Auxiliary Material for Econometrics

© Dimitrios D. Thomakos[1]

This version: October 17, 2009

[1] Preliminary material. Not to be quoted without permission from the author.

Chapter 1

Review of Matrix Algebra

1.1 Notational Conventions

In this section we outline some notational conventions used throughout the text for scalars, vectors, matrices etc. We follow these conventions consistently and any deviations from them are clearly indicated.

1. Scalars are given in lowercase roman italic type. For example, x and θ are to be taken as scalars.

2. Vectors are given in lowercase bold italic type. For example, $\boldsymbol{x}$ and $\boldsymbol{\theta}$ are to be taken as vectors. All vectors are taken as column vectors unless otherwise noted. The elements of a vector are denoted by lowercase italic type subscripted by italic roman letters. For example, the elements of the $(m \times 1)$ vector $\boldsymbol{x}$ are denoted by $x_1, x_2, ..., x_m$ and we

have that:

$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{bmatrix} \tag{1.1}$$

3. Matrices are given in uppercase bold type. For example, $\boldsymbol{X}$ and $\boldsymbol{\Sigma}$ are to be taken as matrices. The elements of a matrix are denoted by lowercase italic type (doubly) subscripted by italic roman letters. For example, the elements of the $(m \times n)$ matrix $\boldsymbol{X}$ are denoted by x_{ij}, for $i = 1, 2, ..., m$ and $j = 1, 2, ..., n$ and we have that:

$$\boldsymbol{X} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1n} \\ x_{21} & x_{21} & \dots & x_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ x_{m1} & x_{m2} & \dots & x_{mn} \end{bmatrix} \tag{1.2}$$

4. The transpose of a vector or a matrix is denoted by the traspose symbol, $^\top$. For example, the transpose of an arbitrary matrix $\boldsymbol{X}$ is denoted by $\boldsymbol{X}^\top$. Note that in many other textbooks a prime $'$ is used in place of $^\top$.

5. The rows of a matrix are denoted by lowercase bold italic type, subscripted by italic roman letters and primed. The columns of a matrix are denoted by uppercase bold type, subscripted by italic roman letters. For example, the rows of the $(m \times n)$ matrix $\boldsymbol{X}$ are denoted by $\boldsymbol{x}_i^\top$, for $i = 1, 2, ..., m$; the columns of $\boldsymbol{X}$ are denoted by $\boldsymbol{X}_j$, for $j = 1, 2, ..., n$

and we have that:

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}_1^\top \\ \boldsymbol{x}_2^\top \\ \vdots \\ \boldsymbol{x}_m^\top \end{bmatrix} \quad \text{and} \quad \boldsymbol{X} = [\boldsymbol{X}_1, \boldsymbol{X}_2 \dots \boldsymbol{X}_n] \tag{1.3}$$

6. Sets are given in uppercase italic type, for roman letters, and plain uppercase type for Greek letters. For example, A, B and Θ are to be taken as sets.

7. The set of all integers, including zero, will be denoted by $\mathbb{N}$. The set of all positive integers will be denoted by $\mathbb{N}_+$ and the set of positive integers, including zero will be denoted $\mathbb{N}_0$. The set of all real numbers will be denoted by $\mathbb{R}$ while the set of all positive real numbers will be denoted $\mathbb{R}_+$.

8. The n-dimensional field of real numbers (i.e. the product of $\mathbb{R}$ with itself n times) will be denoted $\mathbb{R}^n$.

9. When first defining a new object, such as a new equation, we use the notation $\overset{\text{def}}{=}$ instead of simply using the $=$ sign.

1.2 Basic Results

A number of useful, and repeatedly used, results are summarized in this section. We mainly use definitions, propositions and examples to set forth these results.

Definition 1. Let $\boldsymbol{e}$ be an $(n \times 1)$ vector with all of its elements equal to unity, that is $\boldsymbol{e} \overset{\text{def}}{=} [1, 1, \ldots, 1]^\top$. We call $\boldsymbol{e}$ the unit vector.

Definition 2. Let $\boldsymbol{I}$ be an $(n \times n)$ matrix with all of its elements equal to zero, except those on the main diagonal that are equal to unity, that is:

$$\boldsymbol{I} \overset{\text{def}}{=} \begin{bmatrix} 1 & 0 & \ldots & 0 \\ 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \ldots & 1 \end{bmatrix} = \begin{bmatrix} \boldsymbol{i}_1^\top \\ \boldsymbol{i}_2^\top \\ \vdots \\ \boldsymbol{i}_n^\top \end{bmatrix} \tag{1.4}$$

We call $\boldsymbol{I}$ the identity matrix.

Definition 3. For any two $(n \times 1)$ vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ we define their inner product as $\langle \boldsymbol{x}, \boldsymbol{y} \rangle \overset{\text{def}}{=} \boldsymbol{x}^\top \boldsymbol{y} = \boldsymbol{y}^\top \boldsymbol{x} = \sum_{i=1}^{n} x_i y_i$. Note that the inner product of two vectors is a scalar. The inner product allows for a concise representation of many formulas involving summations. Here are some examples.

1. Let $\boldsymbol{x} = \boldsymbol{e}$; then $\langle \boldsymbol{e}, \boldsymbol{y} \rangle = \sum_{i=1}^{n} y_i$. Equivalently, $n^{-1}\langle \boldsymbol{e}, \boldsymbol{y} \rangle = \bar{y}$, the sample mean of all values of $\boldsymbol{y}$.

2. Let $\boldsymbol{x} = \boldsymbol{y}$; then $\langle \boldsymbol{x}, \boldsymbol{x} \rangle = \sum_{i=1}^{n} x_i^2$.

3. Let $\boldsymbol{z} = \boldsymbol{y} - \boldsymbol{e}\bar{y}$; then $(n-1)^{-1}\langle \boldsymbol{z}, \boldsymbol{z} \rangle = (n-1)^{-1} \sum_{i=1}^{n} (y_i - \bar{y})^2 = s^2$, the sample variance of all values of $\boldsymbol{y}$.

Definition 4. Let $\boldsymbol{I}$ to be an $(n \times n)$ identity matrix and let $\boldsymbol{e}$ be an $(n \times 1)$ unit vector. Define the matrix $\boldsymbol{D} \overset{\text{def}}{=} \left(\boldsymbol{I} - n^{-1}\boldsymbol{e}\boldsymbol{e}^\top\right)$. We call $\boldsymbol{D}$ the demeaning matrix, since for any $(n \times k)$ matrix $\boldsymbol{X}$ we have that:

$$\begin{gathered}
\boldsymbol{D}\boldsymbol{X} = \left(\boldsymbol{I} - n^{-1}\boldsymbol{e}\boldsymbol{e}^\top\right)\boldsymbol{X} = \boldsymbol{X} - \boldsymbol{e}\left(n^{-1}\boldsymbol{e}^\top\boldsymbol{X}\right) = \\
\boldsymbol{X} - \boldsymbol{e}\left[n^{-1}\boldsymbol{e}^\top\boldsymbol{X}_1, n^{-1}\boldsymbol{e}^\top\boldsymbol{X}_2, \ldots, n^{-1}\boldsymbol{e}^\top\boldsymbol{X}_k\right] = \\
\boldsymbol{X} - \boldsymbol{e}\left[\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_k\right]
\end{gathered} \tag{1.5}$$

where $\bar{x}_j$ is the sample mean of the j^{th} column of $\boldsymbol{X}$, $j = 1, 2, \ldots, k$. Thus, applying $\boldsymbol{D}$ to $\boldsymbol{X}$ we substracted from each column the sample mean of that column's observations.

Definition 5. For any $(n \times 1)$ vector $\boldsymbol{x}$ we define its (Euclidean) norm $\|\boldsymbol{x}\|$ as the square root of the inner product of $\boldsymbol{x}$ with itself, $\|\boldsymbol{x}\| \overset{\text{def}}{=} \sqrt{\langle \boldsymbol{x}, \boldsymbol{x} \rangle}$.

Definition 6. If any two $(n \times 1)$ vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ have inner product $\langle \boldsymbol{x}, \boldsymbol{y} \rangle = 0$ we call them orthogonal. If, in addition, $\|\boldsymbol{x}\| = 1$ and $\|\boldsymbol{y}\| = 1$ we call the vectors orthonormal.

