Linear Statistical Models
Math 6010-1; Fall 2016

Davar Khoshnevisan

155 South 1400 East JWB 233, Department of Mathematics, University of Utah, Salt Lake City UT 84112–0090

E-mail address: `davar@math.utah.edu`

URL: `http://www.math.utah.edu/~davar`

Contents

Lecture 1

Linear statistical models

1. Introduction

The goal of this course is, in rough terms, to predict a variable y, given that we have the opportunity to observe variables $x_1, \ldots, x_{p-1}$. This is a very important statistical problem. Therefore, let us spend a bit of time and examine a simple example:

Given the various vital statistics of a newborn baby, you wish to predict his height y at maturity. Examples of those "vital statistics" might be $x_1 :=$ present height, and $x_2 :=$ present weight. Or perhaps there are still more predictive variables $x_3 :=$ your height and $x_4 :=$ your spouse's height, etc.

In actual fact, a visit to your pediatrist might make it appear that this prediction problem is trivial. But that is not so [though it is nowadays fairly well understood in this particular case]. A reason the problem is nontrivial is that there is no *a priori* way to know "how" y depends on $x_1, \ldots, x_4$ [say, if we plan to use all 4 predictive variables]. In such a situation, one resorts to writing down a reasonable model for this dependence structure, and then analyzing that model. [Finally, there might be need for model verification as well.]

In this course, we study the general theory of "linear statistical models." That theory deals with the simplest possible nontrivial setting where such problems arises in various natural ways. Namely, in that theory we *posit* that y is a linear function of $(x_1, \ldots, x_4)$, possibly give or take some "noise." In other words, the theory of linear statistical models posits that there exist unknown parameters $\beta_0, \ldots, \beta_{p-1}$ [here, $p = 5$] such that

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_{p-1} x_{p-1} + \varepsilon, \tag{1}$$

where ε is a random variable. The problem is still not fully well defined [for instance, what should be the distribution of ε, etc.?]. But this is roughly the starting point of the theory of linear statistical models. And one begins asking natural questions such as, "how can we estimate $\beta_0, \ldots, \beta_4$?," or "can we perform inference for these parameters?" [for instance, can we test to see if y does not depend on x_1 in this model; i.e., test for $H_0 : \beta_1 = 0$?]. And so on.

We will also see, at some point, how the model can be used to improve itself. For instance, suppose we have only one predictive variable x, but believe to have a nonlinear dependence between y and x. Then we could begin by thinking about *polynomial regression*; i.e., a linear statistical model of the form

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_{p-1} x^{p-1} + \varepsilon.$$

Such a model fits in the general form (1) of linear statistical models, as well: We simply define new predictive variables $x_j := x^j$ for all $1 \le j < p$. One of the conclusions of this discussion is that we are studying models that are linear functions of unknown parameters $\beta_0, \ldots, \beta_{p-1}$ and not $x_1, \ldots, x_{p-1}$. This course studies statistical models with such properties. And as it turns out, not only these models are found in a great number of diverse applications, but also they have a rich mathematical structure.

2. The method of least squares

Suppose we have observed n data points in pairs: $(x_1 , y_1), \ldots (x_n , y_n)$. The basic problem here is, what is the best straight line that fits this data? There is of course no unique sensible answer, because "best" might mean different things.

We will use the *method of least squares*, introduced by C.-F. Gauss. Here is how the method works: If we used the line $y = \beta_0 + \beta_1 x$ to describe how the x_i's affect the y_i's, then the error of approximation, at $x = x_i$, is $y_i - (\beta_0 + \beta_1 x_i)$; this is called the ith *residual error*. The sum of the squared residual errors is $\text{SSE} := \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2$, and the method of least squares is to find the line with the smallest SSE. That is, we need to find the optimal β_0 and β_1—written $\hat\beta_0$ and $\hat\beta_1$—that solve the following optimization problem:

$$\min_{\beta_0, \beta_1} \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2 . \tag{2}$$

Theorem 1 (Gauss). *The least-squares solution to* (2) *is given by*

$$\hat\beta_1 := \frac{\sum_{i=1}^n (x_i - \bar x)(y_i - \bar y)}{\sum_{i=1}^n (x_i - \bar x)^2} \quad \textit{and} \quad \hat\beta_0 := \bar y - \hat\beta_1 \bar x.$$

Proof. Define

$$L(\beta_0, \beta_1) := \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2 .$$

Our goal is to minimize the function L. An inspection of the graph of L shows that L has a unique minimum; multivariable calculus then tells us that it suffices to set $\partial L/\partial\beta_j = 0$ for $j = 1, 2$ and solve. Because

$$\begin{aligned}\frac{\partial}{\partial\beta_0}L(\beta_0,\beta_1) &= -2\sum_{i=1}^{n}\left(y_i - \beta_0 - \beta_1 x_i\right),\\ \frac{\partial}{\partial\beta_1}L(\beta_0,\beta_1) &= -2\sum_{i=1}^{n}x_i\left(y_i - \beta_0 - \beta_1 x_i\right),\end{aligned}$$

a few lines of simple arithmetic finish the derivation. □

The preceding implies that, given the points $(x_1\,,y_1),\ldots,(x_n\,,y_n)$, the best line of fit through these points—in the sense of least squares—is

$$y = \hat\beta_0 + \hat\beta_1 x. \tag{3}$$

For all real numbers x and y, define x_{SU} and y_{SU} to be their respective "standardizations." That is,

$$x_{\text{SU}} := \frac{x-\bar x}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar x)^2}}, \qquad y_{\text{SU}} := \frac{y-\bar y}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar y)^2}}.$$

Then, (3) can be re-written in the following equivalent form:

$$\begin{aligned}y_{\text{SU}} &= \frac{\hat\beta_0-\bar y}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar y)^2}} + \frac{\hat\beta_1 x}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar y)^2}}\\ &= \frac{\hat\beta_1}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar y)^2}}\left(x-\bar x\right),\end{aligned}$$

the last line following from the identity $\hat\beta_0 = \bar y - \hat\beta_1\bar x$. We re-write the preceding again:

$$y_{\text{SU}} = \frac{\hat\beta_1\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar x)^2}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar y)^2}}\,x_{\text{SU}}.$$

Because

$$\hat\beta_1 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar x)(y_i-\bar y)}{\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar y)^2},$$

we can re-write the best line of fit, yet another time, this time as the following easy-to-remember formula:

$$y_{\text{SU}} = r x_{\text{SU}},$$

where r [a kind of "correlation coefficient"] is defined as

$$r := \frac{\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar y)(x_i-\bar x)}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar x)^2}\cdot\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar y)^2}}.$$

3. Simple linear regression

Suppose $Y_1, \ldots, Y_n$ are observations from a distribution, and they satisfy

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \qquad (1 \le i \le n), \tag{4}$$

where $\varepsilon_1, \ldots, \varepsilon_n$ are [unobserved] i.i.d. $\mathrm{N}(0\,,\sigma^2)$ for a fixed [possibly unknown] $\sigma > 0$. We assume that $x_1, \ldots, x_n$ are known, and seek to find the "best" β_0 and β_1.[1]

In other words, we believe that we are observing a certain linear function of the variable x at the x_i's, but our measurement [and/or modeling] contains noise sources [ε_i's] which we cannot observe.

Theorem 2 (Gauss)**.** *Suppose $\varepsilon_1, \ldots, \varepsilon_n$ are i.i.d. with common distribution $\mathrm{N}(0\,,\sigma^2)$, where $\sigma > 0$ is fixed. Then the maximum likelihood estimators of β_1 and β_0 are, respectively,*

$$\hat\beta_1 = \frac{\sum_{j=1}^n (x_j - \bar x)(Y_j - \bar Y)}{\sum_{j=1}^n (x_j - \bar x)^2} \quad \textit{and} \quad \hat\beta_0 = \bar Y - \hat\beta_1 \bar x.$$

Therefore, based on the data $(x_1\,,Y_1), \ldots, (x_n\,,Y_n)$, we predict the y-value at $x = x_*$ to be $y_* := \hat\beta_0 + \hat\beta_1 x_*$.

Proof. Note that $Y_1, \ldots, Y_n$ are independent [though not i.i.d.], and the distribution of Y_j is $\mathrm{N}(\beta_0 + \beta_1 x_j\,,\sigma^2)$. Therefore, the joint probability density function of $(Y_1\,, \ldots, Y_n)$ is

$$f(y_1\,, \ldots, y_n) := \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left[-\frac{1}{2\sigma^2} \sum_{j=1}^n (y_j - \beta_0 - \beta_1 x_j)^2 \right].$$

According to the MLE principle we should maximize $f(Y_1\,, \ldots, Y_n)$ over all choices of β_0 and β_0. But this is equivalent to minimizing $L(\beta_0\,,\beta_1) := \sum_{i=1}^n (Y_i - \beta_0 - \beta_1 x_i)^2$. But this was exactly what we did in Theorem 1 [except for real variables $y_1, \ldots, y_n$ in place of the random variables $Y_1, \ldots, Y_n$]. □

In this way we have the following "regression equation," which uses the observed data $(x_1\,,Y_1), \ldots, (x_n\,,Y_n)$ in order to predict a y-value corresponding to $x = x_*$:

$$Y(x) = \hat\beta_0 + \hat\beta_1 x.$$

But as we shall see, this method is good not only for prediction, but also for inference. Perhaps a first important question is, "do the y's depend linearly on the x's"? Mathematically speaking, we are asking to test the hypothesis that $\beta_1 = 0$. If we could compute the distribution of $\hat\beta_1$, then standard methods can be used to accomplish this. We will see later on how this can be accomplished.

[1] In actual applications, the x_i's are often random. In such a case, we assume that the model holds after conditioning on the x_i's.

But before we develop the theory of linear inference we need to know a few things about linear algebra and some of its probabilistic consequences.

Lecture 2

Random vectors

It will be extremely helpful to us if we worked directly with random vectors and not a group of individual random variables. Throughout, all vectors are written columnwise; and so are random ones. Thus, for instance, a random vector $\boldsymbol{X} \in \mathbf{R}^n$ is written columnwise as

$$\boldsymbol{X} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix} = (X_1\,, \ldots, X_n)'.$$

And even more generally, we might sometimes be interested in random matrices. For instance, a random $m \times n$ matrix is written coordinatewise as

$$\boldsymbol{X} = \begin{pmatrix} X_{1,1} & \cdots & X_{1,n} \\ \vdots & & \vdots \\ X_{m,1} & \cdots & X_{m,n} \end{pmatrix}.$$

1. Expectation

If $\boldsymbol{X}$ is a random $m \times n$ matrix, then we define its *expectation* in the most natural possible way as

$$\mathrm{E}\boldsymbol{X} := \begin{pmatrix} \mathrm{E}X_{1,1} & \cdots & \mathrm{E}X_{1,n} \\ \vdots & & \vdots \\ \mathrm{E}X_{m,1} & \cdots & \mathrm{E}X_{m,n} \end{pmatrix}.$$

Many of the properties of expectations continue to hold in the setting of random vectors and/or matrices. The following summarizes some of those properties.

Proposition 1. *Suppose* $\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{C}$, *and* $\boldsymbol{D}$ *are nonrandom matrices, and* $\boldsymbol{X}$ *and* $\boldsymbol{Y}$ *are random matrices. Then,*

$$\mathrm{E}\left(\boldsymbol{AXB}+\boldsymbol{CYD}\right)=\boldsymbol{A}\left(\mathrm{E}\boldsymbol{X}\right)\boldsymbol{B}+\boldsymbol{C}\left(\mathrm{E}\boldsymbol{Y}\right)\boldsymbol{D},$$

provided that the matrix dimensions are sensible.

