.3 (Law of total variance or analysis of variance) *We have*

$$\text{var}(Y) = E\{\text{var}(Y \mid X)\} + \text{var}\{E(Y \mid X)\}.$$

Independence

Random variables $(X_1, \ldots, X_n)$ are mutually independent if

$$f_{X_1 \cdots X_n}(x_1, \ldots, x_n) = f_{X_1}(x_1) \cdots f_{X_n}(x_n).$$

In the definition of independence, each of $(X_1, \ldots, X_n)$ can be vectors. We have the following rules under independence.

Proposition B.10 *If $X \perp\!\!\!\perp Y$, then $h(X) \perp\!\!\!\perp g(Y)$ for any functions $h(\cdot)$ and $g(\cdot)$.*

Proposition B.11 *If $X \perp\!\!\!\perp Y$, then*

$$\begin{aligned} f_{XY}(x, y) &= f_X(x) f_Y(y), \\ f_{Y|X}(y \mid x) &= f_Y(y), \\ E\{g(Y) \mid X\} &= E\{g(Y)\}, \\ E\{g(Y)h(X)\} &= E\{g(Y)\} E\{h(X)\}. \end{aligned}$$

Expectations of random vectors or random matrices

For a random matrix $W = (w_{ij})$, we define $E(W) = (E(w_{ij}))$. For constant matrices A and C, we can verify that

$$\begin{aligned} E(AW + C) &= AE(W) + C, \\ E(AWC) &= AE(W)C. \end{aligned}$$

Covariance between two random vectors

If $W \in \mathbb{R}^r$ and $Y \in \mathbb{R}^s$, then their covariance

$$\text{cov}(W, Y) = E\left[\{W - E(W)\}\{Y - E(Y)\}^{\text{T}}\right]$$

is an $r \times s$ matrix. As a special case,

$$\text{cov}(Y) = \text{cov}(Y, Y) = E\left[\{Y - E(Y)\}\{Y - E(Y)\}^{\text{T}}\right] = E(YY^{\text{T}}) - E(Y)E(Y)^{\text{T}}.$$

For a scalar random variable, $\text{cov}(Y) = \text{var}(Y)$.

Proposition B.12 *For $A \in \mathbb{R}^{r \times n}, Y \in \mathbb{R}^n$ and $C \in \mathbb{R}^r$, we have*

$$\text{cov}(AY + C) = A\text{cov}(Y)A^{\text{T}}.$$

Using Proposition B.12, we can verify that for any n-dimensional random vector, $\text{cov}(Y) \succeq 0$ because for all $x \in \mathbb{R}^n$, we have

$$x^{\text{T}}\text{cov}(Y)x = \text{cov}(x^{\text{T}}Y) = \text{var}(x^{\text{T}}Y) \geq 0.$$

Proposition B.13 *For two random vectors W and Y, we have*

$$\text{cov}(AW + C, BY + D) = A\text{cov}(W, Y)B^{\text{T}}$$

and

$$\text{cov}(AW + BY) = A\text{cov}(W)A^{\text{T}} + B\text{cov}(Y)B^{\text{T}} + A\text{cov}(W, Y)B^{\text{T}} + B\text{cov}(Y, W)A^{\text{T}}.$$

Similar to Theorem B.3, we have the following decomposition of the covariance.

Theorem B.4 (Law of total covariance) *We have*

$$\text{cov}(Y, W) = E\{\text{cov}(Y, W \mid X)\} + \text{cov}\{E(Y \mid X), E(W \mid X)\}.$$