Definition 7. A square $(n \times n)$ matrix $\boldsymbol{X}$ is said to be orthogonal if its columns are orthonormal vectors.

Definition 8. A square $(n \times n)$ matrix $\boldsymbol{X}$ is said to be symmetric if it equals its transpose, i.e. $\boldsymbol{X} = \boldsymbol{X}^\top$.

Definition 9. A square $(n \times n)$ matrix $\boldsymbol{X}$ is said to be diagonal if $x_{ij} = 0$ for all $i \neq j$, $i, j = 1, 2, ..., n$.

Definition 10. A square $(n \times n)$ matrix $\boldsymbol{X}$ is said to be scalar if it is diagonal and if $x_{ii} = x$ for all $i = 1, 2, ..., n$. Note that the identity matrix is a scalar matrix.

Definition 11. For any square $(n \times n)$ matrix $\boldsymbol{X}$ we define its trace

$\mathsf{tr}\,[\boldsymbol{X}]$ as the sum of its diagonal elements, $\mathsf{tr}\,[\boldsymbol{X}] \overset{\text{def}}{=} \sum_{i=1}^{n} x_{ii}$.

Definition 12. For any $(n \times k)$ matrix $\boldsymbol{X}$, with $n \geq k$, we define the $(k \times k)$ moment matrix (of squares and cross-products) as $\boldsymbol{S}_{\boldsymbol{X}} \overset{\text{def}}{=} (\boldsymbol{X}'\boldsymbol{X})$. Using the representations of equation (1.3) we also have that $\boldsymbol{S}_{\boldsymbol{X}}$ is given by:

$$\boldsymbol{S}_{\boldsymbol{X}} = \sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^\top = \begin{bmatrix} \boldsymbol{X}_1^\top \boldsymbol{X}_1 & \boldsymbol{X}_1^\top \boldsymbol{X}_2 & \dots & \boldsymbol{X}_1^\top \boldsymbol{X}_k \\ \boldsymbol{X}_2^\top \boldsymbol{X}_1 & \boldsymbol{X}_2^\top \boldsymbol{X}_2 & \dots & \boldsymbol{X}_2^\top \boldsymbol{X}_k \\ \vdots & \vdots & \vdots & \vdots \\ \boldsymbol{X}_k^\top \boldsymbol{X}_1 & \boldsymbol{X}_k^\top \boldsymbol{X}_2 & \dots & \boldsymbol{X}_k^\top \boldsymbol{X}_k \end{bmatrix} \tag{1.6}$$

Note that $\boldsymbol{S}_{\boldsymbol{X}}$ is a symmetric matrix.

Definition 13. For any $(n \times k)$ matrix $\boldsymbol{X}$, with $n \geq k$, we define its norm $\|\boldsymbol{X}\|$ as the square root of the trace of the moment matrix $\boldsymbol{S}_{\boldsymbol{X}}$, i.e. $\|\boldsymbol{X}\| \overset{\text{def}}{=} \sqrt{\mathsf{tr}\,[\boldsymbol{S}_{\boldsymbol{X}}]}$.

Definition 14. For any square $(n \times n)$ matrix $\boldsymbol{X}$ we define its characteristic roots as the numbers $\{r\}_{i=1}^{n}$ that are solutions (roots) to the determinantal equation $|r\boldsymbol{I} - \boldsymbol{X}| = 0$.

Proposition 1. For any $(n \times k)$ matrix $\boldsymbol{X}$, with $n \geq k$, the characteristic roots of its moment matrix $\boldsymbol{S}_{\boldsymbol{X}}$ are non-negative.

Definition 15. For any $(n \times k)$ matrix $\boldsymbol{X}$, with $n \geq k$, we define its rank $\varrho\,[\boldsymbol{X}]$ to be the number of positive characteristic roots of the moment matrix $\boldsymbol{S}_{\boldsymbol{X}}$. We say that $\boldsymbol{X}$ is of full rank if all k characteristic roots of the moment matrix $\boldsymbol{S}_{\boldsymbol{X}}$ are positive.

Definition 16. For any $(n \times k)$ matrix $\boldsymbol{X}$, with $n \geq k$, let $\{r\}_{i=1}^{k}$ be the

k characteristic roots of its moment matrix $\boldsymbol{S_X}$, arranged in ascending order. The condition number of $\boldsymbol{S_X}$ is given by $c_* \overset{\text{def}}{=} \sqrt{r_k/r_1}$.

Definition 17. For any full rank, square $(n \times n)$ matrix $\boldsymbol{X}$ we define its inverse $\boldsymbol{X}^{-1}$ as the matrix satisfying $\boldsymbol{X}^{-1}\boldsymbol{X} = \boldsymbol{X}\boldsymbol{X}^{-1} = \boldsymbol{I}$. If $\boldsymbol{X}$ is not of full rank then its inverse does not exist and we say that $\boldsymbol{X}$ is singular.

Definition 18. Let $\boldsymbol{X}$ be an $(n \times k)$ matrix, with $n \geq k$, and of full rank k. Then, the square $(n \times n)$ matrix $\boldsymbol{P_X} \overset{\text{def}}{=} \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}' = \boldsymbol{X}\boldsymbol{S}_{\boldsymbol{X}}^{-1}\boldsymbol{X}^\top$ is called the projection matrix of $\boldsymbol{X}$.

Proposition 2. Let $\boldsymbol{y}$ be an $(n \times 1)$ vector and let $\boldsymbol{X}$ be an $(n \times k)$ matrix, with $n \geq k$. There exists an $(n \times 1)$ vector $\widehat{\boldsymbol{y}} \overset{\text{def}}{=} \boldsymbol{P_X}\boldsymbol{y}$ called the projection of $\boldsymbol{y}$ on (the space spanned by the columns of) $\boldsymbol{X}$ such that $\|\boldsymbol{y} - \widehat{\boldsymbol{y}}\|^2$ is minimized.

Proposition 3. If $\boldsymbol{X}$ is an orthogonal matrix then is of full rank and $\boldsymbol{X}^\top = \boldsymbol{X}^{-1}$.

Proposition 4. The trace of a square $(n \times n)$ matrix $\boldsymbol{X}$ equals the sum of its characteristic roots, $\mathsf{tr}\left[\boldsymbol{X}\right] = \sum_{i=1}^{n} r_i$.

Definition 19. Let $\boldsymbol{A}$ be a square $(n \times n)$, symmetric matrix and $\boldsymbol{x}$ be a $(n \times 1)$ vector. We call the scalar $q_{\boldsymbol{A}} = \boldsymbol{x}^\top\boldsymbol{A}\boldsymbol{x}$ a quadratic form in $\boldsymbol{A}$.

Proposition 5. Let $\boldsymbol{A}$ be a square $(n \times n)$, symmetric matrix and $\boldsymbol{x}$ be a $(n \times 1)$ vector. Then, the following is true: $q_{\boldsymbol{A}} \equiv \mathsf{tr}\left[\boldsymbol{x}^\top\boldsymbol{A}\boldsymbol{x}\right] \equiv \mathsf{tr}\left[\boldsymbol{A}\boldsymbol{x}\boldsymbol{x}^\top\right]$.

Definition 20. Let $\boldsymbol{A}$ be a square $(n \times n)$, symmetric matrix and $\boldsymbol{x}$ be a non-zero $(n \times 1)$ vector. We say that $\boldsymbol{A}$ is positive semi-definite if the quadratic $q_{\boldsymbol{A}} = \boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x} \geq 0$. We say that $\boldsymbol{A}$ is positive definite if strict inequality holds. We say that $\boldsymbol{A}$ is a negative (semi) definite matrix if $-\boldsymbol{A}$ is positive (semi) definite. A positive (or negative) definite matrix is a full rank matrix.

Proposition 6. Let $\boldsymbol{X}$ be a square $(n \times n)$, positive definite matrix. Then, there exists a square $(n \times n)$, positive definite matrix called the square root of $\boldsymbol{X}$, denoted by $\boldsymbol{X}^{1/2}$, such that $\boldsymbol{X}$ can be decomposed as $\boldsymbol{X} \overset{\text{def}}{=} \boldsymbol{X}^{1/2^\top} \boldsymbol{X}^{1/2}$. If $\boldsymbol{X}$ is a symmetric matrix then $\boldsymbol{X}^{1/2}$ is also symmetric.