Proof. Because

$$(\boldsymbol{AXB}+\boldsymbol{CYD})_{i,j}=(\boldsymbol{AXB})_{i,j}+(\boldsymbol{CYD})_{i,j}\,,$$

it suffices to prove that $\mathrm{E}(\boldsymbol{AXB})=\boldsymbol{A}\mathrm{E}(\boldsymbol{X})\boldsymbol{B}$. But then, we can work coordinatewise as follows:

$$\begin{aligned}\mathrm{E}\left[(\boldsymbol{AXB})_{i,j}\right]&=\mathrm{E}\sum_{k=1}^{n}\sum_{\ell=1}^{n}A_{i,k}X_{k,\ell}B_{\ell,j}=\sum_{k=1}^{n}\sum_{\ell=1}^{n}A_{i,k}\mathrm{E}(X_{k,\ell})B_{\ell,j}\\&=\sum_{k=1}^{n}\sum_{\ell=1}^{n}A_{i,k}\left[\mathrm{E}\boldsymbol{X}\right]_{k,\ell}B_{\ell,j}=\left[\boldsymbol{A}\mathrm{E}(\boldsymbol{X})\boldsymbol{B}\right]_{i,j}.\end{aligned}$$

That is, $\mathrm{E}(\boldsymbol{AXB})=\boldsymbol{A}\mathrm{E}(\boldsymbol{X})\boldsymbol{B}$ coordinatewise. This proves the result. □

2. Covariance

Suppose $\boldsymbol{X}=(X_1\,,\ldots,X_m)'$ and $\boldsymbol{Y}=(Y_1\,,\ldots,Y_n)'$ are two jointly distributed random vectors. We define their *covariance* as

$$\mathrm{Cov}(\boldsymbol{X}\,,\boldsymbol{Y}):=\begin{pmatrix}\mathrm{Cov}(X_1\,,Y_1)&\cdots&\mathrm{Cov}(X_1\,,Y_n)\\\vdots&&\vdots\\\mathrm{Cov}(X_m\,,Y_1)&\cdots&\mathrm{Cov}(X_m\,,Y_n)\end{pmatrix}.$$

Proposition 2. *We always have*

$$\mathrm{Cov}(\boldsymbol{X}\,,\boldsymbol{Y})=\mathrm{E}\left[(\boldsymbol{X}-\mathrm{E}\boldsymbol{X})\,(\boldsymbol{Y}-\mathrm{E}\boldsymbol{Y})'\right].$$

Proof. The $(i\,,j)$th entry of the matrix $(\boldsymbol{X}-\mathrm{E}\boldsymbol{X})(\boldsymbol{Y}-\mathrm{E}\boldsymbol{Y})'$ is $(X_i-\mathrm{E}X_i)(Y_j-\mathrm{E}Y_j)$, whose expectation is $\mathrm{Cov}(X_i\,,X_j)$. Because this is true for all $(i\,,j)$, the result holds coordinatewise. □

Warning. Note where the transpose is: Except in the case that n and m are the same integer, $(\boldsymbol{X}-\mathrm{E}\boldsymbol{X})'(\boldsymbol{Y}-\mathrm{E}\boldsymbol{Y})$ does not even make sense, whereas $(\boldsymbol{X}-\mathrm{E}\boldsymbol{X})(\boldsymbol{Y}-\mathrm{E}\boldsymbol{Y})'$ is always a random $m\times n$ matrix. □

An important special case occurs when we have $\boldsymbol{X}=\boldsymbol{Y}$. In that case we write

$$\mathrm{Var}(\boldsymbol{X}):=\mathrm{Cov}(\boldsymbol{X}\,,\boldsymbol{X}).$$

We call $\text{Var}(\boldsymbol{X})$ the *variance-covariance* matrix of $\boldsymbol{X}$. The terminology is motivated by the fact that

$$\text{Var}(\boldsymbol{X}) = \begin{pmatrix} \text{Var}(X_1) & \text{Cov}(X_1\,, X_2) & \text{Cov}(X_1\,, X_3) & \cdots & \text{Cov}(X_1\,, X_m) \\ \text{Cov}(X_2\,, X_1) & \text{Var}(X_2) & \text{Cov}(X_2\,, X_3) & \cdots & \text{Cov}(X_2\,, X_n) \\ \text{Cov}(X_3\,, X_1) & \text{Cov}(X_3\,, X_2) & \text{Var}(X_3) & \cdots & \text{Cov}(X_3\,, X_m) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \text{Cov}(X_m\,, X_1) & \text{Cov}(X_m\,, X_2) & \text{Cov}(X_m\,, X_3) & \cdots & \text{Var}(X_m) \end{pmatrix}.$$

Note that $\text{Var}(\boldsymbol{X})$ is *always* a square and symmetric matrix; its dimension is $m \times m$ when $\boldsymbol{X}$ is m-dimensional. On-diagonal entries of $\text{Var}(\boldsymbol{X})$ are always nonnegative; off-diagonal entries can be arbitrary real numbers.

3. Mathematical properties of variance and covariance

- Because $(\boldsymbol{X}-\text{E}\boldsymbol{X})(\boldsymbol{X}-\text{E}\boldsymbol{X})' = \boldsymbol{X}\boldsymbol{X}'-\boldsymbol{X}(\text{E}\boldsymbol{X})'-(\text{E}\boldsymbol{X})\boldsymbol{X}'+(\text{E}\boldsymbol{X})(\text{E}\boldsymbol{X})'$, it follows that

$$\begin{aligned}\text{Var}(\boldsymbol{X}) &= \text{E}\left(\boldsymbol{X}\boldsymbol{X}'\right) - 2(\text{E}\boldsymbol{X})(\text{E}\boldsymbol{X})' + (\text{E}\boldsymbol{X})(\text{E}\boldsymbol{X})' \\ &= \text{E}\left(\boldsymbol{X}\boldsymbol{X}'\right) - (\text{E}\boldsymbol{X})(\text{E}\boldsymbol{X})',\end{aligned}$$

  after expansion. This is a multidimensional extension of the formula $\text{Var}(Z) = \text{E}(Z^2) - (\text{E}Z)^2$, valid for every [univariate] random variable Z.

- If $\boldsymbol{a} \in \mathbf{R}^n$ is nonrandom, then $(\boldsymbol{X} - \boldsymbol{a}) - \text{E}(\boldsymbol{X} - \boldsymbol{a}) = \boldsymbol{X} - \text{E}\boldsymbol{X}$. Therefore,

$$\text{Var}(\boldsymbol{X} - \boldsymbol{a}) = \text{E}\left[(\boldsymbol{X} - \text{E}\boldsymbol{X})(\boldsymbol{X} - \text{E}\boldsymbol{X})'\right] = \text{Var}(\boldsymbol{X}).$$

  This should be a familiar property in the one-dimensional case.

- If $\boldsymbol{X}$, $\boldsymbol{Y}$, and $\boldsymbol{Z}$ are three jointly-distributed random vectors [with the same dimensions], then $\boldsymbol{X}((\boldsymbol{Y} + \boldsymbol{Z}) - \text{E}(\boldsymbol{Y} + \boldsymbol{Z}))' = \boldsymbol{X}(\boldsymbol{Y} - \text{E}\boldsymbol{Y})' + \boldsymbol{X}(\boldsymbol{Z} - \text{E}\boldsymbol{Z})'$. Therefore,

$$\text{Cov}(\boldsymbol{X}\,, \boldsymbol{Y} + \boldsymbol{Z}) = \text{Cov}(\boldsymbol{X}\,, \boldsymbol{Y}) + \text{Cov}(\boldsymbol{X}\,, \boldsymbol{Z}).$$

- Suppose $\boldsymbol{A}\,, \boldsymbol{B}$ are nonrandom matrices. Then, $(\boldsymbol{A}\boldsymbol{X} - \text{E}(\boldsymbol{A}\boldsymbol{X}))(\boldsymbol{B}\boldsymbol{Y} - \text{E}(\boldsymbol{B}\boldsymbol{Y}))' = \boldsymbol{A}(\boldsymbol{X} - \text{E}\boldsymbol{X})(\boldsymbol{Y} - \text{E}\boldsymbol{Y})'\boldsymbol{B}'$. Therefore,

$$\text{Cov}(\boldsymbol{A}\boldsymbol{X}\,, \boldsymbol{B}\boldsymbol{Y}) = \boldsymbol{A}\text{Cov}(\boldsymbol{X}\,, \boldsymbol{Y})\boldsymbol{B}'.$$

  The special case that $\boldsymbol{X} = \boldsymbol{Y}$ is worth pointing out: In that case we obtain the identity,

$$\text{Var}(\boldsymbol{A}\boldsymbol{X}) = \boldsymbol{A}\text{Var}(\boldsymbol{X})\boldsymbol{A}'.$$

4. A relation to positive-semidefinite matrices

Let $\boldsymbol{a} \in \mathbf{R}^n$ be a nonrandom vector and $\boldsymbol{X}$ be an n-dimensional random vector. Then, the properties of variance-covariance matrices ensure that

$$\mathrm{Var}\left(\boldsymbol{a}'\boldsymbol{X}\right) = \boldsymbol{a}'\mathrm{Var}(\boldsymbol{X})\boldsymbol{a}.$$

Because $\boldsymbol{a}'\boldsymbol{X} = \sum_{j=1}^n a_j X_j$ is univariate, $\mathrm{Var}(\boldsymbol{a}'\boldsymbol{X}) \geq 0$, and hence

$$\boldsymbol{a}'\mathrm{Var}(\boldsymbol{X})\boldsymbol{a} \geq 0 \qquad \text{for all } \boldsymbol{a} \in \mathbf{R}^n. \tag{1}$$

A real and symmetric $n \times n$ matrix $\boldsymbol{A}$ is said to be *positive semidefinite* if $\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x} \geq 0$ for all $\boldsymbol{x} \in \mathbf{R}^n$. And $\boldsymbol{A}$ is *positive definite* if $\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x} > 0$ for every nonzero $\boldsymbol{x} \in \mathbf{R}^n$.

Proposition 3. *If* $\boldsymbol{X}$ *is an n-dimensional random vector, then* $\mathrm{Var}(\boldsymbol{X})$ *is positive semidefinite. If* $\mathrm{P}\{\boldsymbol{a}'\boldsymbol{X} = b\} = 0$ *for every* $\boldsymbol{a} \in \mathbf{R}^n$ *and* $b \in \mathbf{R}$, *then* $\mathrm{Var}(\boldsymbol{X})$ *is positive definite.*

Proof. We have seen already in (1) that $\mathrm{Var}(\boldsymbol{X})$ is positive semidefinite. Now suppose that $\mathrm{P}\{\boldsymbol{a}'\boldsymbol{X} = b\} = 0$, as indicated. Then, $\boldsymbol{a}'\boldsymbol{X}$ is a genuine random variable and hence $\boldsymbol{a}'\mathrm{Var}(\boldsymbol{X})\boldsymbol{a} = \mathrm{Var}(\boldsymbol{a}'\boldsymbol{X}) > 0$ for all $\boldsymbol{a} \in \mathbf{R}^n$. □

Remark 4. The very same argument can be used to prove the following improvement: Suppose $\mathrm{P}\{\boldsymbol{a}'\boldsymbol{X} = b\} < 1$ for all $b \in \mathbf{R}$ and $\boldsymbol{a} \in \mathbf{R}^n$. Then $\mathrm{Var}(\boldsymbol{X})$ is positive definite. The proof is the same because $\mathrm{P}\{\boldsymbol{a}'\boldsymbol{X} = \mathrm{E}(\boldsymbol{a}'\boldsymbol{X})\} < 1$ implies that the variance of the random variable $\boldsymbol{a}'\boldsymbol{X}$ cannot be zero when $\boldsymbol{a} \neq \mathbf{0}$. □

Some linear algebra

Recall the convention that, for us, all vectors are column vectors.