B.3 Multivariate Normal and its properties

I use a generative definition of the multivariate Normal random vector. First, Z is a standard Normal random vector if $Z = (Z_1, \ldots, Z_n)^{\mathrm{T}}$ has components $Z_i \overset{\mathrm{IID}}{\sim} \mathrm{N}(0,1)$. Given a mean vector μ and a positive semi-definite covariance matrix Σ, define a Normal random vector $Y \sim \mathrm{N}(\mu, \Sigma)$ with mean μ and covariance Σ if Y can be represented as

$$Y = \mu + AZ, \tag{B.3}$$

where A satisfies $\Sigma = AA^{\mathrm{T}}$. We can verify that $\mathrm{cov}(Y) = \Sigma$, so indeed Σ is its covariance matrix. If $\Sigma \succ 0$, then we can verify that Y has density

$$f_Y(y) = (2\pi)^{-n/2} \left\{\det(\Sigma)\right\}^{-1/2} \exp\left\{-(y-\mu)^{\mathrm{T}}\Sigma^{-1}(y-\mu)/2\right\}. \tag{B.4}$$

We can easily verify the following result by calculating the density.

Proposition B.14 *If $Z \sim \mathrm{N}(0, I_n)$ and Γ is an orthogonal matrix, then $\Gamma Z \sim \mathrm{N}(0, I_n)$.*

I do not define multivariate Normal based on the density (B.4) because it is only well defined with a positive definite Σ. I do not define multivariate Normal based on the characteristic function because it is more advanced than the level of this book. Definition (B.3) does not require Σ to be positive definite and is more elementary. However, it has a subtle issue of uniqueness. Although the decomposition $\Sigma = AA^{\mathrm{T}}$ is not unique, the resulting distribution $Y = \mu + AZ$ is. We can verify this using the Polar decomposition in (A.8). Because $A = \Sigma^{1/2}\Gamma$ where Γ is an orthogonal matrix, we have $Y = \mu + \Sigma^{1/2}\Gamma Z = \mu + \Sigma^{1/2}\widetilde{Z}$ where $\widetilde{Z} = \Gamma Z$ is a standard Normal random vector by Proposition B.14. Importantly, although the definition (B.3) can be general, we usually use the following representation

$$Y = \mu + \Sigma^{1/2} Z.$$

Theorem B.5 *Assume that*

$$\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \right).$$

Then $Y_1 \perp\!\!\!\perp Y_2$ if and only if $\Sigma_{12} = 0$.

I leave the proof of Theorem B.5 as Problem B.8.

Proposition B.15 *If $Y \sim \mathrm{N}(\mu,\Sigma)$, then $BY + C \sim \mathrm{N}(B\mu + C, B\Sigma B^{\mathrm{T}})$, that is, any linear transformation of a Normal random vector is also a Normal random vector.*

I leave the proof of Proposition B.15 as Problem B.9.

An obvious corollary of Proposition B.15 is that if $X_1 \sim \mathrm{N}(\mu_1, \sigma_1^2)$ and $X_2 \sim \mathrm{N}(\mu_2, \sigma_2^2)$ are independent, then $X_1 + X_2 \sim \mathrm{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$. So the summation of two independent Normals is also Normal. Remarkably, the reverse of the result is also true.

Theorem B.6 (Levy–Cramer) *If $X_1 \perp\!\!\!\perp X_2$ and $X_1 + X_2$ is Normal, then both X_1 and X_2 must be Normal.*

The statement of Theorem B.6 is extremely simple. But its proof is non-trivial and beyond the scope of this book. See Benhamou et al. (2018) for a proof.

Theorem B.7 *Assume*

$$\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \right).$$

1. *The marginal distributions are Normal:*

$$Y_1 \sim \mathrm{N}\left(\mu_1, \Sigma_{11}\right),$$
$$Y_2 \sim \mathrm{N}\left(\mu_2, \Sigma_{22}\right).$$

2. *If $\Sigma_{22} \succ 0$, then the conditional distribution is Normal:*

$$Y_1 \mid Y_2 = y_2 \sim \mathrm{N}\left(\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(y_2 - \mu_2), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\right);$$

Y_2 is independent of the residual

$$Y_1 - \Sigma_{12}\Sigma_{22}^{-1}(Y_2 - \mu_2) \sim \mathrm{N}\left(\mu_1, \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\right).$$

I review some other results of the multivariate Normal below.