Definition 21. Let $\boldsymbol{\beta}$ be an $(k \times 1)$ vector of parameters and let $g(\boldsymbol{\beta})$ be a scalar function of $\boldsymbol{\beta}$. Then, we define the gradient vector to be the $(k \times 1)$ vector whose j^{th} element is the partial derivate of $g(\boldsymbol{\beta})$ with respect to the j^{th} element of $\boldsymbol{\beta}$. We denote the gradient vector by $\mathbb{G}\left[g(\boldsymbol{\beta})\right]$. We have:

$$\mathbb{G}\left[g(\boldsymbol{\beta})\right] \overset{\text{def}}{=} \begin{bmatrix} \frac{\partial g(\boldsymbol{\beta})}{\partial \beta_1} \\ \frac{\partial g(\boldsymbol{\beta})}{\partial \beta_2} \\ \vdots \\ \frac{\partial g(\boldsymbol{\beta})}{\partial \beta_k} \end{bmatrix} \tag{1.7}$$

Definition 22. Let $\boldsymbol{\beta}$ be an $(k \times 1)$ vector of parameters and let $g(\boldsymbol{\beta})$ be a scalar function of $\boldsymbol{\beta}$. Then, we define the Hessian matrix to be the $(k \times k)$ matrix whose $(i,j)^{th}$ element is the cross-partial derivate of $g(\boldsymbol{\beta})$ with respect to the i^{th} and the j^{th} elements of $\boldsymbol{\beta}$. We denote the Hessian matrix by $\mathbb{H}\left[g(\boldsymbol{\beta})\right]$. We have:

$$\mathbb{H}\left[g(\boldsymbol{\beta})\right] \overset{\text{def}}{=} \begin{bmatrix} \frac{\partial^2 g(\boldsymbol{\beta})}{\partial \beta_1^2} & \frac{\partial^2 g(\boldsymbol{\beta})}{\partial \beta_1 \partial \beta_2} & \cdots & \frac{\partial^2 g(\boldsymbol{\beta})}{\partial \beta_1 \partial \beta_k} \\ \frac{\partial^2 g(\boldsymbol{\beta})}{\partial \beta_2 \partial \beta_1} & \frac{\partial^2 g(\boldsymbol{\beta})}{\partial \beta_2^2} & \cdots & \frac{\partial^2 g(\boldsymbol{\beta})}{\partial \beta_2 \partial \beta_k} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial^2 g(\boldsymbol{\beta})}{\partial \beta_k \partial \beta_1} & \frac{\partial^2 g(\boldsymbol{\beta})}{\partial \beta_k \partial \beta_2} & \cdots & \frac{\partial^2 g(\boldsymbol{\beta})}{\partial \beta_k^2} \end{bmatrix} \tag{1.8}$$

Definition 23. Let $\boldsymbol{\beta}$ be an $(k \times 1)$ vector of parameters and let $g(\boldsymbol{\beta})$ be a scalar function of $\boldsymbol{\beta}$. Then, we define the second-order Taylor series approximation of $g(\boldsymbol{\beta})$ around a vector $\boldsymbol{\beta}_0$ as:

$$\begin{aligned} g(\boldsymbol{\beta}) &\approx g(\boldsymbol{\beta}_0) + (\boldsymbol{\beta} - \boldsymbol{\beta}_0)^\top \mathbb{G}\left[g(\boldsymbol{\beta}0)\right] + (\boldsymbol{\beta} - \boldsymbol{\beta}_0)^\top \mathbb{H}\left[g(\boldsymbol{\beta}_0)\right](\boldsymbol{\beta} - \boldsymbol{\beta}_0) \Rightarrow \\ g(\boldsymbol{\beta}) &\approx g(\boldsymbol{\beta}_0) + \sum_{i=1}^{k}(\beta_i - \beta_{i0})\frac{\partial g(\boldsymbol{\beta}_0)}{\partial \beta_i} + \sum_{i=1}^{k}\sum_{j=1}^{k}(\beta_i - \beta_{i0})(\beta_j - \beta_{j0})\frac{\partial^2 g(\boldsymbol{\beta}_0)}{\partial \beta_i \partial \beta_j} \end{aligned} \tag{1.9}$$

Definition 24. Let $\boldsymbol{\beta}$ be an $(k \times 1)$ vector of parameters and let $\boldsymbol{g}(\boldsymbol{\beta})$ be an $(n \times 1)$ vector function of $\boldsymbol{\beta}$. Then, we define the Jacobian matrix to be the $(n \times k)$ matrix whose i^{th} row is the (transpose of the) gradient vector of the i^{th} row of $\boldsymbol{g}(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$. We denote the Jacobian matrix by $\mathbb{J}\left[\boldsymbol{g}(\boldsymbol{\beta})\right]$. We have:

$$\mathbb{J}\left[\boldsymbol{g}(\boldsymbol{\beta})\right] \overset{\text{def}}{=} \begin{bmatrix} \mathbb{G}\left[\boldsymbol{g}_1(\boldsymbol{\beta})\right]^\top \\ \mathbb{G}\left[\boldsymbol{g}_2(\boldsymbol{\beta})\right]^\top \\ \vdots \\ \mathbb{G}\left[\boldsymbol{g}_n(\boldsymbol{\beta})\right]^\top \end{bmatrix} \tag{1.10}$$

Proposition 7. Let $\boldsymbol{\beta}$ be an $(k \times 1)$ vector of parameters and let $\boldsymbol{g}(\boldsymbol{\beta})$ be an $(n \times 1)$ vector function of $\boldsymbol{\beta}$. Let $\boldsymbol{A}$ be a square $(n \times n)$, symmetric matrix (not depending on $\boldsymbol{\beta}$) and define the quadratic form $q_{\boldsymbol{A}}(\boldsymbol{\beta}) = \boldsymbol{g}(\boldsymbol{\beta})^\top \boldsymbol{A}\boldsymbol{g}(\boldsymbol{\beta})$.

Then, the gradient vector of $q_{\boldsymbol{A}}(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$ is given by:

$$\mathbb{G}\left[q_{\boldsymbol{A}}(\boldsymbol{\beta})\right] \overset{\text{def}}{=} 2\mathbb{J}\left[\boldsymbol{g}(\boldsymbol{\beta})\right]^{\top} \boldsymbol{A}\boldsymbol{g}(\boldsymbol{\beta}) \tag{1.11}$$

Proposition 8. Let $\boldsymbol{\beta}$ be an $(k \times 1)$ vector of parameters and let $\boldsymbol{g}(\boldsymbol{\beta})$ be an $(n \times 1)$ vector function of $\boldsymbol{\beta}$. Let $\boldsymbol{A}$ be a square $(n \times n)$, symmetric matrix (not depending on $\boldsymbol{\beta}$) and define the quadratic form $q_{\boldsymbol{A}}(\boldsymbol{\beta}) = \boldsymbol{g}(\boldsymbol{\beta})^{\top}\boldsymbol{A}\boldsymbol{g}(\boldsymbol{\beta})$. Assume that the Jacobian $\mathbb{J}\left[\boldsymbol{g}(\boldsymbol{\beta})\right] = \mathbb{J}$, so that it does not depend on $\boldsymbol{\beta}$. Then, the Hessian matrix of $q_{\boldsymbol{A}}(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$ is given by:

$$\mathbb{H}\left[q_{\boldsymbol{A}}(\boldsymbol{\beta})\right] \overset{\text{def}}{=} 2\mathbb{J}^{\top}\boldsymbol{A}\mathbb{J} \tag{1.12}$$

1.3 Exercises

1. Let $\boldsymbol{X}$ be an $(n \times k)$ matrix, with $n \geq k$. Consider the projection matrix $\boldsymbol{P_X} = \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'$. Show that the following are true: (a) $\boldsymbol{P_X}^{\top} = \boldsymbol{P_X}$, (b) $\boldsymbol{P_X}^{\top}\boldsymbol{P_X} = \boldsymbol{P_X}\boldsymbol{P_X}^{\top} = \boldsymbol{P_X}$.