1. Symmetric matrices

Let $\boldsymbol{A}$ be a real $n \times n$ matrix. Recall that a complex number λ is an *eigenvalue* of $\boldsymbol{A}$ if there exists a real and nonzero vector $\boldsymbol{x}$—called an eigenvector for λ—such that $\boldsymbol{A}\boldsymbol{x} = \lambda\boldsymbol{x}$. Whenever $\boldsymbol{x}$ is an eigenvector for λ, so is $a\boldsymbol{x}$ for every real number a.

The *characteristic polynomial* $\chi_{\boldsymbol{A}}$ of matrix $\boldsymbol{A}$ is the function

$$\chi_{\boldsymbol{A}}(\lambda) := \det(\lambda\boldsymbol{I} - \boldsymbol{A}),$$

defined for all complex numbers λ, where $\boldsymbol{I}$ denotes the $n \times n$ identity matrix. It is not hard to see that a complex number λ is an eigenvalue of $\boldsymbol{A}$ if and only if $\chi_{\boldsymbol{A}}(\lambda) = 0$. We see by direct computation that $\chi_{\boldsymbol{A}}$ is an nth-order polynomial. Therefore, $\boldsymbol{A}$ has precisely n eigenvalues, thanks to the fundamental theorem of algebra. We can write them as $\lambda_1, \ldots, \lambda_n$, or sometimes more precisely as $\lambda_1(\boldsymbol{A}), \ldots, \lambda_n(\boldsymbol{A})$.

1. The spectral theorem. The following important theorem is the starting point of our discussion. It might help to recall that vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k \in \mathbf{R}^n$ are *orthonormal* if $\boldsymbol{x}_i'\boldsymbol{x}_j = 0$ when $i \neq j$ and $\boldsymbol{x}_i'\boldsymbol{x}_i = \|\boldsymbol{x}_i\|^2 = 1$.

Theorem 1. *If $\boldsymbol{A}$ is a real and symmetric $n \times n$ matrix, then $\lambda_1, \ldots, \lambda_n$ are real numbers. Moreover, there exist n orthonormal eigenvectors $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n$ that correspond respectively to $\lambda_1, \ldots, \lambda_n$.*

Proof. Let λ be an eigenvalue of $\boldsymbol{A} := (a_{ij})$, with corresponding eigenvector $\boldsymbol{v}$ and observe that $\overline{\lambda}$ is an eigenvalue of $\overline{\boldsymbol{A}} := (\overline{a_{ij}})$, with corresponding eigenvector $\overline{\boldsymbol{v}}$. Therefore,

$$\lambda\|\boldsymbol{v}\|^2 = \lambda\boldsymbol{v}'\overline{\boldsymbol{v}} = (\lambda\boldsymbol{v})'\overline{\boldsymbol{v}} = (\boldsymbol{A}\boldsymbol{v})'\overline{\boldsymbol{v}} = \boldsymbol{v}'\boldsymbol{A}'\overline{\boldsymbol{v}}.$$

Because $\boldsymbol{A}$ is symmetric [and real], $\boldsymbol{A}'\boldsymbol{v} = \overline{\boldsymbol{A}\boldsymbol{v}} = \overline{\lambda}\overline{\boldsymbol{v}}$. This proves that

$$\lambda\|\boldsymbol{v}\|^2 = \overline{\lambda}\boldsymbol{v}'\overline{\boldsymbol{v}} = \overline{\lambda}\|\boldsymbol{v}\|^2.$$

Divide by $\|\boldsymbol{v}\|^2 \neq 0$ to deduce that $\lambda = \overline{\lambda}$, which means that the eigenvalue λ is real valued. Since λ is arbitrary, this proves that all eigenvalues of $\boldsymbol{A}$ are real. □

Theorem 2 (The spectral theorem)**.** *Let* $\boldsymbol{A}$ *denote a symmetric* $n \times n$ *matrix with real eigenvalues* $\lambda_1, \ldots, \lambda_n$ *and corresponding orthonormal eigenvectors* $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n$*. Define* $\boldsymbol{D} := \text{diag}(\lambda_1, \ldots, \lambda_n)$ *to be the diagonal matrix of the* λ_i*'s and* $\boldsymbol{P}$ *to be the matrix whose columns are* $\boldsymbol{v}_1$ *though* $\boldsymbol{v}_n$ *respectively; that is,*

$$\boldsymbol{D} := \begin{pmatrix} \lambda_1 & 0 & 0 & \cdots & 0 \\ 0 & \lambda_2 & 0 & \cdots & 0 \\ 0 & 0 & \lambda_3 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \lambda_n \end{pmatrix}, \quad \boldsymbol{P} := (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n).$$

Then $\boldsymbol{P}$ *is* orthogonal $[\boldsymbol{P}' = \boldsymbol{P}^{-1}]$ *and* $\boldsymbol{A} = \boldsymbol{P}\boldsymbol{D}\boldsymbol{P}^{-1} = \boldsymbol{P}\boldsymbol{D}\boldsymbol{P}'$.

Proof. $\boldsymbol{P}$ is orthogonal because the orthonormality of the $\boldsymbol{v}_i$'s implies that

$$\boldsymbol{P}'\boldsymbol{P} = \begin{pmatrix} \boldsymbol{v}_1 \\ \vdots \\ \boldsymbol{v}_n \end{pmatrix}' (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n) = \boldsymbol{I}.$$

Furthermore, because $\boldsymbol{A}\boldsymbol{v}_j = \lambda_j\boldsymbol{v}_j$, it follows that $\boldsymbol{A}\boldsymbol{P} = \boldsymbol{P}\boldsymbol{D}$, which is another way to say that $\boldsymbol{A} = \boldsymbol{P}\boldsymbol{D}\boldsymbol{P}^{-1}$. □

Recall that the *trace* of an $n \times n$ matrix $\boldsymbol{A}$ is the sum $A_{1,1} + \cdots + A_{n,n}$ of its diagonal entries.

Corollary 3. *If* A *is a real and symmetric* $n \times n$ *matrix with real eigenvalues* $\lambda_1, \ldots, \lambda_n$*, then*

$$\text{tr}(\boldsymbol{A}) = \lambda_1 + \cdots + \lambda_n \quad \textit{and} \quad \det(\boldsymbol{A}) = \lambda_1 \times \cdots \times \lambda_n.$$

Proof. Write $\boldsymbol{A}$, in spectral form, as $\boldsymbol{P}\boldsymbol{D}\boldsymbol{P}^{-1}$. Since the determinant of $\boldsymbol{P}^{-1}$ is the reciprocal of that of $\boldsymbol{A}$, it follows that $\det(\boldsymbol{A}) = \det(\boldsymbol{D})$, which is clearly

$\lambda_1 \times \cdots \times \lambda_n$. In order to compute the trace of $\boldsymbol{A}$ we compute directly also:

$$\begin{aligned}\mathrm{tr}(\boldsymbol{A}) &= \sum_{i=1}^n \sum_{j=1}^n P_{i,j} \left(\boldsymbol{D}\boldsymbol{P}^{-1}\right)_{i,j} = \sum_{i=1}^n \sum_{j=1}^n \sum_{k=1}^n P_{i,j} D_{i,k} P^{-1}_{j,k} \\ &= \sum_{i=1}^n \sum_{k=1}^n \left(\boldsymbol{P}\boldsymbol{P}^{-1}\right)_{i,k} D_{i,k} = \sum_{i=1}^n D_{i,i} = \mathrm{tr}(\boldsymbol{D}),\end{aligned}$$

which is $\lambda_1 + \cdots + \lambda_n$. □

2. The square-root matrix. Let $\boldsymbol{A}$ continue to denote a real and symmetric $n \times n$ matrix.

Proposition 4. *There exists a complex and symmetric $n \times n$ matrix $\boldsymbol{B}$—called the* square root *of $\boldsymbol{A}$ and written as $\boldsymbol{A}^{1/2}$ or even sometimes as $\sqrt{\boldsymbol{A}}$—such that $\boldsymbol{A} = \boldsymbol{B}^2 := \boldsymbol{B}\boldsymbol{B}$.*

The proof of Proposition 4 is more important than its statement. So let us prove this result.

Proof. Apply the spectral theorem and write $\boldsymbol{A} = \boldsymbol{P}\boldsymbol{D}\boldsymbol{P}^{-1}$. Since $\boldsymbol{D}$ is a diagonal matrix, its square root can be defined unambiguously as the following complex-valued $n \times n$ diagonal matrix:

$$\boldsymbol{D}^{1/2} := \begin{pmatrix} \lambda_1^{1/2} & 0 & 0 & \cdots & 0 \\ 0 & \lambda_2^{1/2} & 0 & \cdots & 0 \\ 0 & 0 & \lambda_3^{1/2} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \lambda_n^{1/2} \end{pmatrix}.$$

Define $\boldsymbol{B} := \boldsymbol{P}\boldsymbol{D}^{1/2}\boldsymbol{P}^{-1}$, and note that

$$\boldsymbol{B}^2 = \boldsymbol{P}\boldsymbol{D}^{1/2}\boldsymbol{P}^{-1}\boldsymbol{P}\boldsymbol{D}^{1/2}\boldsymbol{P}^{-1} = \boldsymbol{P}\boldsymbol{D}\boldsymbol{P}^{-1} = \boldsymbol{A},$$

since $\boldsymbol{P}^{-1}\boldsymbol{P} = \boldsymbol{I}$ and $(\boldsymbol{D}^{1/2})^2 = \boldsymbol{D}$. □

2. Positive-semidefinite matrices

Recall that an $n \times n$ matrix $\boldsymbol{A}$ is *positive semidefinite* if it is symmetric and

$$\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x} \geq 0 \qquad \text{for all } \boldsymbol{x} \in \mathbf{R}^n.$$

Recall that $\boldsymbol{A}$ is *positive definite* if it is symmetric and

$$\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x} > 0 \qquad \text{for all nonzero } \boldsymbol{x} \in \mathbf{R}^n.$$

Theorem 5. *A symmetric matrix $\boldsymbol{A}$ is positive semidefinite if and only if all of its eigenvalues are ≥ 0. $\boldsymbol{A}$ is positive definite if and only if all of its eigenvalues are > 0. In the latter case, $\boldsymbol{A}$ is also nonsingular.*

The following is a ready consequence.

Corollary 6. *All of the eigenvalues of a variance-covariance matrix are always* ≥ 0.

Now let us establish the theorem.

Proof of Theorem 5. Suppose $\boldsymbol{A}$ is positive semidefinite, and let λ denote one of its eignenvalues, together with corresponding eigenvector $\boldsymbol{x}$. Since $0 \leq \boldsymbol{x}'\boldsymbol{A}\boldsymbol{x} = \lambda\|\boldsymbol{x}\|^2$ and $\|\boldsymbol{x}\| > 0$, it follows that $\lambda \geq 0$. This proves that all of the eigenvalues of $\boldsymbol{A}$ are nonnegative. If $\boldsymbol{A}$ is positive definite, then the same argument shows that all of its eigenvalues are > 0. Because $\det(\boldsymbol{A})$ is the product of all n eigenvalues of $\boldsymbol{A}$ (Corollary 3), it follows that $\det(\boldsymbol{A}) > 0$, whence $\boldsymbol{A}$ is nonsingular.