Proposition B.16 *Assume $Y \sim \mathrm{N}(\mu, \sigma^2 I_n)$. If $AB^{\mathrm{T}} = 0$, then $AY \perp\!\!\!\perp BY$.*

Proposition B.17 *Assume*

$$\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} \right),$$

where ρ is the correlation coefficient defined as

$$\rho = \frac{\mathrm{cov}(Y_1, Y_2)}{\sqrt{\mathrm{var}(Y_1)\mathrm{var}(Y_2)}}.$$

Then the conditional distribution is

$$Y_1 \mid Y_2 = y_2 \sim \mathrm{N}\left(\mu_1 + \rho\frac{\sigma_1}{\sigma_2}(y_2 - \mu_2), \sigma_1^2(1 - \rho^2)\right).$$

B.4 Quadratic forms of random vectors

Given a random vector Y and a symmetric matrix A, we can define the quadratic form $Y^{\mathrm{T}}AY$, which is a random variable playing an important role in statistics. The first theorem is about its mean.

Theorem B.8 *If Y has mean μ and covariance Σ, then*

$$E(Y^{\mathrm{T}}AY) = \mathrm{trace}(A\Sigma) + \mu^{\mathrm{T}}A\mu.$$

The proof below uses the following three basic facts.

(F1) $E(YY^{\mathrm{T}}) = \mathrm{cov}(Y) + E(Y)E(Y^{\mathrm{T}}) = \Sigma + \mu\mu^{\mathrm{T}}$.

(F2) For an $n \times n$ symmetric random matrix $W = (w_{ij})$, we have $E\{\mathrm{trace}(W)\} = \mathrm{trace}\{E(W)\}$ because $E\left(\sum_{i=1}^n w_{ii}\right) = \sum_{i=1}^n E(w_{ii})$.

(F3) If BC and CB are both well defined, then $\text{trace}(BC) = \text{trace}(CB)$.

Proof of Theorem B.8: The conclusion follows from

$$\begin{aligned} E(Y^{\mathrm{T}}AY) &= E\{\text{trace}(Y^{\mathrm{T}}AY)\} \quad (\text{because } Y^{\mathrm{T}}AY \text{ is a scalar}) \\ &= E\{\text{trace}(AYY^{\mathrm{T}})\} \quad (\text{by (F3)}) \\ &= \text{trace}\{E(AYY^{\mathrm{T}})\} \quad (\text{by (F2)}) \\ &= \text{trace}\{AE(YY^{\mathrm{T}})\} \\ &= \text{trace}\{A(\Sigma + \mu\mu^{\mathrm{T}})\} \quad (\text{by (F1)}) \\ &= \text{trace}(A\Sigma) + \text{trace}(A\mu\mu^{\mathrm{T}}) \\ &= \text{trace}(A\Sigma) + \text{trace}(\mu^{\mathrm{T}}A\mu) \quad (\text{by (F3)}) \\ &= \text{trace}(A\Sigma) + \mu^{\mathrm{T}}A\mu. \quad (\text{because } \mu^{\mathrm{T}}A\mu \text{ is a scalar}) \end{aligned}$$

□

The variance of the quadratic form is much more complicated for a general random vector. For the multivariate Normal random vector, we have the following formula.

Theorem B.9 *If $Y \sim \mathrm{N}(\mu, \Sigma)$, then*

$$\text{var}(Y^{\mathrm{T}}AY) = 2\text{trace}(A\Sigma A\Sigma) + 4\mu^{\mathrm{T}}A\Sigma A\mu.$$

I relegate the proof as Problem B.15.

From its definition, χ^2_n is the summation of the squares of n IID standard Normal random variables. It is closely related to quadratic forms of multivariate Normals.