2. Let $\boldsymbol{X}$ be an $(n \times k)$ matrix, with $n \geq k$. Consider the projection matrix $\boldsymbol{P_X}$ of the previous exercise and define the matrix $\boldsymbol{M_X} \overset{\text{def}}{=} \boldsymbol{I} - \boldsymbol{P_X}$. Show that the following are true: (a) $\boldsymbol{M_X}\boldsymbol{X} = \boldsymbol{0}$, (b) $\boldsymbol{M_X}\boldsymbol{P_X} = \boldsymbol{0}$, (c) $\boldsymbol{M_X}^{\top} = \boldsymbol{M_X}$, and (d) $\boldsymbol{M_X}^{\top}\boldsymbol{M_X} = \boldsymbol{M_X}\boldsymbol{P_X}^{\top} = \boldsymbol{M_X}$.

3. For the demeaning matrix $\boldsymbol{D}$ is it true that $\boldsymbol{D}\boldsymbol{e} = \boldsymbol{0}$?

Chapter 2

Review of Probability and Statistics

2.1 Notational Conventions

In this section we outline some notational conventions used throughout the text for denoting samples, expectations, variances, probability densities, estimators etc. We follow these conventions consistently and any deviations from them are clearly indicated.

1. A random sample of size n or T for a $(k \times 1)$ random vector $\boldsymbol{x}$ will be denoted by $\{\boldsymbol{x}_i\}_{i=1}^{n}$ or $\{\boldsymbol{x}_t\}_{t=1}^{T}$.

2. The expectation operator for any $(n \times 1)$ random vector $\boldsymbol{x}$ will be denoted by $\mathsf{E}\left[\boldsymbol{x}\right]$.

3. The variance operator for any random variable x is to be denoted by $\mathsf{Var}\left[x\right]$ while the variance-covariance operator for any $(n \times 1)$ random

vector $\boldsymbol{x}$ is to be denoted by $\mathsf{Cov}\left[\boldsymbol{x}\right]$.

4. The conditional expectation operator for any $(n \times 1)$ random vector $\boldsymbol{x}$, conditional on some collection (σ-algebra) of random vectors $\mathcal{G}$ is to be denoted by $\mathsf{E}\left[\boldsymbol{x}|\mathcal{G}\right]$.

5. The probability density function (pdf) of an arbitrary random variable x, possibly depending on a vector of unknown parameters $\boldsymbol{\theta}$, will be denoted by $f(x; \boldsymbol{\theta})$ and its corresponding cumulative density function (cdf) will be denoted by $F(x; \boldsymbol{\theta})$.

6. The joint density (or Likelihood Function [LF]) of an arbitrary $(n \times 1)$ random vector $\boldsymbol{x}$, possibly depending on vector of unknown parameters $\boldsymbol{\theta}$, is to be denoted by $\mathcal{L}_n(\boldsymbol{x}; \boldsymbol{\theta})$. The logarithm of the LF will be denoted by $\ell_n(\boldsymbol{x}; \boldsymbol{\theta})$.

7. An estimator for a $(k \times 1)$ vector of unknown parameters $\boldsymbol{\theta}$ will be denoted either by $\widehat{\boldsymbol{\theta}}$ or $\tilde{\boldsymbol{\theta}}$. If we want to emphasize the fact that the estimator was obtained from a sample of size, say, n we then write $\widehat{\boldsymbol{\theta}}_n$ or $\tilde{\boldsymbol{\theta}}_n$.

2.2 Basic Results

A number of useful, and repeatedly used, results are summarized in this section. We mainly use definitions, propositions and examples to set forth these results. The reader is assumed thouroughly familiar with basic statistics for one random variable as well as with the univariate normal distribution.

Definition 25. Let Ω denote the collection of all possible outcomes of an experiment of chance (e.g. tossing a coin). We call Ω the sample space and we call the individual outcomes in Ω, say $\omega \in \Omega$ the elementary events. In the coin tossing example we have that $\Omega = \{\text{Heads}, \ \text{Tails}\}$.

Definition 26. A random variable $x(\omega)$ is a (random) function defined on a sample space Ω and taking values on a subset $\mathcal{R}$ of the real line $\mathcal{R} \subseteq \mathbb{R}$. We write $x(\omega) : \Omega \rightarrow \mathcal{R} \subseteq \mathbb{R}$. We usually write x instead of $x(\omega)$, the dependence on some sample space being assumed implicitly. If the elements of $\mathcal{R}$ are countable then we say that x is a discrete random variable; if the elements of $\mathcal{R}$ are uncountable then we say that x is a continuous random variable. In our discussion we will focus almost exclusively in continuous random variables.

Definition 27. A $(k \times 1)$ random vector $\boldsymbol{x}(\omega)$ is a correspondence defined on a sample space Ω and taking values on a subset of the k-dimensional Euclidean space $\mathcal{R} \subseteq \mathbb{R}^k$. We write $\boldsymbol{x}(\omega) : \Omega \rightarrow \mathcal{R} \subseteq \mathbb{R}^k$. We usually write $\boldsymbol{x}$ instead of $\boldsymbol{x}(\omega)$, the dependence on some sample space being assumed implicitly. Equivalently, $\boldsymbol{x}$ can be thought of as a collection of k random variables.

Definition 28. The probability density function (pdf) of a continuous random variable x, denoted by $f(x; \boldsymbol{\theta})$ with $\boldsymbol{\theta}$ being a $(k \times 1)$ vector of unknown parameters, is a function defined on the range of values $\mathcal{R}$ of the random variable and taking values in the unit interval $[0, 1]$. We write

$f(x;\boldsymbol{\theta}) : \mathcal{R} \rightarrow [0,1]$. The pdf has the following properties:

$$\begin{aligned} & f(x;\boldsymbol{\theta}) \geq 0 && \forall x \in \mathcal{R} \\ & \textstyle\int_{\mathcal{R}} f(x;\boldsymbol{\theta})dx = 1 && \\ & \mathsf{Prob}(a \leq x \leq b) \overset{\text{def}}{=} \int_a^b f(x;\boldsymbol{\theta})dx && \forall a,b \in \mathcal{R} \end{aligned} \tag{2.1}$$

Definition 29. The cumulative density function (cdf) of a continuous random variable x, denoted by $F(x;\boldsymbol{\theta})$ with $\boldsymbol{\theta}$ being a $(k \times 1)$ vector of unknown parameters, is a function defined on the range of values $\mathcal{R}$ of the random variable and taking values in the unit interval $[0,1]$. We write $F(x;\boldsymbol{\theta}) : \mathcal{R} \rightarrow [0,1]$. The cdf is defined as:

$$F(z;\boldsymbol{\theta}) \overset{\text{def}}{=} \int_{x \in \mathcal{R} \,:\, x \leq z} f(x;\boldsymbol{\theta})dx = \mathsf{Prob}(x \leq z) \tag{2.2}$$

Definition 30. The p^{th} quantile of the values of a random variable x, with cdf $F(x;\boldsymbol{\theta})$, is defined as that value x_p that satisfies $F(x_p;\boldsymbol{\theta}) \overset{\text{def}}{=} p$, for some probability $p \in [0,1]$. The 50% quantile is called the *median* of the values of the random variable. For every symmetric distribution we have that $x_{0.5} = \mathsf{E}[x]$, i.e. the mean equals the median.

Definition 31. The mean (or expected value) of a continuous random variable x, denoted by $\mathsf{E}[x;\boldsymbol{\theta}]$ with $\boldsymbol{\theta}$ being a $(k \times 1)$ vector of unknown parameters, is a constant (perhaps a function of the elements of $\boldsymbol{\theta}$) defined as $\mathsf{E}[x;\boldsymbol{\theta}] \overset{\text{def}}{=} \int_{\mathcal{R}} xf(x;\boldsymbol{\theta})dx$. Thus, the expected value of a random variable is a weighted average of all the values of the random variable with the weights being given by the pdf of x. The mean measures the central tendency of the values of x and is considered a measure of location for the underlying pdf.