This proves slightly more than half of the proposition. Now let us suppose that all eigenvalues of $\boldsymbol{A}$ are ≥ 0. We write $\boldsymbol{A}$ in spectral form $\boldsymbol{A} = \boldsymbol{P}\boldsymbol{D}\boldsymbol{P}'$, and observe that $\boldsymbol{D}$ is a diagonal matrix of nonnegative numbers. By virute of its construction. $\boldsymbol{A}^{1/2} = \boldsymbol{P}\boldsymbol{D}^{1/2}\boldsymbol{P}'$, and hence for all $\boldsymbol{x} \in \mathbf{R}^n$,

$$\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x} = \left(\boldsymbol{D}^{1/2}\boldsymbol{P}\boldsymbol{x}\right)'\left(\boldsymbol{D}^{1/2}\boldsymbol{P}\boldsymbol{x}\right) = \left\|\boldsymbol{D}^{1/2}\boldsymbol{P}\boldsymbol{x}\right\|^2, \tag{1}$$

which is ≥ 0. Therefore, $\boldsymbol{A}$ is positive semidefinite.

If all of the eigenvalues of $\boldsymbol{A}$ are > 0, then (1) tells us that

$$\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x} = \left\|\boldsymbol{D}^{1/2}\boldsymbol{P}\boldsymbol{x}\right\|^2 = \sum_{j=1}^n \left(\left[\boldsymbol{D}^{1/2}\boldsymbol{P}\boldsymbol{x}\right]_j\right)^2 = \sum_{j=1}^n \lambda_j \left([\boldsymbol{P}\boldsymbol{x}]_j\right)^2, \tag{2}$$

where $\lambda_j > 0$ for all j. Therefore,

$$\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x} \geq \min_{1\leq j\leq n} \lambda_j \cdot \sum_{j=1}^n \left([\boldsymbol{P}\boldsymbol{x}]_j\right)^2 = \min_{1\leq j\leq n} \lambda_j \cdot \boldsymbol{x}'\boldsymbol{P}'\boldsymbol{P}\boldsymbol{x} = \min_{1\leq j\leq n} \lambda_j \cdot \|\boldsymbol{x}\|^2.$$

Since $\min_{1\leq j\leq n} \lambda_j > 0$, it follows that $\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x} > 0$ for all nonzero $\boldsymbol{x}$. This completes the proof. □

Let us pause and point out a consequence of the proof of this last result.

Corollary 7. *If $\boldsymbol{A}$ is positive semidefinite, then its extremal eigenvalues satisfy*

$$\min_{1\leq j\leq n} \lambda_j = \min_{\|\boldsymbol{x}\|>0} \frac{\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x}}{\|\boldsymbol{x}\|^2}, \qquad \max_{1\leq j\leq n} \lambda_j = \max_{\|\boldsymbol{x}\|>0} \frac{\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x}}{\|\boldsymbol{x}\|^2}.$$

Proof. We saw, during the course of the previous proof, that

$$\min_{1\leq j\leq n} \lambda_j \cdot \|\boldsymbol{x}\|^2 \leq \boldsymbol{x}'\boldsymbol{A}\boldsymbol{x} \quad \text{for all } \boldsymbol{x} \in \mathbf{R}^n. \tag{3}$$

Optimize over all $\boldsymbol{x}$ to see that

$$\min_{1\leq j\leq n} \lambda_j \leq \min_{\|\boldsymbol{x}\|>0} \frac{\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x}}{\|\boldsymbol{x}\|^2}. \tag{4}$$

But $\min_{1\le j\le n}\lambda_j$ is an eigenvalue for $\boldsymbol{A}$; let $\boldsymbol{z}$ denote a corresponding eigenvector in order to see that

$$\min_{1\le j\le n}\lambda_j \le \min_{\|\boldsymbol{x}\|>0}\frac{\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x}}{\|\boldsymbol{x}\|^2} \le \frac{\boldsymbol{z}'\boldsymbol{A}\boldsymbol{z}}{\|\boldsymbol{z}\|^2} = \min_{1\le j\le n}\lambda_j.$$

So both inequalities are in fact equalities, and hence follows the formula for the minimum eigenvalue. The one for the maximum eigenvalue is proved similarly. □

Finally, a word about the square root of positive semidefinite matrices:

Proposition 8. *If $\boldsymbol{A}$ is positive semidefinite, then so is $\boldsymbol{A}^{1/2}$. If $\boldsymbol{A}$ is positive definite, then so is $\boldsymbol{A}^{1/2}$.*

Proof. We write, in spectral form, $\boldsymbol{A} = \boldsymbol{PDP}'$ and observe [by squaring it] that $\boldsymbol{A}^{1/2} = \boldsymbol{PD}^{1/2}\boldsymbol{P}'$. Note that $\boldsymbol{D}^{1/2}$ is a real diagonal matrix since the eigenvalues of $\boldsymbol{A}$ are ≥ 0. Therefore, we may apply (1) to $\boldsymbol{A}^{1/2}$ [in place of $\boldsymbol{A}$] to see that $\boldsymbol{x}'\boldsymbol{A}^{1/2}\boldsymbol{x} = \|\boldsymbol{D}^{1/4}\boldsymbol{Px}\|^2 \ge 0$ where $\boldsymbol{D}^{1/4}$ denotes the [real] square root of $\boldsymbol{D}^{1/2}$. This proves that if $\boldsymbol{A}$ is positive semidefinite, then so is $\mathbf{A}^{1/2}$. Now suppose there exists a positive definite $\boldsymbol{A}$ whose square root is not positive definite. It would follow that there necessarily exists a nonzero $\boldsymbol{x} \in \mathbf{R}^n$ such that $\boldsymbol{x}'\boldsymbol{A}^{1/2}\boldsymbol{x} = \|\boldsymbol{D}^{1/4}\boldsymbol{Px}\|^2 = 0$. Since $\boldsymbol{D}^{1/4}\boldsymbol{Px} = \mathbf{0}$,

$$\boldsymbol{D}^{1/2}\boldsymbol{Px} = \mathbf{D}^{1/4}\boldsymbol{D}^{1/4}\boldsymbol{Px} = \mathbf{0} \qquad \Rightarrow \qquad \boldsymbol{x}'\boldsymbol{A}\boldsymbol{x} = \left\|\boldsymbol{D}^{1/2}\boldsymbol{Px}\right\|^2 = 0.$$

And this contradicts the assumption that $\boldsymbol{A}$ is positive definite. □

3. The rank of a matrix

Recall that vectors $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k$ are *linearly independent* if

$$c_1\boldsymbol{v}_1 + \cdots + c_k\boldsymbol{v}_k = \mathbf{0} \qquad \Rightarrow \qquad c_1 = \cdots = c_k = 0.$$

For instance, $\boldsymbol{v}_1 := (1\,,0)'$ and $\boldsymbol{v}_2 := (0\,,1)'$ are linearly independent 2-vectors.

The *column rank* of a matrix $\boldsymbol{A}$ is the maximum number of linearly independent column vectors of $\boldsymbol{A}$. The *row rank* of a matrix $\boldsymbol{A}$ is the maximum number of linearly independent row vectors of $\boldsymbol{A}$. We can interpret these definitions geometrically as follows: First, suppose $\boldsymbol{A}$ is $m \times n$ and define $\mathcal{C}(\boldsymbol{A})$ denote the linear space of all vectors of the form $c_1\boldsymbol{v}_1 + \cdots + c_n\boldsymbol{v}_n$, where $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n$ are the column vectors of $\boldsymbol{A}$ and $c_1, \ldots, c_n$ are real numbers. We call $\mathcal{C}(\boldsymbol{A})$ the *column space* of $\boldsymbol{A}$.

We can define the *row space* $\mathcal{R}(\boldsymbol{A})$, of $\boldsymbol{A}$ similarly, or simply define $\mathcal{R}(\boldsymbol{A}) := \mathcal{C}(\boldsymbol{A}')$.

Lemma 9. *For every $m \times n$ matrix $\boldsymbol{A}$,*

$$\mathcal{C}(\boldsymbol{A}) = \{\boldsymbol{Ax} :\ \boldsymbol{x} \in \mathbf{R}^n\}, \quad \mathcal{R}(\boldsymbol{A}) := \left\{\boldsymbol{x}'\boldsymbol{A} :\ \boldsymbol{x} \in \mathbf{R}^m\right\}.$$

We can think of an $m \times n$ matrix $\boldsymbol{A}$ as a mapping from $\mathbf{R}^n$ into $\mathbf{R}^m$; namely, we can think of matrix $\boldsymbol{A}$ also as the function $\boldsymbol{f_A(x)} := \boldsymbol{x} \mapsto \boldsymbol{Ax}$. In this way we see that $\mathcal{C}(\boldsymbol{A})$ is also the "range" of the function $\boldsymbol{f_A}$.

Proof. Let us write the columns of $\boldsymbol{A}$ as $\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_n$. Note that $\boldsymbol{y} \in \mathcal{C}(\boldsymbol{A})$ if and only if there exist $c_1, \ldots, c_n$ such that $\boldsymbol{y} = c_1\boldsymbol{a}_1 + \cdots + c_n\boldsymbol{a}_n = \boldsymbol{Ac}$, where $\boldsymbol{c} := (c_1\, , \ldots , c_n)'$. This shows that $\mathcal{C}(\boldsymbol{A})$ is the collection of all vectors of the form $\boldsymbol{Ax}$, for $\boldsymbol{x} \in \mathbf{R}^n$. The second assertion [about $\mathcal{R}(\boldsymbol{A})$] follows from the definition of $\mathcal{R}(\boldsymbol{A}$ equalling $\mathcal{C}(\boldsymbol{A}')$ and the already-proven first assertion. □

It then follows, from the definition of dimension, that

$$\text{column rank of } \boldsymbol{A} = \dim \mathcal{C}(\boldsymbol{A}), \qquad \text{row rank of } \boldsymbol{A} = \dim \mathcal{R}(\boldsymbol{A}).$$

Proposition 10. *Given any matrix $\boldsymbol{A}$, its row rank and column rank are the same. We write their common value as* $\text{rank}(\boldsymbol{A})$.

Proof. Suppose $\boldsymbol{A}$ is $m \times n$ and its column rank is r. Let $\boldsymbol{b}_1, \ldots, \boldsymbol{b}_r$ denote a basis for $\mathcal{C}(\boldsymbol{A})$ and consider the matrix $m \times r$ matrix $\boldsymbol{B} := (\boldsymbol{b}_1\, , \ldots , \boldsymbol{b}_r)$. Write $\boldsymbol{A}$, columnwise , as $\boldsymbol{A} := (\boldsymbol{a}_1\, , \ldots , \boldsymbol{a}_n)$. For every $1 \le j \le n$, there exists $c_{1,j}, \ldots, c_{r,j}$ such that $\boldsymbol{a}_j = c_{1,j}\boldsymbol{b}_1 + \cdots + c_{r,j}\boldsymbol{b}_r$. Let $\boldsymbol{C} := (c_{i,j})$ be the resulting $r \times n$ matrix, and note that $\boldsymbol{A} = \boldsymbol{BC}$. Because $A_{i,j} = \sum_{k=1}^r B_{i,k}C_{k,j};$, every row of $\boldsymbol{A}$ is a linear combination of the rows of $\boldsymbol{C}$. In other words, $\mathcal{R}(\boldsymbol{A}) \subseteq \mathcal{R}(\boldsymbol{C})$ and hence the row rank of $\boldsymbol{A}$ is $\le \dim \mathcal{R}(\boldsymbol{C}) = r =$ the column rank of $\boldsymbol{A}$. Apply this fact to $\boldsymbol{A}'$ to see that also the row rank of $\boldsymbol{A}'$ is $\le$ the column rank of $\boldsymbol{A}'$; equivalently that the column rank of $\boldsymbol{A}$ is $\le$ the row rank of $\boldsymbol{A}$. □

Proposition 11. *If $\boldsymbol{A}$ is $n \times m$ and $\boldsymbol{B}$ is $m \times k$, then*

$$\text{rank}(\boldsymbol{AB}) \le \min\left(\text{rank}(\boldsymbol{A})\, , \text{rank}(\boldsymbol{B})\right).$$

Proof. The proof uses an idea that we exploited already in the proof of Proposition 10: Since $(\boldsymbol{AB})_{j,l} = \sum_{\nu=1}^n A_{j,\nu}B_{\nu,l}$, the rows of $\boldsymbol{AB}$ are linear combinations of the rows of $\boldsymbol{B}$; that is $\mathcal{R}(\boldsymbol{AB}) \subseteq \mathcal{R}(\boldsymbol{B})$, whence $\text{rank}(\boldsymbol{AB}) \le \text{rank}(\boldsymbol{B})$. Also, $\mathcal{C}(\boldsymbol{AB}) \subseteq \mathcal{C}(\boldsymbol{A})$, whence $\text{rank}(\boldsymbol{AB}) \le \text{rank}(\boldsymbol{A})$. These observations complete the proof. □

Proposition 12. *If $\boldsymbol{A}$ and $\boldsymbol{C}$ are nonsingular, then*

$$\text{rank}(\boldsymbol{ABC}) = \text{rank}(\boldsymbol{B}),$$

provided that the dimensions match up so that $\boldsymbol{ABC}$ makes sense.