Theorem B.10 *We have the following results on the χ^2 random variables.*

1. *If $Y \sim \mathrm{N}(\mu, \Sigma)$ is an n-dimensional random vector with $\Sigma \succ 0$, then*

$$(Y - \mu)^{\mathrm{T}}\Sigma^{-1}(Y - \mu) \sim \chi^2_n.$$

If $\text{rank}(\Sigma) = k \le n$, then

$$(Y - \mu)^{\mathrm{T}}\Sigma^{+}(Y - \mu) \sim \chi^2_k.$$

2. *If $Y \sim \mathrm{N}(0, I_n)$ and H is a projection matrix of rank K, then*

$$Y^{\mathrm{T}}HY \sim \chi^2_K.$$

3. *If $Y \sim \mathrm{N}(0, H)$ where H is a projection matrix of rank K, then*

$$Y^{\mathrm{T}}Y \sim \chi^2_K.$$

Proof of Theorem B.10:

1. I only prove the general result with $\text{rank}(\Sigma) = k \le n$. By definition, $Y = \mu + \Sigma^{1/2}Z$ where Z is a standard Normal random vector, then

$$\begin{aligned} (Y - \mu)^{\mathrm{T}}\Sigma^{+}(Y - \mu) &= Z^{\mathrm{T}}\Sigma^{1/2}\Sigma^{+}\Sigma^{1/2}Z \\ &= \sum_{i=1}^{k} Z_i^2 \sim \chi^2_k. \end{aligned}$$

2. Using the eigendecomposition of the projection matrix
$$H = P\text{diag}\{1, \dots, 1, 0, \dots, 0\} P^{\mathrm{T}}$$
with K 1's in the diagonal matrix, we have
$$\begin{aligned} Y^{\mathrm{T}}HY &= Y^{\mathrm{T}}P\text{diag}\{1, \dots, 1, 0, \dots, 0\} P^{\mathrm{T}}Y \\ &= Z^{\mathrm{T}}\text{diag}\{1, \dots, 1, 0, \dots, 0\} Z, \end{aligned}$$
where $Z = (Z_1, \dots, Z_n)^{\mathrm{T}} = P^{\mathrm{T}}Y \sim \mathrm{N}(0, P^{\mathrm{T}}P) = \mathrm{N}(0, I_n)$ is a standard Normal random vector. So
$$Y^{\mathrm{T}}HY = \sum_{i=1}^{K} Z_i^2 \sim \chi^2_K.$$

3. Writing $Y = H^{1/2}Z$ where Z is a standard Normal random vector, we have
$$Y^{\mathrm{T}}Y = Z^{\mathrm{T}}H^{1/2}H^{1/2}Z = Z^{\mathrm{T}}HZ \sim \chi^2_K$$
using the second result.

□

B.5 Homework problems

B.1 Uniform moments

Let X be Uniform$(0, 1)$. Find $E(X^k)$ for $k = 1, 2, \dots$.

B.2 Beta-Gamma duality

Prove Theorem B.1.

Remark: Calculate the joint density of $(X + Y, X/(X + Y))$.

B.3 Gamma and Beta moments

Prove Propositions B.2–B.5.

B.4 Memoryless property of Exponential

Prove Proposition B.6.

B.5 Minimum of independent Exponentials

Prove Proposition B.7.

B.6 Laplace as the difference between two IID Exponentials

Prove Proposition B.8.

B.7 Maximums of Gumbels

Prove Proposition B.9.

B.8 Independence and uncorrelatedness in the multivariate Normal

Prove Theorem B.5.

B.9 Linear transformation of Normal

Prove Proposition B.15.

B.10 Transformation of bivariate Normal

Prove that if $(Y_1, Y_2)^{\mathrm{T}}$ follows a bivariate Normal distribution

$$\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right),$$

then

$$Y_1 + Y_2 \perp\!\!\!\perp Y_1 - Y_2.$$

Remark: This result holds for arbitrary ρ.

B.11 Normal conditional distributions

Suppose that (X_1, X_2) has the joint distribution

$$f_{X_1 X_2}(x_1, x_2) \propto C_0 \exp\left\{ -\frac{1}{2}\left(Ax_1^2 x_2^2 + x_1^2 + x_2^2 - 2Bx_1x_2 - 2C_1x_1 - 2C_2x_2 \right) \right\},$$

where C_0 is the normalizing constant depending on (A, B, C_1, C_2). To ensure that this is a well-defined density, we need $A \geq 0$, and if $A = 0$ then $|B| < 1$.