Definition 32. The j^{th} moment of a random variable x $(j = 1, 2, 3, \dots)$, denoted by $\mathsf{M}_j\left[x; \boldsymbol{\theta}\right]$ with $\boldsymbol{\theta}$ being a $(k \times 1)$ vector of unknown parameters, is defined as the expected value of the x^j, that is $\mathsf{M}_j\left[x; \boldsymbol{\theta}\right] \overset{\text{def}}{=} \mathsf{E}\left[x^j; \boldsymbol{\theta}\right] = \int_{\mathcal{R}} x^j f(x; \boldsymbol{\theta}) dx$. Thus, the mean of a random variable is the first moment.

Definition 33. Let x be a continuous random variable such that its mean exists. Define the continuous random variable $y \overset{\text{def}}{=} x - \mathsf{E}\left[x; \boldsymbol{\theta}\right]$. Then, the j^{th} moment of x *around the mean* $(j = 1, 2, 3, \dots)$, denoted by $\mathsf{D}_j\left[x; \boldsymbol{\theta}\right]$ with $\boldsymbol{\theta}$ being a $(k \times 1)$ vector of unknown parameters, is defined as the expected value of the y^j, that is $\mathsf{D}_j\left[x; \boldsymbol{\theta}\right] \overset{\text{def}}{=} \mathsf{E}\left[y^j; \boldsymbol{\theta}\right] = \int_{\mathcal{R}} y^j f(x; \boldsymbol{\theta}) dx$. The second moment around the mean is called the variance of x and its denoted by $\mathsf{Var}\left[x; \boldsymbol{\theta}\right] \overset{\text{def}}{=} \mathsf{D}_2\left[x; \boldsymbol{\theta}\right]$. The variance measures the dispersion of values of x around their mean and, consequently, is considered a measure of dispersion of the underlying pdf. The square root of the variance is called the standard deviation.

Definition 34. Let x be a continuous random variable with mean $\mu \overset{\text{def}}{=} \mathsf{E}\left[x; \boldsymbol{\theta}\right]$ and variance $\sigma^2 \overset{\text{def}}{=} \mathsf{Var}\left[x; \boldsymbol{\theta}\right]$, where both μ and σ^2 are elements (or functions of the elements) of $\boldsymbol{\theta}$. Then, the skewness coefficient of x is defined as $\mathcal{S} \overset{\text{def}}{=} \mathsf{D}_3\left[x; \boldsymbol{\theta}\right]/\sigma^3$ and the kurtosis coefficient of x is defined as $\mathcal{K} \overset{\text{def}}{=} \mathsf{D}_4\left[x; \boldsymbol{\theta}\right]/\sigma^4$. The skewness coefficient measures the symmetry of the underlying pdf around μ and equals zero if the pdf is symmetric. The kurtosis coefficient measures the thickness of the tails of the pdf and equals 3 for the standard normal distribution. We usually define $\mathcal{K} - 3$ to be the *degree of excees* kurtosis, with respect to the standard normal distribution.

Definition 35. Let y and x be two continuous random variables with joint pdf given by $f(y,x;\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is a $(k \times 1)$ vector of unknown parameters, and let $h(x;\boldsymbol{\theta}) = \int_{\mathcal{R}} f(y,x;\boldsymbol{\theta})dy$ be the marginal pdf of x. Then, the conditional pdf of y given $x = x^*$, denoted by $f(y|x^*;\boldsymbol{\theta})$, is given by:

$$f(y|x^*;\boldsymbol{\theta}) \overset{\text{def}}{=} \frac{f(y,x^*;\boldsymbol{\theta})}{h(x^*;\boldsymbol{\theta})} \tag{2.3}$$

The conditional mean or conditional expectation of y given $x = x^*$, denoted by $\mathsf{E}\left[y|x^*;\boldsymbol{\theta}\right]$, is given by:

$$\mathsf{E}\left[y|x^*;\boldsymbol{\theta}\right] \overset{\text{def}}{=} \int_{\mathcal{R}} y f(y|x^*;\boldsymbol{\theta})dy \tag{2.4}$$

Note that since x changes values in repeated sampling, according to $h(x;\boldsymbol{\theta})$, we must have that the conditional expectation is a random variable with pdf given by $h(x;\boldsymbol{\theta})$ (since a random variable is a function, the conditional expectation is also called the *regression function*). If we write the conditional expectation without indicating a particular value for x, like $\mathsf{E}\left[y|x^*;\boldsymbol{\theta}\right]$, we have the conditional expectation taking different values given a different values of x. From this observation we obtain the important law of iterated expectations, which we state in the following proposition. [NOTE: this definition remains essentially unchanged if instead of a random variable x we have a random vector $\boldsymbol{x}$.]

Proposition 10. Let y and x be two continuous random variables with joint pdf given by $f(y,x;\boldsymbol{\theta})$ where $\boldsymbol{\theta}$ is a $(k \times 1)$ vector of unknown parameters. Denote by $\mathcal{G}_x$ the collection of all possible combinations of values of x (σ-algebra) and let $h(x;\boldsymbol{\theta})$ be the marginal pdf of x. Also, let $\mathsf{E}\left[y|x;\boldsymbol{\theta}\right]$ denote the conditional expectation of y given x. Then, the unconditional expectation of y, $\mathsf{E}\left[y;\boldsymbol{\theta}\right]$, is given by:

$$\mathsf{E}\left[y;\boldsymbol{\theta}\right] \overset{\text{def}}{=} \mathsf{E}\left[\mathsf{E}\left[y|x;\boldsymbol{\theta}\right]|\mathcal{G}_x\right] = \int_{\mathcal{R}} \mathsf{E}(y|x;\boldsymbol{\theta})h(x;\boldsymbol{\theta})dx \tag{2.5}$$

[NOTE: this proposition remains essentially unchanged if instead of a random variable x we have a random vector $\boldsymbol{x}$.]

Definition 35. Let y and x be two continuous random variable with means μ_y and μ_x and variances σ_y^2 and σ_x^2 respectively. Then, their covariance is defined as the expected value of the first cross-moment around the means $\sigma_{yx} \overset{\text{def}}{=} \mathsf{Cov}\left[y,x\right] = \mathsf{E}\left[(y-\mu_y)(x-\mu_x)\right]$. Their correlation is defined as the ratio of the covariance to the product of the standard deviations $\rho_{yx} \overset{\text{def}}{=} \mathsf{Corr}\left[y,x\right] = \sigma_{yx}/(\sigma_y \cdot \sigma_x)$.

Definition 36. Let $\boldsymbol{x}$ be an $(n \times 1)$ random vector with elements x_i that are independent and identically distributed random variabels drawn from some underlying pdf $f(x_i;\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is a $(k \times 1)$ vector of unknown parameters. Then, the joint pdf of all n elements of the random vector is called the likelihood function (LF) and equals the product of the individual pdf. We have:

$$\mathcal{L}_n(\boldsymbol{x};\boldsymbol{\theta}) \overset{\text{def}}{=} \prod_{i=1}^{n} f(x_i;\boldsymbol{\theta}) \tag{2.6}$$

Definition 37. Let $\boldsymbol{x}$ be an $(n \times 1)$ random vector whose individual elements have means $\mu_i \overset{\text{def}}{=} \mathsf{E}\left[x_i;\boldsymbol{\theta}\right]$ with $\boldsymbol{\theta}$ being a $(k \times 1)$ vector of unknown parameters and $i = 1, 2, \ldots, n$. Then, the mean vector $\boldsymbol{\mu}$ is defined as:

$$\boldsymbol{\mu} \overset{\text{def}}{=} \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{bmatrix} \tag{2.7}$$

Definition 38. Let $\boldsymbol{x}$ be an $(n \times 1)$ random vector with mean $\boldsymbol{\mu} \overset{\text{def}}{=} \mathsf{E}\left[\boldsymbol{x}\right]$. Then, its $(n \times n)$ variance-covariance matrix $\boldsymbol{\Sigma}$ is given by:

$$\boldsymbol{\Sigma} \overset{\text{def}}{=} \mathsf{Cov}\left[\boldsymbol{x}\right] = \mathsf{E}\left[\left(\boldsymbol{x} - \boldsymbol{\mu}\right)\left(\boldsymbol{x} - \boldsymbol{\mu}\right)^{\top}\right] \tag{2.8}$$

The variances of the individual elements of $\boldsymbol{x}$ are given in the diagonal of $\boldsymbol{\Sigma}$, that is $\sigma_{ii} \overset{\text{def}}{=} \mathsf{Var}\left[x_i\right]$. The pairwise covariances are given by the off-diagonal elements of $\boldsymbol{\Sigma}$, that is $\sigma_{ij} \overset{\text{def}}{=} \mathsf{Cov}\left[x_i, x_j\right]$, for $i \neq j$ and for $i, j = 1, 2, \ldots, n$.