Proof. Let $\boldsymbol{D} := \boldsymbol{ABC}$; our goal is to show that $\text{rank}(\boldsymbol{D}) = \text{rank}(\boldsymbol{B})$.

Two applications of the previous proposition together yield $\text{rank}(\boldsymbol{D}) \le \text{rank}(\boldsymbol{AB}) \le \text{rank}(\boldsymbol{B})$. And since $\boldsymbol{B} = \boldsymbol{A}^{-1}\boldsymbol{D}\boldsymbol{C}^{-1}$, we have also $\text{rank}(\boldsymbol{B}) \le \text{rank}(\boldsymbol{A}^{-1}\boldsymbol{D}) \le \text{rank}(\boldsymbol{D})$. □

Corollary 13. *If $\boldsymbol{A}$ is an $n \times n$ real and symmetric matrix, then* $\text{rank}(\boldsymbol{A}) =$ *the total number of nonzero eigenvalues of $\boldsymbol{A}$. In particular, $\boldsymbol{A}$ has full rank if and only if $\boldsymbol{A}$ is nonsingular. Finally, $\mathcal{C}(\boldsymbol{A})$ is the linear space spanned by the eigenvectors of $\boldsymbol{A}$ that correspond to nonzero eigenvalues.*

Proof. We write $\boldsymbol{A}$, in spectral form, as $\boldsymbol{A} = \boldsymbol{PDP}^{-1}$, and apply the preceding proposition to see that $\text{rank}(\boldsymbol{A}) = \text{rank}(\boldsymbol{D})$, which is clearly the total number of nonzero eigenvalue of $\boldsymbol{A}$. Since $\boldsymbol{A}$ is nonsingular if and only if all of its eigenvalues are nonzero, $\boldsymbol{A}$ has full rank if and only if $\boldsymbol{A}$ is nonsingular.

Finally, suppose $\boldsymbol{A}$ has rank $k \leq n$; this is the number of its nonzero eigenvalues $\lambda_1, \ldots, \lambda_k$. Let $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n$ denote orthonormal eigenvectors such that $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k$ are eigenvectors that correspond to $\lambda_1, \ldots, \lambda_k$ and $\boldsymbol{v}_{k+1}, \ldots, \boldsymbol{v}_n$ are eigenvectors that correspond to eigenvalues 0 [Gram–Schmidt]. And define $\mathcal{E}$ to be the span of $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k$; i.e.,

$$\mathcal{E} := \{c_1\boldsymbol{v}_1 + \cdots + c_k\boldsymbol{v}_k : \; c_1, \ldots, c_k \in \mathbf{R}\}.$$

Our final goal is to prove that $\mathcal{E} = \mathcal{C}(\boldsymbol{A})$, which we know is equal to the linear space of all vectors of the form $\boldsymbol{Ax}$.

Clearly, $c_1\boldsymbol{v}_1 + \cdots + c_k\boldsymbol{v}_k = \boldsymbol{Ax}$, where $\boldsymbol{x} = \sum_{j=1}^k (c_j/\lambda_j)\boldsymbol{v}_j$. Therefore, $\mathcal{E} \subseteq \mathcal{C}(\boldsymbol{A})$. If $k = n$, then this suffices because in that case $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k$ is a basis for $\mathbf{R}^n$, hence $\mathcal{E} = \mathcal{C}(\boldsymbol{A}) = \mathbf{R}^n$. If $k < n$, then we can write every $\boldsymbol{x} \in \mathbf{R}^n$ as $a_1\boldsymbol{v}_1 + \cdots + a_n\boldsymbol{v}_n$, so that $\mathbf{Ax} = \sum_{j=1}^k a_j\lambda_j\boldsymbol{v}_j \in \mathcal{E}$. Thus, $\mathcal{C}(\boldsymbol{A}) \subseteq \mathcal{E}$ and we are done. □

Let $\boldsymbol{A}$ be $m \times n$ and define the *null space* [or "kernel"] of $\boldsymbol{A}$ as

$$\mathcal{N}(\boldsymbol{A}) := \{\boldsymbol{x} \in \mathbf{R}^n : \; \boldsymbol{Ax} = \mathbf{0}\}.$$

Note that $\mathcal{N}(\boldsymbol{A})$ is the linear span of the eigenvectors of $\boldsymbol{A}$ that correspond to eigenvalue 0. The other eigenvectors can be chosen to be orthogonal to these, and hence the preceding proof contains the facts that: (i) Nonzero elements of $\mathcal{N}(\boldsymbol{A})$ are orthogonal to nonzero elements of $\mathcal{C}(\boldsymbol{A})$; and (ii)

$$\dim \mathcal{N}(\boldsymbol{A}) + \text{rank}(\boldsymbol{A}) = n \; (= \text{the number of columns of } \boldsymbol{A}). \tag{5}$$

Proposition 14. $\text{rank}(\boldsymbol{A}) = \text{rank}(\boldsymbol{A}'\boldsymbol{A}) = \text{rank}(\boldsymbol{AA}')$ *for every $m \times n$ matrix $\boldsymbol{A}$.*

Proof. If $\boldsymbol{Ax} = \mathbf{0}$ then $\boldsymbol{A}'\boldsymbol{Ax} = \mathbf{0}$, and if $\boldsymbol{A}'\boldsymbol{Ax} = \mathbf{0}$, then $\|\boldsymbol{Ax}\|^2 = \boldsymbol{x}'\boldsymbol{A}'\boldsymbol{Ax} = 0$. In other words, $\mathcal{N}(\boldsymbol{A}) = \mathcal{N}(\boldsymbol{A}'\boldsymbol{A})$. Because $\boldsymbol{A}'\boldsymbol{A}$ and $\boldsymbol{A}$ both have n columnes, it follows from (5) that $\text{rank}(\boldsymbol{A}'\boldsymbol{A}) = \text{rank}(\boldsymbol{A})$. Apply this observation to $\boldsymbol{A}'$ to see that $\text{rank}(\boldsymbol{A}') = \text{rank}(\boldsymbol{AA}')$ as well. The result follows from this and the fact that $\boldsymbol{A}$ and $\boldsymbol{A}'$ have the same rank (Proposition 10). □

4. Projection matrices

A matrix $\boldsymbol{A}$ is said to be a *projection* matrix if: (i) $\boldsymbol{A}$ is symmetric; and (ii) $\boldsymbol{A}$ is "idempotent"; that is, $\boldsymbol{A}^2 = \boldsymbol{A}$.

Note that projection matrices are always positive semidefinite. Indeed, $\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x} = \boldsymbol{x}'\boldsymbol{A}^2\boldsymbol{x} = \boldsymbol{x}'\boldsymbol{A}'\boldsymbol{A}\boldsymbol{x} = \|\boldsymbol{A}\boldsymbol{x}\|^2 \geq 0$

Proposition 15. *If $\boldsymbol{A}$ is an $n \times n$ projection matrix, then so is $\boldsymbol{I} - \boldsymbol{A}$. Moreover, all eigenvalues of $\boldsymbol{A}$ are zeros and ones, and* $\text{rank}(\boldsymbol{A})$ *= the number of eigenvalues that are equal to one.*

Proof. $(\boldsymbol{I} - \boldsymbol{A})^2 = \boldsymbol{I} - 2\boldsymbol{A} + \boldsymbol{A}^2 = \boldsymbol{I} - \boldsymbol{A}$. Since $\boldsymbol{I} - \boldsymbol{A}$ is symmetric also, it is a projection. If λ is an eigenvalue of $\boldsymbol{A}$ and $\boldsymbol{x}$ is a corresponding eigenvector, then $\lambda\boldsymbol{x} = \boldsymbol{A}\boldsymbol{x} = \boldsymbol{A}^2\boldsymbol{x} = \lambda\boldsymbol{A}\boldsymbol{x} = \lambda^2\boldsymbol{x}$. Multiply both sides by $\boldsymbol{x}'$ to see that $\lambda\|\boldsymbol{x}\|^2 = \lambda^2\|\boldsymbol{x}\|^2$. Since $\|\boldsymbol{x}\| > 0$, it follows that $\lambda \in \{0\,,1\}$. The total number of nonzero eigenvalues is then the total number of eigenvalues that are ones. Therefore, the rank of $\boldsymbol{A}$ is the total number of eigenvalues that are one. □

Corollary 16. *If $\boldsymbol{A}$ is a projection matrix, then* $\text{rank}(\boldsymbol{A}) = \text{tr}(\boldsymbol{A})$.

Proof. Simply recall that $\text{tr}(\boldsymbol{A})$ is the sum of the eigenvalues, which for a projection matrix, is the total number of eigenavalues that are one. □

Why are they called "projection" matrices? Or, perhaps even more importantly, what is a "projection"?

Lemma 17. *Let Ω denote a linear subspace of $\mathbf{R}^n$, and $\boldsymbol{x} \in \mathbf{R}^n$ be fixed. Then there exists a unique element $\boldsymbol{y} \in \Omega$ that is closest to $\boldsymbol{x}$; that is,*

$$\|\boldsymbol{y} - \boldsymbol{x}\| = \min_{z \in \Omega} \|\boldsymbol{z} - \boldsymbol{x}\|.$$

The point $\boldsymbol{y}$ is called the projection *of $\boldsymbol{x}$ onto Ω.*

Proof. Let $k := \dim \Omega$, so that there exists an orthonormal basis $\boldsymbol{b}_1, \ldots, \boldsymbol{b}_k$ for Ω. Extend this to a basis $\boldsymbol{b}_1, \ldots, \boldsymbol{b}_n$ for all of $\mathbf{R}^n$ by the Gram–Schmidt method.