Prove that the conditional distributions are

$$\begin{aligned} X_1 \mid X_2 = x_2 &\sim \mathrm{N}\left(\frac{Bx_2 + C_1}{Ax_2^2 + 1}, \frac{1}{Ax_2^2 + 1} \right), \\ X_2 \mid X_1 = x_1 &\sim \mathrm{N}\left(\frac{Bx_1 + C_2}{Ax_1^2 + 1}, \frac{1}{Ax_1^2 + 1} \right). \end{aligned}$$

Remark: For a bivariate Normal distribution, the two conditional distributions are both Normal. The converse of the statement is not true. That is, even if the two conditional distributions are both Normal, the joint distribution may not be bivariate Normal. Gelman and Meng (1991) reported this result.

B.12 Inverse of covariance matrix and conditional independence in multivariate Normal

Assume $X = (X_1, \ldots, X_p)^{\mathrm{T}} \sim \mathrm{N}(\mu, \Sigma)$. Denote the inverse of its covariance matrix by $\Sigma^{-1} = (\sigma^{jk})_{1 \leq j,k \leq p}$.

Prove that for any pair of $j \neq k$, we have

$$\sigma^{jk} = 0 \Longleftrightarrow X_j \perp\!\!\!\perp X_k \mid X_{\setminus(j,k)},$$

where $X_{\setminus(j,k)}$ contains all the variables except X_j and X_k.

Remark: This basic property of multivariate Normal motivates the Gaussian Graphical Model, which uses an undirected graph to illustrate the conditional independence relationship among random variables $X_1, \ldots, X_p$ (Dempster, 1972). In particular, $\sigma^{jk} = 0$ if and only if the edge between X_j and X_k is missing.

B.13 Independence of linear and quadratic functions of the multivariate Normal

Assume $Y \sim \mathrm{N}(\mu, \sigma^2 I_n)$. For an n dimensional vector a and two $n \times n$ symmetric matrices A and B, prove that

1. if $Aa = 0$, then $a^{\mathrm{T}}Y \perp\!\!\!\perp Y^{\mathrm{T}}AY$;
2. if $AB = BA = 0$, then $Y^{\mathrm{T}}AY \perp\!\!\!\perp Y^{\mathrm{T}}BY$.

Remark: To simplify the proof, you can use the pseudoinverse of A which satisfies $AA^{+}A = A$. In fact, a strong result holds. Ogasawara and Takahashi (1951) proved the following theorem; see also Styan (1970, Theorem 5).

Theorem B.11 *Assume $Y \sim N(\mu, \Sigma)$. Define quadratic forms $Y^{\mathrm{T}}AY$ and $Y^{\mathrm{T}}BY$ for two symmetric matrices A and B. The $Y^{\mathrm{T}}AY$ and $Y^{\mathrm{T}}BY$ are independent if and only if*

$$\Sigma A \Sigma B \Sigma = 0, \quad \Sigma A \Sigma B \mu = \Sigma B \Sigma A \mu = 0, \quad \mu^{\mathrm{T}} A \Sigma B \mu = 0.$$

B.14 Independence of the sample mean and variance of IID Normals

Theorem B.12 below is a fundamental result on IID Normals. Prove Theorem B.12.

Theorem B.12 *If $X_1, \ldots, X_n \overset{\mathrm{IID}}{\sim} \mathrm{N}(\mu, \sigma^2)$, then $\bar{X} \perp\!\!\!\perp S^2$, where $\bar{X} = n^{-1}\sum_{i=1}^{n} X_i$ and $S^2 = (n-1)^{-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$.*

Remark: A remarkable result due to Geary (1936) ensures the reverse of Theorem B.12. That is, if $X_1, \ldots, X_n$ are IID and $\bar{X} \perp\!\!\!\perp S^2$, then $X_1, \ldots, X_n$ must be Normals. See Lukacs (1942) and Benhamou et al. (2018) for proofs.