Proposition 11. Let $\boldsymbol{x}$ be an $(n \times 1)$ random vector with elements x_i that are independent and identically distributed random variables drawn from a standard normal distribution, i.e. $x_i \sim \mathcal{N}\left[0, 1\right]$ for all $i = 1, 2, ..., n$. Remember that the standard normal pdf is given by $f(x_i) = 1/\sqrt{2\pi}\exp(-x_i^2/2)$. We then say that $\boldsymbol{x}$ follows a multivariate standard normal distribution with mean vector $\mathsf{E}\left[\boldsymbol{x}\right] = \mathbf{0}$ and variance-covariance matrix $\mathsf{Cov}\left[\boldsymbol{x}\right] = \boldsymbol{I}$, and we write $\boldsymbol{x} \sim \mathcal{N}\left[\mathbf{0}, \boldsymbol{I}\right]$. The LF of $\boldsymbol{x}$ can then be obtained using the previous definition as:

$$\begin{aligned} \mathcal{L}_n(\boldsymbol{x}) &\overset{\text{def}}{=} \prod_{i=1}^{n} \tfrac{1}{\sqrt{2\pi}} \exp(-x_i^2/2) \\ &= \left(\tfrac{1}{\sqrt{2\pi}}\right)^n \exp\left(-\tfrac{1}{2}\sum_{i=1}^{n} x_i^2\right) \\ &= \left(\tfrac{1}{\sqrt{2\pi}}\right)^n \exp\left(-\tfrac{1}{2}\|\boldsymbol{x}\|^2\right) \end{aligned} \tag{2.9}$$

The logarithm of the above LF is given by:

$$\ell_n(\boldsymbol{x}) \overset{\text{def}}{=} -\frac{n}{2}\ln(2\pi) - \frac{1}{2}\|\boldsymbol{x}\|^2 \tag{2.10}$$

Proposition 12. Let $\boldsymbol{x}$ be an $(n \times 1)$ random vector that follows a multivariate standard normal distribution, $\boldsymbol{x} \sim \mathcal{N}[\mathbf{0}, \boldsymbol{I}]$. Let $\boldsymbol{\mu}$ be an $(n \times 1)$ vector of constants and let $\boldsymbol{A}$ be a positive definite, $(n \times n)$ matrix of constants. Define the $(n \times 1)$ random vector $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{\mu}$. Then, $\boldsymbol{y}$ follows a multivariate normal distribution with mean vector $\mathsf{E}[\boldsymbol{y}] = \boldsymbol{\mu}$ and variance-covariance matrix $\mathsf{Cov}[\boldsymbol{y}] = \boldsymbol{A}\boldsymbol{A}^\top \overset{\text{def}}{=} \boldsymbol{\Sigma}$, and we write $\boldsymbol{y} \sim \mathcal{N}[\boldsymbol{\mu}, \boldsymbol{\Sigma}]$. Note that $\boldsymbol{\Sigma}$ is a symmetric matrix.

To show these first note that $\mathsf{E}[\boldsymbol{y}] = \boldsymbol{A}\mathsf{E}[\boldsymbol{x}] + \boldsymbol{\mu} = \boldsymbol{\mu}$, since $\mathsf{E}[\boldsymbol{x}] = \mathbf{0}$. Then, note that $\mathsf{Cov}[\boldsymbol{y}] = \mathsf{E}\left[(\boldsymbol{y} - \boldsymbol{\mu})(\boldsymbol{y} - \boldsymbol{\mu})^\top\right] = \boldsymbol{A}\mathsf{E}\left[\boldsymbol{x}\boldsymbol{x}^\top\right]\boldsymbol{A}^\top = \boldsymbol{A}\boldsymbol{A}^\top$, since $\mathsf{Cov}[\boldsymbol{x}] = \boldsymbol{I}$.

Definition 39. Let $\boldsymbol{x}$ be an $(n \times 1)$ random vector that follows a multivariate normal distribution, $\boldsymbol{x} \sim \mathcal{N}[\boldsymbol{\mu}, \boldsymbol{\Sigma}]$. The logarithm of the LF of the elements of $\boldsymbol{x}$ is given by:

$$\ell_n(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = -\frac{n}{2}\ln(2\pi) - \frac{1}{2}\ln|\boldsymbol{\Sigma}| - \frac{1}{2}(\boldsymbol{y} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{y} - \boldsymbol{\mu}) \tag{2.11}$$

Proposition 13. Let $\boldsymbol{x}$ be an $(n \times 1)$ random vector that follows a multivariate normal distribution $\boldsymbol{x} \sim \mathcal{N}[\boldsymbol{\mu}, \boldsymbol{\Sigma}]$. Let $\boldsymbol{A}$ be an arbitrary $(m \times n)$ matrix of constants and define the $(m \times 1)$ random vector $\boldsymbol{y} \overset{\text{def}}{=} \boldsymbol{A}\boldsymbol{x}$. Then $\boldsymbol{y}$ is also normally distributed as $\boldsymbol{y} \sim \mathcal{N}\left[\boldsymbol{A}\boldsymbol{\mu}, \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^\top\right]$.

Proposition 14. Let $\boldsymbol{y}$ be an $(n \times 1)$ random vector that follows a multivariate normal distribution $\boldsymbol{y} \sim \mathcal{N}[\boldsymbol{\mu}, \boldsymbol{\Sigma}]$. Define the vector $\boldsymbol{x} \overset{\text{def}}{=} \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{y} - \boldsymbol{\mu})$. Then, $\boldsymbol{x}$ follows a multivariate standard normal distribution $\boldsymbol{x} \sim \mathcal{N}[\mathbf{0}, \boldsymbol{I}]$.

Proposition 15. Let $\boldsymbol{z}$ be an $(n \times 1)$ random vector that follows a multivariate normal distribution $\boldsymbol{z} \sim \mathcal{N}[\boldsymbol{\mu}, \boldsymbol{\Sigma}]$. Split $\boldsymbol{z}$ into two components, a scalar component $y \overset{\text{def}}{=} z_1$ and an $(n-1 \times 1)$ component $\boldsymbol{x} \overset{\text{def}}{=} [z_2, z_3, \ldots, z_{n-1}]^\top$. Partition $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ conformably as:

$$\boldsymbol{\mu} \overset{\text{def}}{=} \begin{bmatrix} \mu_y \\ \boldsymbol{\mu}_x \end{bmatrix}, \boldsymbol{\Sigma} \overset{\text{def}}{=} \begin{bmatrix} \sigma_y^2 & \boldsymbol{\sigma}_{xy}^\top \\ \boldsymbol{\sigma}_{xy} & \boldsymbol{\Sigma}_{xx} \end{bmatrix} \tag{2.12}$$

and note that $\boldsymbol{\sigma}_{xy}$ is $(n-1 \times 1)$ and $\boldsymbol{\Sigma}_{xx}$ is $(n-1 \times n-1)$. Define the regression coefficient vector $\boldsymbol{\beta} \overset{\text{def}}{=} \boldsymbol{\Sigma}_{xx}^{-1}\boldsymbol{\sigma}_{xy}$ (note that $\boldsymbol{\beta}$ is a constant vector [i.e. not random] since it depends exclusively on the variance-covaraince parameters). Then, the conditional distribution of y given $\boldsymbol{x}$ is (univariate) normal with conditional mean $\mathsf{E}[y|\boldsymbol{x}; \mu_y, \boldsymbol{\mu}_x, \boldsymbol{\beta}] \overset{\text{def}}{=} \mu_y + (\boldsymbol{x} - \boldsymbol{\mu}_x)^\top \boldsymbol{\beta}$ and conditional variance $\mathsf{Var}\left[y|\boldsymbol{x}; \sigma_y^2, \boldsymbol{\sigma}_{xy}, \boldsymbol{\beta}\right] \overset{\text{def}}{=} \sigma_y^2 - \boldsymbol{\sigma}_{xy}^\top \boldsymbol{\beta}$.