Given a fixed vector $\boldsymbol{x} \in \mathbf{R}^n$, we can write it as $\boldsymbol{x} := c_1\boldsymbol{b}_1 + \cdots + c_n\boldsymbol{b}_n$ for some $c_1, \ldots, c_n \in \mathbf{R}$. Define $\boldsymbol{y} := c_1\boldsymbol{b}_1 + \cdots + c_k\boldsymbol{b}_k$. Clearly, $\boldsymbol{y} \in \Omega$ and $\|\boldsymbol{y} - \boldsymbol{x}\|^2 = \sum_{i=k+1}^n c_i^2$. Any other $\boldsymbol{z} \in \Omega$ can be written as $\boldsymbol{z} = \sum_{i=1}^k d_i\boldsymbol{b}_i$, and hence $\|\boldsymbol{z} - \boldsymbol{x}\|^2 = \sum_{i=1}^k (d_i - c_i)^2 + \sum_{i=k+1}^n c_i^2$, which is strictly greater than $\|\boldsymbol{y} - \boldsymbol{x}\|^2 = \sum_{i=k+1}^n c_i^2$ unless $d_i = c_i$ for all $i = 1, \ldots, k$; i.e., unless $\boldsymbol{z} = \boldsymbol{y}$. □

Usually, we have a k-dimensional linear subspace Ω of $\mathbf{R}^n$ that is the range of some $n \times k$ matrix $\boldsymbol{A}$. That is, $\Omega = \{\boldsymbol{A}\boldsymbol{y} :\ \boldsymbol{y} \in \mathbf{R}^k\}$. Equivalently, $\Omega = \mathcal{C}(\boldsymbol{A})$. In that case,

$$\min_{z \in \Omega} \|\boldsymbol{z} - \boldsymbol{x}\|^2 = \min_{y \in \mathrm{R}^k} \|\boldsymbol{A}\boldsymbol{y} - \boldsymbol{x}\|^2 = \min_{y \in \mathrm{R}^k} \left[\boldsymbol{y}'\boldsymbol{A}'\boldsymbol{A}\boldsymbol{y} - \boldsymbol{y}'\boldsymbol{A}'\boldsymbol{x} - \boldsymbol{x}'\boldsymbol{A}\boldsymbol{y} + \boldsymbol{x}'\boldsymbol{x}\right].$$

Because $\boldsymbol{y}'\boldsymbol{A}'\boldsymbol{x}$ is a scalar, the preceding is simplified to

$$\min_{\boldsymbol{z}\in\Omega}\|\boldsymbol{z}-\boldsymbol{x}\|^2=\min_{\boldsymbol{y}\in\mathrm{R}^k}\left[\boldsymbol{y}'\boldsymbol{A}'\boldsymbol{A}\boldsymbol{y}-2\boldsymbol{y}'\boldsymbol{A}'\boldsymbol{x}+\boldsymbol{x}'\boldsymbol{x}\right].$$

Suppose that the $k\times k$ positive semidefinite matrix $\boldsymbol{A}'\boldsymbol{A}$ is nonsingular [so that $\boldsymbol{A}'\boldsymbol{A}$ and hence also $(\boldsymbol{A}\boldsymbol{A}')^{-1}$ are both positive definite]. Then, we can relabel variables [$\boldsymbol{\alpha}:=\boldsymbol{A}'\boldsymbol{A}\boldsymbol{y}$] to see that

$$\min_{\boldsymbol{z}\in\Omega}\|\boldsymbol{z}-\boldsymbol{x}\|^2=\min_{\boldsymbol{\alpha}\in\mathrm{R}^k}\left[\boldsymbol{\alpha}'(\boldsymbol{A}'\boldsymbol{A})^{-1}\boldsymbol{\alpha}-2\boldsymbol{\alpha}'(\boldsymbol{A}'\boldsymbol{A})^{-1}\boldsymbol{A}'\boldsymbol{x}+\boldsymbol{x}'\boldsymbol{x}\right].$$

A little arithmetic shows that

$$\begin{aligned}&(\boldsymbol{\alpha}-\boldsymbol{A}'\boldsymbol{x})'(\boldsymbol{A}'\boldsymbol{A})^{-1}(\boldsymbol{\alpha}-\boldsymbol{A}'\boldsymbol{x})\\&\qquad=\boldsymbol{\alpha}'(\boldsymbol{A}'\boldsymbol{A})^{-1}\boldsymbol{\alpha}-2\boldsymbol{\alpha}'(\boldsymbol{A}'\boldsymbol{A})^{-1}\boldsymbol{A}'\boldsymbol{x}+\boldsymbol{x}'\boldsymbol{A}(\boldsymbol{A}'\boldsymbol{A})^{-1}\boldsymbol{A}'\boldsymbol{x}.\end{aligned}$$

Consequently,

$$\begin{aligned}&\min_{\boldsymbol{z}\in\Omega}\|\boldsymbol{z}-\boldsymbol{x}\|^2\\&=\min_{\boldsymbol{\alpha}\in\mathrm{R}^k}\left[(\boldsymbol{\alpha}-\boldsymbol{A}'\boldsymbol{x})'(\boldsymbol{A}'\boldsymbol{A})^{-1}(\boldsymbol{\alpha}-\boldsymbol{A}'\boldsymbol{x})-\boldsymbol{x}'\boldsymbol{A}(\boldsymbol{A}'\boldsymbol{A})^{-1}\boldsymbol{A}'\boldsymbol{x}+\boldsymbol{x}'\boldsymbol{x}\right].\end{aligned}$$

The first term in the parentheses is ≥ 0; in fact it is >0 unless we select $\boldsymbol{\alpha}=\boldsymbol{A}'\boldsymbol{x}$. This proves that the projection of $\boldsymbol{x}$ onto Ω is obtained by setting $\boldsymbol{\alpha}:=\boldsymbol{A}'\boldsymbol{x}$, in which case the projection itself is $\boldsymbol{A}\boldsymbol{y}=\boldsymbol{A}(\boldsymbol{A}'\boldsymbol{A})^{-1}\boldsymbol{A}'\boldsymbol{x}$ and the distance between $\boldsymbol{y}$ and $\boldsymbol{x}$ is the square root of $\|\boldsymbol{x}\|^2-\boldsymbol{x}'\boldsymbol{A}(\boldsymbol{A}'\boldsymbol{A})^{-1}\boldsymbol{A}'\boldsymbol{x}$.

Let $\boldsymbol{P}_\Omega:=\boldsymbol{A}(\boldsymbol{A}'\boldsymbol{A})^{-1}\boldsymbol{A}'$. It is easy to see that $\boldsymbol{P}_\Omega$ is a projection matrix. The preceding shows that $\boldsymbol{P}_\Omega\boldsymbol{x}$ is the projection of $\boldsymbol{x}$ onto Ω for every $\boldsymbol{x}\in\mathbf{R}^n$. That is, we can think of $\boldsymbol{P}_\Omega$ as the matrix that projects onto Ω. Moreover, the distance between $\boldsymbol{x}$ and the linear subspace Ω [i.e., $\min_{\boldsymbol{z}\in\mathrm{R}^k}\|\boldsymbol{z}-\boldsymbol{x}\|$] is exactly the square root of $\boldsymbol{x}'\boldsymbol{x}-\boldsymbol{x}'\boldsymbol{P}_\Omega\boldsymbol{x}=\boldsymbol{x}'(\boldsymbol{I}-\boldsymbol{P}_\Omega)\boldsymbol{x}=\|(\boldsymbol{I}-\boldsymbol{P}_\Omega)\boldsymbol{x}\|^2$, because $\boldsymbol{I}-\boldsymbol{P}_\Omega$ is a projection matrix. What space does it project into?

Let $\Omega^\perp$ denote the collection of all n-vectors that are perpendicular to every element of Ω. If $\boldsymbol{z}\in\Omega^\perp$, then we can write, for all $\boldsymbol{x}\in\mathbf{R}^n$,

$$\begin{aligned}\|\boldsymbol{z}-\boldsymbol{x}\|^2&=\|\boldsymbol{z}-(\boldsymbol{I}-\boldsymbol{P}_\Omega)\boldsymbol{x}+\boldsymbol{P}_\Omega\boldsymbol{x}\|^2\\&=\|\boldsymbol{z}-(\boldsymbol{I}-\boldsymbol{P}_\Omega)\boldsymbol{x}\|^2+\|\boldsymbol{P}_\Omega\boldsymbol{x}\|^2-2\left\{\boldsymbol{z}-(\boldsymbol{I}-\boldsymbol{P}_\Omega)\boldsymbol{x}\right\}'\boldsymbol{P}_\Omega\boldsymbol{x}\\&=\|\boldsymbol{z}-(\boldsymbol{I}-\boldsymbol{P}_\Omega)\boldsymbol{x}\|^2+\|\boldsymbol{P}_\Omega\boldsymbol{x}\|^2,\end{aligned}$$

since $\boldsymbol{z}$ is orthogonal to every element of Ω including $\boldsymbol{P}_\Omega\boldsymbol{x}$, and $\boldsymbol{P}_\Omega=\boldsymbol{P}_\Omega^2$. Take the minimum over all $\boldsymbol{z}\in\Omega^\perp$ to find that $\boldsymbol{I}-\boldsymbol{P}_\Omega$ is the projection onto $\Omega^\perp$. Let us summarize our findings.

Proposition 18. *If $\boldsymbol{A}'\boldsymbol{A}$ is nonsingular [equivalently, has full rank], then $\boldsymbol{P}_{\mathcal{C}(\boldsymbol{A})}:=\boldsymbol{A}(\boldsymbol{A}'\boldsymbol{A})^{-1}\boldsymbol{A}'$ is the projection onto $\mathcal{C}(\boldsymbol{A})$, $\boldsymbol{I}-\boldsymbol{P}_{\mathcal{C}(\boldsymbol{A})}=\boldsymbol{P}_{\mathcal{C}(\boldsymbol{A})^\perp}$ is the projection onto $\Omega^\perp$, and we have*

$$\boldsymbol{x}=\boldsymbol{P}_{\mathcal{C}(\boldsymbol{A})}\boldsymbol{x}+\boldsymbol{P}_{\mathcal{C}(\boldsymbol{A})^\perp}\boldsymbol{x},\quad\textit{and}\quad\|\boldsymbol{x}\|^2=\left\|\boldsymbol{P}_{\mathcal{C}(\boldsymbol{A})}\boldsymbol{x}\right\|^2+\left\|\boldsymbol{P}_{\mathcal{C}(\boldsymbol{A})^\perp}\boldsymbol{x}\right\|^2.$$

The last result is called the "Pythagorean property."

Quadratic forms

Let $\boldsymbol{A}$ be a real and symmetric $n\times n$ matrix. Then the *quadratic form* associated to $\boldsymbol{A}$ is the function $Q_{\boldsymbol{A}}$ defined by

$$Q_{\boldsymbol{A}}(\boldsymbol{x}) := \boldsymbol{x}'\boldsymbol{A}\boldsymbol{x} \qquad (\boldsymbol{x} \in \mathbf{R}^n).$$

We have seen quadratic forms already, particularly in the context of positive-semidefinite matrices.

1. Random quadratic forms

Let $\boldsymbol{X} := (X_1, \ldots, X_n)'$ be an n-dimensional random vector. We are interested in the random quadratic form $Q_{\boldsymbol{A}}(\boldsymbol{X}) := \boldsymbol{X}'\boldsymbol{A}\boldsymbol{X}$.

Proposition 1. *If* $\mathrm{E}\boldsymbol{X} := \boldsymbol{\mu}$ *and* $\mathrm{Var}(\boldsymbol{X}) := \boldsymbol{\Sigma}$, *then*

$$\mathrm{E}\left(\boldsymbol{X}'\boldsymbol{A}\boldsymbol{X}\right) = \mathrm{tr}(\boldsymbol{A}\boldsymbol{\Sigma}) + \boldsymbol{\mu}'\boldsymbol{A}\boldsymbol{\mu}.$$

In symbols, $\mathrm{E}(Q_{\boldsymbol{A}}(\boldsymbol{X})) = \mathrm{tr}(\boldsymbol{A}\boldsymbol{\Sigma}) + Q_{\boldsymbol{A}}(\boldsymbol{\mu})$.