B.15 Variance of the quadratic form of the multivariate Normal

First prove Theorem B.9 with a symmetric A. Then prove Theorem B.13 below.

Theorem B.13 *Assume A_1 and A_2 are symmetric matrices. If $Y \sim \mathrm{N}(\mu, \Sigma)$, then*

$$\mathrm{cov}(Y^{\mathrm{T}}A_1Y, Y^{\mathrm{T}}A_2Y) = 2\mathrm{trace}(A_1\Sigma A_2\Sigma) + 4\mu^{\mathrm{T}}A_1\Sigma A_2\mu.$$

Remark: Theorem B.9 is a special case of Theorem B.13. You can prove Theorem B.13 and then Theorem B.9 follows immediately. You can also first prove Theorem B.9 and then use it to prove Theorem B.13. You can write $Y = \mu + \Sigma^{1/2}Z$ and reduce the problem to calculating the moments of standard Normals.

Bibliography

Belkin, M., Hsu, D., Ma, S., and Mandal, S. (2019). Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854.

Benhamou, E., Guez, B., and Paris, N. (2018). Three remarkable properties of the normal distribution for sample variance. *Theoretical Mathematics and Applications*, 8:1792–9709.

Berman, M. (1988). A theorem of Jacobi and its generalization. *Biometrika*, 75:779–783.

Chen, J. and Li, X. (2019). Model-free nonconvex matrix completion: Local minima analysis and applications in memory-efficient kernel PCA. *Journal of Machine Learning Research*, 20:1–39.

Dempster, A. P. (1972). Covariance selection. *Biometrics*, 28:157–175.

Ding, P. (2024). Linear model and extensions. *arXiv preprint arXiv:2401.00649*.

Farebrother, R. W. (1976). Further results on the mean square error of ridge regression. *Journal of the Royal Statistical Society. Series B (Methodological)*, 38:248–250.

Galton, F. (1886). Regression towards mediocrity in hereditary stature. *The Journal of the Anthropological Institute of Great Britain and Ireland*, 15:246–263.

Geary, R. C. (1936). The distribution of "Student's" ratio for non-normal samples. *Supplement to the Journal of the Royal Statistical Society*, 3:178–184.

Gelman, A. and Meng, X.-L. (1991). A note on bivariate distributions that are conditionally normal. *American Statistician*, 45:125–126.

Gelman, A. and Park, D. K. (2009). Splitting a predictor at the upper quarter or third and the lower quarter or third. *American Statistician*, 63:1–8.

Hastie, T., Montanari, A., Rosset, S., and Tibshirani, R. J. (2022). Surprises in high-dimensional ridgeless least squares interpolation. *Annals of statistics*, 50(2):949.

Lukacs, E. (1942). A characterization of the normal distribution. *Annals of Mathematical Statistics*, 13:91–93.

Ogasawara, T. and Takahashi, M. (1951). Independence of quadratic quantities in a normal system. *Journal of Science of the Hiroshima University, Series A*, 15:1–9.

Ruhe, A. (1970). Perturbation bounds for means of eigenvalues and invariant subspaces. *BIT Numerical Mathematics*, 10:343–354.

Stigler, S. M. (1981). Gauss and the invention of least squares. *Annals of Statistics*, 9:465–474.

Styan, G. P. H. (1970). Notes on the distribution of quadratic forms in singular normal variables. *Biometrika*, 57:567–572.

Subrahmanyam, M. (1972). A property of simple least squares estimates. *Sankhyā*, 34:355–356.

Theobald, C. M. (1974). Generalizations of mean square error applied to ridge regression. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 36:103–106.

Von Neumann, J. (1937). Some matrix-inequalities and metrization of matric-space. *Tomsk Univ. Rev*, 1:286–300.

Woolley, E. B. (1941). The method of minimized areas as a basis for correlation analysis. *Econometrica*, 9:38–62.

Wu, C.-F. J. (1986). Jackknife, bootstrap and other resampling methods in regression analysis. *Annals of Statistics*, 14:1261–1295.