This proposition allows us to express any element of a normal random vector as a linear function of the remaining elements. This is so since we can always define a random variable $u \overset{\text{def}}{=} y - \mathsf{E}[y|\boldsymbol{x}; \mu_y, \boldsymbol{\mu}_x, \boldsymbol{\beta}]$, with conditional mean zero, so that we can express y as a linear regression of the form:

$$y = \mu_y + (\boldsymbol{x} - \boldsymbol{\mu}_x)^\top \boldsymbol{\beta} + u \tag{2.13}$$

Definition 40. Let $\boldsymbol{x}$ be an $(n \times 1)$ random vector that follows a multivariate standard normal distribution. Then, $\chi \overset{\text{def}}{=} \langle \boldsymbol{x}, \boldsymbol{x} \rangle = \|\boldsymbol{x}\|^2 \sim \chi^2_{(n)}$, that is, the scalar random variable χ (that equals the sum of squares of the x_i's) follows a chi-squared distribution with n degrees of freedom.

Proposition 16. Let $\boldsymbol{y}$ be an $(n \times 1)$ random vector that follows a multivariate normal distribution $\boldsymbol{y} \sim \mathcal{N}[\boldsymbol{\mu}, \boldsymbol{\Sigma}]$. Define $\boldsymbol{x} \overset{\text{def}}{=} \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{y} - \boldsymbol{\mu}) \sim \mathcal{N}[\boldsymbol{0}, \boldsymbol{I}]$. Then, the scalar random variable $\chi \overset{\text{def}}{=} \langle \boldsymbol{x}, \boldsymbol{x} \rangle = \|\boldsymbol{x}\|^2 \sim \chi^2_{(n)}$.

Definition 41. Let Let $\widehat{\boldsymbol{\theta}}_n$ be an estimator of the $(k \times 1)$ vector of unknown parameters $\boldsymbol{\theta}$. We define the mean-squared error (MSE) matrix of the estimator as $\mathsf{MSE}\left[\widehat{\boldsymbol{\theta}}_n\right] \overset{\text{def}}{=} \mathsf{E}\left[\left(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}\right)\left(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}\right)^\top\right]$.

Definition 42. Let $\widehat{\boldsymbol{\theta}}_n$ be an estimator of the $(k \times 1)$ vector of unknown parameters $\boldsymbol{\theta}$. We say that $\widehat{\boldsymbol{\theta}}_n$ is an unbiased estimator if $\mathsf{E}\left[\widehat{\boldsymbol{\theta}}_n\right] = \boldsymbol{\theta}$. If the estimator is not unbiased we define $\boldsymbol{b}\left[\widehat{\boldsymbol{\theta}}_n\right] \overset{\text{def}}{=} \mathsf{E}\left[\widehat{\boldsymbol{\theta}}_n\right] - \boldsymbol{\theta}$ to be the bias of the estimator.

Proposition 17. Let Let $\widehat{\boldsymbol{\theta}}_n$ be an estimator of the $(k \times 1)$ vector of unknown parameters $\boldsymbol{\theta}$. It can be shown that the MSE matrix of $\widehat{\boldsymbol{\theta}}_n$ is equal to $\mathsf{MSE}\left[\widehat{\boldsymbol{\theta}}_n\right] \overset{\text{def}}{=} \mathsf{Cov}\left[\widehat{\boldsymbol{\theta}}_n\right] + \boldsymbol{b}\left[\widehat{\boldsymbol{\theta}}_n\right]\boldsymbol{b}\left[\widehat{\boldsymbol{\theta}}_n\right]^\top$.

Proposition 18. Let Let $\widehat{\boldsymbol{\theta}}_n$ be an estimator of the $(k \times 1)$ vector of unknown parameters $\boldsymbol{\theta}$. We say that $\widehat{\boldsymbol{\theta}}_n$ converges in quadratic mean to $\boldsymbol{\theta}$ if $\lim_{n\to\infty} \mathsf{MSE}\left[\widehat{\boldsymbol{\theta}}_n\right] = \mathbf{0}$.

Definition 43. Let Let $\widehat{\boldsymbol{\theta}}_n$ be an estimator of the $(k \times 1)$ vector of unknown parameters $\boldsymbol{\theta}$. We say that $\widehat{\boldsymbol{\theta}}_n$ is a consistent estimator for $\boldsymbol{\theta}$ if $\widehat{\boldsymbol{\theta}}_n$ converges in quadratic mean. Note that, from the definition of the MSE matrix, a necessary and sufficient condition for consistency for an unbiased estimator is that $\lim_{n\to\infty} \mathsf{Cov}\left[\widehat{\boldsymbol{\theta}}_n\right] = \mathbf{0}$.

Definition 44. Let Let $\widehat{\boldsymbol{\theta}}_n$ and $\tilde{\boldsymbol{\theta}}_n$ be two unbiased estimators of the $(k \times 1)$ vector of unknown parameters $\boldsymbol{\theta}$. We say that $\widehat{\boldsymbol{\theta}}_n$ is relatively efficient with respect to $\tilde{\boldsymbol{\theta}}_n$ if the matrix $\boldsymbol{C} \overset{\text{def}}{=} \mathsf{Cov}\left[\tilde{\boldsymbol{\theta}}_n\right] - \mathsf{Cov}\left[\widehat{\boldsymbol{\theta}}_n\right]$ is positive semi-definite, i.e. $\boldsymbol{C} \geq 0$.

Definition 45. Let $\{\boldsymbol{x}_i\}_{i=1}^n$ be a random sample of size n for the $(k \times 1)$

random vector $\boldsymbol{x}$, drawn from some underlying joint pdf with common mean vector $\mathsf{E}\left[\boldsymbol{x}_i\right] = \boldsymbol{\mu}$ and common variance-covariance matrix $\mathsf{Cov}\left[\boldsymbol{x_i}\right] = \boldsymbol{\Sigma}$. The sample mean vector $\widehat{\boldsymbol{\mu}}_n$ is defined as $\widehat{\boldsymbol{\mu}}_n = n^{-1} \sum\limits_{i=1}^{n} \boldsymbol{x}_i$. The sample variance-covariance matrix $\widehat{\boldsymbol{\Sigma}}_n$ is defined as $\widehat{\boldsymbol{\Sigma}}_n = n^{-1} \sum\limits_{i=1}^{n} \left(\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}}_n\right) \left(\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}}_n\right)^{\top}$. Let $\widehat{\boldsymbol{V}}_n$ be a diagonal matrix having the sample variances on its main diagonal (i.e. the elements $\widehat{\sigma}_{ii}$ of the diagonal of $\widehat{\boldsymbol{\Sigma}}_n$). Then, the sample correlation matrix is defined as $\widehat{\boldsymbol{R}}_n = \widehat{\boldsymbol{V}}_n^{-1/2} \widehat{\boldsymbol{\Sigma}}_n \widehat{\boldsymbol{V}}_n^{-1/2}$.

Definition 46. Let $\{\boldsymbol{x}_i\}_{i=1}^{n}$ be a random sample of size n for the $(k \times 1)$ random vector $\boldsymbol{x}$, following a multivariate normal distribution $\boldsymbol{x}_i \sim \mathcal{N}\left[\boldsymbol{\mu}, \boldsymbol{\Sigma}\right]$. Then, the logarithm of the LF of the sample observations is given by:

$$\ell_n(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = -\frac{nk}{2} \ln(2\pi) - \frac{n}{2} |\boldsymbol{\Sigma}| - \frac{1}{2} \sum_{i=1}^{n} \left(\boldsymbol{x}_i - \boldsymbol{\mu}\right)^{\top} \boldsymbol{\Sigma}^{-1} \left(\boldsymbol{x}_i - \boldsymbol{\mu}\right) \quad (2.14)$$

Proposition 19. Let $\{\boldsymbol{x}_i\}_{i=1}^{n}$ be a random sample of size n for the random vector $\boldsymbol{x}$, drawn from a multivariate normal distribution $\boldsymbol{x}_i \sim \mathcal{N}\left[\boldsymbol{\mu}, \boldsymbol{\Sigma}\right]$. Then, the maximum likelihood (ML) estimators for $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are given by the sample mean $\widehat{\boldsymbol{\mu}}_n$ and the sample variance-covariance matrix $\widehat{\boldsymbol{\Sigma}}_n$, as defined before. Note that the ML estimators are obtained by maximizing the logarithm of the LF given in the previous equation.