Proof. We can write

$$\begin{aligned}\boldsymbol{X}'\boldsymbol{A}\boldsymbol{X} &= (\boldsymbol{X}-\boldsymbol{\mu})'\boldsymbol{A}\boldsymbol{X} + \boldsymbol{\mu}'\boldsymbol{A}\boldsymbol{X}\\ &= (\boldsymbol{X}-\boldsymbol{\mu})'\boldsymbol{A}(\boldsymbol{X}-\boldsymbol{\mu}) + \boldsymbol{\mu}'\boldsymbol{A}\boldsymbol{X} + (\boldsymbol{X}-\boldsymbol{\mu})'\boldsymbol{A}\boldsymbol{\mu}.\end{aligned}$$

If we take expectations, then the last term vanishes and we obtain

$$\mathrm{E}\left(\boldsymbol{X}'\boldsymbol{A}\boldsymbol{X}\right) = \mathrm{E}\left[(\boldsymbol{X}-\boldsymbol{\mu})'\boldsymbol{A}(\boldsymbol{X}-\boldsymbol{\mu})\right] + \boldsymbol{\mu}'\boldsymbol{A}\boldsymbol{\mu}.$$

It suffices to verify that the expectation on the right-hand side is the trace of $\boldsymbol{A}\boldsymbol{\Sigma}$. But this is a direct calculation: Let $Y_j := X_i - \mu_j$, so that $\boldsymbol{Y} = \boldsymbol{X} - \boldsymbol{\mu}$

and hence

$$\begin{aligned}
\mathrm{E}\left[(\boldsymbol{X}-\boldsymbol{\mu})'\boldsymbol{A}(\boldsymbol{X}-\boldsymbol{\mu})\right] &= \mathrm{E}\left(\boldsymbol{Y}'\boldsymbol{A}\boldsymbol{Y}\right)\\
&= \sum_{i=1}^n\sum_{j=1}^n \mathrm{E}\left(Y_iA_{i,j}Y_j\right) = \sum_{i=1}^n\sum_{j=1}^n A_{i,j}\left[\mathrm{Var}(\boldsymbol{Y})\right]_{i,j}\\
&= \sum_{i=1}^n\sum_{j=1}^n A_{i,j}\left[\mathrm{Var}(\boldsymbol{X}-\boldsymbol{\mu})\right]_{i,j} = \sum_{i=1}^n\sum_{j=1}^n A_{i,j}\left[\mathrm{Var}(\boldsymbol{X})\right]_{i,j}\\
&= \sum_{i=1}^n\sum_{j=1}^n A_{i,j}\Sigma_{i,j} = \sum_{i=1}^n\sum_{i=1}^n A_{i,j}\Sigma_{j,i}\\
&= \sum_{i=1}^n \left[\boldsymbol{A\Sigma}\right]_{i,i} = \mathrm{tr}(\boldsymbol{A\Sigma}),
\end{aligned}$$

as desired. □

We easily get the following by a relabeling ($\boldsymbol{X} \Leftrightarrow \boldsymbol{X}-\boldsymbol{b}$):

Corollary 2. *For every nonrandom* $\boldsymbol{b} \in \mathbf{R}^n$,

$$\mathrm{E}\left[(\boldsymbol{X}-\boldsymbol{b})'\boldsymbol{A}(\boldsymbol{X}-\boldsymbol{b})\right] = \mathrm{tr}(\boldsymbol{A\Sigma}) + (\boldsymbol{\mu}-\boldsymbol{b})'\boldsymbol{A}(\boldsymbol{\mu}-\boldsymbol{b}).$$

In particular, $\mathrm{E}[(\boldsymbol{X}-\boldsymbol{\mu})'\boldsymbol{A}(\boldsymbol{X}-\boldsymbol{\mu})] = \mathrm{tr}(\boldsymbol{A\Sigma})$.

2. Examples of quadratic forms

What do quadratic forms look like? It is best to proceed by example.

Example 3. If $\boldsymbol{A} := \boldsymbol{I}_{n\times n}$, then $Q_{\boldsymbol{A}}(\boldsymbol{x}) = \sum_{i=1}^n x_i^2$. Because $\mathrm{tr}(\boldsymbol{A\Sigma}) = \mathrm{tr}(\boldsymbol{\Sigma}) = \sum_{i=1}^n \mathrm{Var}(X_i)$, it follows that

$$\mathrm{E}\left(\sum_{i=1}^n X_i^2\right) = \sum_{i=1}^n \mathrm{Var}(X_i) + \left(\sum_{i=1}^n \mu_i^2\right).$$

This ought to be a familiar formula. □

Example 4. If

$$\boldsymbol{A} := \mathbf{1}_{m\times m} := \begin{pmatrix} 1 & \cdots & 1\\ \vdots & & \vdots\\ 1 & \cdots & 1\end{pmatrix}_{m\times m},$$

then $Q_{\boldsymbol{A}}(\boldsymbol{x}) = (\sum_{i=1}^n x_i)^2$. Note that

$$\mathrm{tr}(\boldsymbol{A\Sigma}) = \sum_{i=1}^n\sum_{j=1}^n A_{i,j}\Sigma_{i,j} = \sum_{i=1}^n\sum_{j=1}^n \mathrm{Cov}(X_i\,,X_j).$$

Therefore,

$$\mathrm{E}\left[\left(\sum_{i=1}^n X_i\right)^2\right] = \sum_{i=1}^n\sum_{j=1}^n \mathrm{Cov}(X_i\,, X_j) + \left(\sum_{i=1}^n \mu_i\right)^2 ;$$

this is another familiar formula. □

Example 5. One can combine matrices in a natural way to obtain new quadratic forms from old ones. Namely, if $a, b \in \mathbf{R}$ and $\boldsymbol{A}$ and $\boldsymbol{B}$ are real and symmetric $n \times n$ matrices, then $Q_{a\boldsymbol{A}+b\boldsymbol{B}}(\boldsymbol{x}) = aQ_{\boldsymbol{A}}(\boldsymbol{x}) + bQ_{\boldsymbol{B}}(\boldsymbol{x})$. For instance, suppose $\boldsymbol{A} := \boldsymbol{I}_{n\times n}$ and $\boldsymbol{B} := \mathbf{1}_{n\times n}$. Then,

$$a\boldsymbol{A} + b\boldsymbol{B} = \begin{pmatrix} a+b & b & b & \cdots & b \\ b & a+b & b & \cdots & b \\ b & b & a+b & \cdots & b \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ b & b & b & \cdots & a+b \end{pmatrix},$$

and, thanks to the preceding two examples,

$$Q_{a\boldsymbol{A}+b\boldsymbol{B}}(\boldsymbol{x}) = a\sum_{i=1}^n x_i^2 + b\left(\sum_{i=1}^n x_i\right)^2 .$$

An important special case is when $a := 1$ and $b := -1/n$. In that case,

$$\boldsymbol{A} - \frac{1}{n}\boldsymbol{B} = \begin{pmatrix} 1-{}^1\!/_n & -{}^1\!/_n & -{}^1\!/_n & \cdots & -{}^1\!/_n \\ -{}^1\!/_n & 1-{}^1\!/_n & -{}^1\!/_n & \cdots & -{}^1\!/_n \\ -{}^1\!/_n & -{}^1\!/_n & 1-{}^1\!/_n & \cdots & -{}^1\!/_n \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -{}^1\!/_n & -{}^1\!/_n & -{}^1\!/_n & \cdots & 1-{}^1\!/_n \end{pmatrix},$$

and

$$Q_{\boldsymbol{A}-(1/n)\boldsymbol{B}}(\boldsymbol{x}) = \sum_{i=1}^n x_i^2 - \frac{1}{n}\left(\sum_{i=1}^n x_i\right)^2 = \sum_{i=1}^n (x_i - \bar{x})^2.$$

Note that

$$\mathrm{tr}(\boldsymbol{A\Sigma}) = \sum_{i=1}^n \mathrm{Var}(X_i) + \frac{1}{n}\sum_{i=1}^n\sum_{j=1}^n \mathrm{Cov}(X_i\,, X_j).$$

Consider the special case that the X_i's are uncorrelated. In that case, $\mathrm{tr}(\boldsymbol{A\Sigma}) = (1 - {}^1\!/_n)\sum_{i=1}^n \mathrm{Var}(X_i)$, and hence

$$\mathrm{E}\left[\sum_{i=1}^n (X_i - \bar{X})^2\right] = (1 - {}^1\!/_n)\sum_{i=1}^n \mathrm{Var}(X_i) + \sum_{i=1}^n (\mu_i - \bar{\mu})^2.$$

When the X_i's are i.i.d. this yields $\mathrm{E}\sum_{i=1}^n (X_i - \bar{X})^2 = (n-1)\mathrm{Var}(X_1)$, which is a formula that you have seen in the context of the unbiasedness of the sample variance estimator $S^2 := (n-1)^{-1}\sum_{i=1}^n (X_i - \bar{X})^2$. □

Example 6. Consider a symmetric matrix of the form

$$\boldsymbol{A} := \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 & 0 \\ 1 & 0 & 1 & \cdots & 0 & 0 \\ 0 & 1 & 0 & & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & 1 \\ 0 & 0 & 0 & \cdots & 1 & 0 \end{pmatrix}.$$

That is, the super- and sub-diagonal entires are all ones and all other entries are zeros. Then,

$$Q_{\boldsymbol{A}}(\boldsymbol{x}) = 2\sum_{i=1}^{n-2} x_i x_{i+2}.$$

Other examples can be constructed in this way as well, and by also combining such examples. □

3. The variance of a random quadratic form

In the previous section we computed the expectation of $\boldsymbol{X}'\boldsymbol{A}\boldsymbol{X}$ where $\boldsymbol{X}$ is a random vector. Here let us say a few things about the variance of the same random vector, under some conditions on $\boldsymbol{X}$.

Proposition 7. *Suppose $\boldsymbol{X} := (X_1, \ldots, X_n)'$ where the X_j's are i.i.d. with mean zero and four finite moments. Then,*

$$\text{Var}\left(\boldsymbol{X}'\boldsymbol{A}\boldsymbol{X}\right) = \left(\mu_4 - 3\mu_2^2\right)\sum_{i=1}^{n} A_{i,i}^2 + \left(\mu_2^2 - 1\right)\left(\text{tr}(\boldsymbol{A})\right)^2 + 2\mu_2^2 \text{tr}\left(\boldsymbol{A}^2\right),$$

where $\mu_2 := \text{E}(X_1^2)$ and $\mu_4 := \text{E}(X_1^4)$.

One can generalize this a little more as well, with more or less the same set of techniques, in order to compute the variance of $\boldsymbol{X}'\boldsymbol{A}\boldsymbol{X}$ in the case that the X_i's are independent, with common first four moments, and not necessarily mean zero.