Definition 47. Let $\{x_i\}_{i=1}^{n}$ be a random sample of size n for the random variable x, drawn from some underlying pdf with common mean $\mathsf{E}\left[x_i\right] \overset{\text{def}}{=} \mu$. Then, we define the j^{th} sample moments about the origin and about the mean as:

$$\begin{array}{rcl} \widehat{\mathsf{m}}_j & \overset{\text{def}}{=} & n^{-1} \sum_{i=1}^{n} x_i^j \\ \widehat{\mathsf{d}}_j & \overset{\text{def}}{=} & n^{-1} \sum_{i=1}^{n} (x_i - \widehat{\mathsf{m}}_1)^j \end{array} \quad (2.15)$$

Proposition 20. Let $\{x_i\}_{i=1}^n$ be a random sample of size n for the random variable x, drawn from some underlying normal distribution $x_i \sim \mathcal{N}\left[\mu, \sigma^2\right]$. Denote the sample standard deviation by $s \overset{\text{def}}{=} \sqrt{\widehat{\mathsf{d}}_2}$ and consider the sample skewness $\widehat{\mathcal{S}}_n \overset{\text{def}}{=} \widehat{\mathsf{d}}_3/s^3$ and sample kurtosis coefficients $\widehat{\mathcal{K}}_n \overset{\text{def}}{=} \widehat{\mathsf{d}}_4/s^4$. It can then be shown that the sampling distributions of $\widehat{\mathcal{S}}_n$ and $\widehat{\mathcal{K}}_n$ are given by:

$$\widehat{\mathcal{S}}_n \sim \mathcal{N}\left[0, \frac{6}{n}\right] \qquad \widehat{\mathcal{K}}_n \sim \mathcal{N}\left[3, \frac{24}{n}\right] \tag{2.16}$$

Proposition 21. Let $\{x_i\}_{i=1}^n$ be a random sample of size n for the random variable x, drawn from some underlying normal distribution $x_i \sim \mathcal{N}\left[\mu, \sigma^2\right]$. Denote the sample standard deviation by $s \overset{\text{def}}{=} \sqrt{\widehat{\mathsf{d}}_2}$, the sample median by $\tilde{x}_{0.5}$ and denote by $\widehat{Q}_1$ and $\widehat{Q}_3$ the sample quartiles (where $Q_1 \overset{\text{def}}{=} \Phi^{-1}(0.25)$ and $Q_3 \overset{\text{def}}{=} \Phi^{-1}(0.75)$ are the 25% and 75% quartiles of the normal distribution with cdf denoted by $\Phi(\cdot)$). Then, an estimator for the dispersion of the sample observations around their sample median is given by the ratio of the inter-quartile range $IQR \overset{\text{def}}{=} \widehat{Q}_3 - \widehat{Q}_1$ over 1.34. We denote this estimator by $d \overset{\text{def}}{=} IQR/1.34$.

2.3 Data Transformations to Near Normality

If the sample observations do not appear to be drawn from some underlying normal distribution one may try to appropriately transform them so that they near normality. A class of data transformations that is frequently used for this purpose is that of power transformations, also known as Box-Cox transformations. Assume that you have a random sample of n observations for the random variable x, that is $\{x_i\}_{i=1}^n$. For every sample observation

x_i consider the transformation to $y_i(\lambda)$, given the non-negative parameter $\lambda \geq 0$, as:

$$x_i \rightarrow y_i(\lambda) \overset{\text{def}}{=} \left\{ \begin{array}{l} (x_i^\lambda - 1)/\lambda \text{ for } x_i > 0 \text{ and } \lambda \neq 0 \\ \ln(x_i) \text{ for } x_i > 0 \text{ and } \lambda = 0 \end{array} \right\} \tag{2.17}$$

While the transformation parameter λ can be set to some arbitrary value, one that transforms the data so as to near normality, we can actually estimate λ from the sample observations using the method of maximum likelihood (ML). It can be shown that the log-likelihood function (the log of the joint pdf) of the transformed observations $y_i(\lambda)$ is given by:

$$\ell_n(\boldsymbol{x}_n; \lambda) \overset{\text{def}}{=} -\frac{1}{2} n \ln \left\{ \frac{1}{n} \sum_{i=1}^{n} [y_i(\lambda) - \bar{y}(\lambda)]^2 \right\} + (\lambda - 1) \sum_{i=1}^{n} \ln x_i \tag{2.18}$$

where $\boldsymbol{x}_n \overset{\text{def}}{=} \{x_i\}_{i=1}^n$ and $\bar{y}(\lambda)$ denotes the sample mean of the transformed observations. Maximizing $\ell_n(\boldsymbol{x}_n; \lambda)$ with respect to λ gives us the maximum likelihood estimator (MLE) $\widehat{\lambda}_n$. Note that the likelihood function is non-linear in λ and therefore we cannot obtain a closed-form expression for the MLE. However, we can employ a simple search procedure, over a grid of possible values for λ, and select that value that maximizes the likelihood function.[1]

A formal hypothesis test for the null hypothesis $H_0 : \lambda = 0$, corresponding to a logarithmic transformation of the data, can be based on the likelihood ratio principle. It can be shown that $2(\mathcal{L}_n^U - \mathcal{L}_n^R) \sim \chi^2_{(1)}$, where $\mathcal{L}_n^U \overset{\text{def}}{=} \ell_n(\boldsymbol{x}_n; \widehat{\lambda}_n)$ denotes the (unrestricted) maximum value of the log-likelihood

[1] This can be easily accomplished in a spreadsheet program and, of course, with any programming language.

function under the estimated $\widehat{\lambda}_n$ and $\mathcal{L}_n^R \stackrel{\text{def}}{=} \ell_n(\boldsymbol{x}_n; 0)$ denotes the (restricted) maximum value of the log-likelihood function under the null value of $\lambda = 0$.

2.4 Exercises

1. Propose estimators for the median $x_{0.5}$ and the quartiles $Q_1 \stackrel{\text{def}}{=} x_{0.25}$ and $Q_3 \stackrel{\text{def}}{=} x_{0.75}$.

2. Is the mean of a positively skewed distribution larger or smaller than its median?

3. Show that for a normal distribution $\mathcal{N}[\mu, \sigma^2]$ we have that following holds: $\sigma \approx IQR/1.34$.

4. Let $\{x_i\}_{i=1}^n$ be a sample of observations for the random variable x drawn from some underlying pdf $f(x_i; \boldsymbol{\theta})$ with common mean $\mathsf{E}[x_i] = \mu$ and common variance $\mathsf{Var}[x_i] = \sigma^2$. (a) Show that the univariate sample mean $\widehat{\mathsf{m}}_1$ is an unbiased and consistent estimator of the true mean μ; (b) Show that $\widehat{\mathsf{d}}_2$ is biased but that $[n/(n-1)]\widehat{\mathsf{d}}_2$ is an unbiased estimator for σ^2.

5. Let $\{\boldsymbol{x}_i\}_{i=1}^n$ be a random sample of size n for the $(k \times 1)$ random vector $\boldsymbol{x}$, drawn from some underlying joint pdf with common mean vector $\mathsf{E}[\boldsymbol{x}_i] = \boldsymbol{\mu}$ and common variance-covariance matrix $\mathsf{Cov}[\boldsymbol{x}_i] = \boldsymbol{\Sigma}$. Show that the multivariate sample mean $\widehat{\boldsymbol{\mu}}_n$ is an unbiased and consistent estimator of the true mean $\boldsymbol{\mu}$.

6. Construct t-type tests for zero skewness and zero excess kurtosis.