Proof. Suppose $\boldsymbol{X} := (X_1, \ldots, X_n)'$, where $X_1, \ldots, X_n$ are independent and mean zero. Suppose $\mu_2 := \text{E}(X_i^2)$ and $\mu_4 := \text{E}(X_i^4)$ do not depend on i [e.g., because the X_j's are independent]. Then we can write

$$\left(\boldsymbol{X}'\boldsymbol{A}\boldsymbol{X}\right)^2 = \sum\sum\sum\sum_{1\le i,j,k,\ell\le n} A_{i,j}A_{k,\ell}X_iX_jX_kX_\ell.$$

Note that

$$\mathrm{E}\left(X_iX_jX_kX_\ell\right) = \begin{cases} \mu_4 & \text{of } i=j=k=\ell, \\ \mu_2^2 & \text{if } i=j\neq k=\ell \text{ or} \\ & \text{if } i=k\neq j=\ell \text{ or} \\ & \text{if } i=\ell\neq k=j, \\ 0 & \text{otherwise.} \end{cases}$$

Therefore,

$$\begin{aligned}\mathrm{E}\left[\left(\boldsymbol{X}'\boldsymbol{A}\boldsymbol{X}\right)^2\right] &= \sum_{i=1}^n A_{i,i}^2\ \mu_4 + \sum\sum_{1\le i\neq k\le n} A_{i,i}A_{k,k}\ \mu_2^2 + \sum\sum_{1\le i\neq j\le n} A_{i,j}A_{j,i}\ \mu_2^2 + \sum\sum_{1\le i\neq k\le n} A_{i,k}A_{k,i}\ \mu_2^2 \\ &= \mu_4\sum_{i=1}^n A_{i,i}^2 + \mu_2^2\left[\sum\sum_{1\le i\neq k\le n} A_{i,i}A_{k,k} + 2\sum\sum_{1\le i\neq j\le n} A_{i,j}^2\right].\end{aligned}$$

Next, we identify the double sums in turn:

$$\begin{aligned}\sum\sum_{1\le i\neq k\le n} A_{i,i}A_{k,k} &= \sum_{i=1}^n A_{i,i}\sum_{k=1}^n A_{k,k} - \sum_{i=1}^n A_{i,i}^2 = (\mathrm{tr}(\boldsymbol{A}))^2 - \sum_{i=1}^n A_{i,i}^2, \\ \sum\sum_{1\le i\neq j\le n} A_{i,j}^2 &= \sum_{i=1}^n\sum_{j=1}^n A_{i,j}^2 - \sum_{i=1}^n A_{i,i}^2 = \sum_{i=1}^n\sum_{j=1}^n A_{i,j}A_{j,i} - \sum_{i=1}^n A_{i,i}^2 \\ &= \sum_{i=1}^n (A^2)_{i,i} - \sum_{i=1}^n A_{i,i}^2 = \mathrm{tr}\left(\boldsymbol{A}^2\right) - \sum_{i=1}^n A_{i,i}^2.\end{aligned}$$

Consequently,

$$\begin{aligned}\mathrm{E}\left[\left(\boldsymbol{X}'\boldsymbol{A}\boldsymbol{X}\right)^2\right] &= \mu_4\sum_{i=1}^n A_{i,i}^2 + \mu_2^2\left[(\mathrm{tr}(\boldsymbol{A}))^2 - \sum_{i=1}^n A_{i,i}^2 + 2\mathrm{tr}\left(\boldsymbol{A}^2\right) - 2\sum_{i=1}^n A_{i,i}^2\right] \\ &= \left(\mu_4 - 3\mu_2^2\right)\sum_{i=1}^n A_{i,i}^2 + \mu_2^2\left[(\mathrm{tr}(\boldsymbol{A}))^2 + 2\mathrm{tr}\left(\boldsymbol{A}^2\right)\right].\end{aligned}$$

Therefore, in this case,

$$\mathrm{Var}\left(\boldsymbol{X}'\boldsymbol{A}\boldsymbol{X}\right) = \left(\mu_4 - 3\mu_2^2\right)\sum_{i=1}^n A_{i,i}^2 + \mu_2^2\left[(\mathrm{tr}(\boldsymbol{A}))^2 + 2\mathrm{tr}\left(\boldsymbol{A}^2\right)\right] - \left[\mathrm{E}\left(\boldsymbol{X}'\boldsymbol{A}\boldsymbol{X}\right)\right]^2.$$

This proves the result because $\mathrm{E}(\boldsymbol{X}'\boldsymbol{A}\boldsymbol{X}) = \mathrm{tr}(\boldsymbol{A})$. □

Moment-generating functions and independence

Let $\boldsymbol{X} := (X_1\,, \ldots, X_n)'$ be a random vector. Its *moment generating function* [written MGF for short] $M_{\boldsymbol{X}}$ is defined as

$$M_{\boldsymbol{X}}(\boldsymbol{t}) := \mathrm{E}e^{\boldsymbol{t}'\boldsymbol{X}} \qquad (\boldsymbol{t} \in \mathbf{R}^n).$$

It is the case that $M_{\boldsymbol{X}}(\boldsymbol{t})$ is a well-defined quantity, but it might be infinite for some, and even all, values of $\boldsymbol{t} \in \mathbf{R}^n$. The following is a hard fact from classical analysis:

Theorem 1 (Uniqueness theorem of MGFs)**.** *Suppose there exists $t > 0$ such that $M_{\boldsymbol{X}}(\boldsymbol{t}) < \infty$ for all $\boldsymbol{t} \in \mathbf{R}^n$ with $\|\boldsymbol{t}\| \le r$. Then, the distribution of $\boldsymbol{X}$ is determined uniquely by the function $M_{\boldsymbol{X}}$. That is, if $\boldsymbol{Y}$ is any random vector whose MGF is the same as $M_{\boldsymbol{X}}$, then $\boldsymbol{Y}$ has the same distribution as $\boldsymbol{X}$.*

We are interested in examples, and primarily those that involve normal distributions in one form or another.

Example 2. If $X \sim \mathrm{N}(\mu\,,\sigma^2)$, then

$$\begin{aligned} M_X(t) = \mathrm{E}e^{tX} &= \int_{-\infty}^{\infty} e^{tx}\, \frac{e^{-(x-\mu)^2/(2\sigma^2)}}{\sqrt{2\pi\sigma^2}}\,\mathrm{d}x \\ &= \int_{-\infty}^{\infty} e^{t(\sigma y+\mu)}\, \frac{e^{-y^2/2}}{\sqrt{2\pi}}\,\mathrm{d}y \qquad\qquad (y := (x-\mu)/\sigma) \\ &= e^{t\mu}\int_{-\infty}^{\infty} \frac{e^{t\sigma y - y^2/2}}{\sqrt{2\pi}} = e^{t\mu}\int_{-\infty}^{\infty} \frac{e^{-\frac12[y^2-2yt\sigma]}}{\sqrt{2\pi}}\,\mathrm{d}y. \end{aligned}$$

We complete the square $[y^2 - 2yt\sigma = (y-t\sigma)^2 - (t\sigma)^2]$ in order to see that

$$M_X(t) = \exp\left(t\mu + \frac{t^2\sigma^2}{2}\right).$$

Therefore, the uniqueness theorem (Theorem 1) tells us that any random variable Y whose MGF is $M_Y(t) = \exp(t\mu + \frac12 t^2\sigma^2)$ is distributed according to $\mathrm{N}(\mu\,,\sigma^2)$. □

Example 3 (MGF of a simple multivariable normal)**.** Suppose $X_i \sim \mathrm{N}(\mu_i\,,\sigma_i^2)$ $(1 \le i \le n)$ are independent. Then, the MGF of $\boldsymbol{X} := (X_1\,,\ldots,X_n)'$ is

$$M_{\boldsymbol{X}}(\boldsymbol{t}) = \mathrm{E}e^{\boldsymbol{t}'\boldsymbol{X}} = \prod_{j=1}^n \mathrm{E}e^{t_jX_j} = \prod_{j=1}^n e^{t_j\mu_j + \frac12 t_j^2\sigma_j^2} = \exp\left(\sum_{j=1}^n t_j\mu_j + \frac12\sum_{j=1}^n t_j^2\sigma_j^2\right)$$

for all $\boldsymbol{t} \in \mathbf{R}^n$. □

Example 4 (MGF of χ_1^2)**.** If X is standard normal, then $Y := X^2 \sim \chi_1^2$. Now

$$\begin{aligned} M_Y(t) = \mathrm{E}e^{tX^2} &= \int_{-\infty}^{\infty} \frac{e^{tx^2 - x^2/2}}{\sqrt{2\pi}}\,\mathrm{d}x \\ &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} \exp\left(-\frac12(1-2t)x^2\right)\,\mathrm{d}x. \end{aligned}$$

If $t \ge 1/2$, then the preceding is infinite. Otherwise, a change of variables $[y = \sqrt{1-2t}\,x]$ tells us that it is equal to

$$\int_{-\infty}^{\infty} \frac{e^{-y^2/2}}{\sqrt{2\pi}}\,\frac{\mathrm{d}y}{\sqrt{1-2t}} = \frac{1}{\sqrt{1-2t}}.$$

In other words, $M_{X^2}(t) = \infty$ if $t \ge 1/2$ and $M_{X^2}(t) = (1-2t)^{-1/2}$ if $t < 1/2$. □

Example 5 (MGF of χ_n^2)**.** Let $X_1,\ldots,X_n \sim \mathrm{N}(0\,,1)$ be independent, and consider the χ_n^2 random variable $Y := \sum_{i=1}^n X_i^2$. Its MGF is

$$M_Y(t) = \prod_{j=1}^n M_{X_j^2}(t) = \begin{cases} (1-2t)^{-n/2} & \text{if } t < 1/2, \\ \infty & \text{if } t \ge 1/2. \end{cases}$$

According to Theorem 1, this is a formula for the MGF of the χ^2_n distribution, and identifies that distribution uniquely. □

Theorem 6 (Independence theorem of MGFs)**.** *Let $\boldsymbol{X}$ be a random n-vector with a MGF that is finite in an open neighborhood of the origin $\mathbf{0} \in \mathbf{R}^n$. Suppose there exists $r = 1, \ldots, n$ such that*

$$M_{\boldsymbol{X}}(\boldsymbol{t}) = M_{\boldsymbol{X}}(t_1\,, \ldots, t_r\,, 0\,, \ldots, 0) \cdot M_{\boldsymbol{X}}(0\,, \ldots, 0\,, t_{r+1}\,, \ldots, t_n)$$

for all $\boldsymbol{t} \in \mathbf{R}^n$. Then, $(X_1\,, \ldots, X_r)$ and $(X_{r+1}\,, \ldots, X_n)$ are independent.

Proof. Let $\tilde{\boldsymbol{X}}$ denote an independent copy of $\boldsymbol{X}$. Define a new random vector $\boldsymbol{Y}$ as follows:

$$\boldsymbol{Y} := \begin{pmatrix} X_1 \\ \vdots \\ X_r \\ \tilde{X}_{r+1} \\ \vdots \\ \tilde{X}_n \end{pmatrix}.$$

Then,

$$M_{\boldsymbol{Y}}(\boldsymbol{t}) = M_{\boldsymbol{X}}(t_1\,, \ldots, t_r\,, 0\,, \ldots, 0) \cdot M_{\boldsymbol{X}}(0\,, \ldots, 0\,, t_{r+1}\,, \ldots, t_n).$$

According to the condition of this theorem, $\boldsymbol{X}$ and $\boldsymbol{Y}$ have the same MGF's, and therefore they have the same distribution (Theorem 1). That is, for all sets $A_1, \ldots, A_n$,

$$\mathrm{P}\left\{X_1 \in A_1\,, \ldots, X_n \in A_n\right\} = \mathrm{P}\left\{Y_1 \in A_1\,, \ldots, Y_n \in A_n\right\},$$

which is, by construction equal to

$$\mathrm{P}\left\{X_1 \in A_1\,, \ldots, X_r \in A_r\right\} \cdot \mathrm{P}\left\{\tilde{X}_{r+1} \in A_{r+1}\,, \ldots, \tilde{X}_n \in A_n\right\}.$$

Since $\tilde{\boldsymbol{X}}$ has the same distribution as $\boldsymbol{X}$, this proves that

$$\mathrm{P}\left\{X_1 \in A_1\,, \ldots, X_n \in A_n\right\} = \mathrm{P}\left\{X_1 \in A_1\,, \ldots, X_r\right\} \cdot \mathrm{P}\left\{X_{r+1} \in A_{r+1}\,, \ldots, X_n \in A_n\right\},$$

which has the desired result. □

Gaussian Random Vectors

1. The multivariate normal distribution

Let $\boldsymbol{X} := (X_1\,, \ldots, X_n)'$ be a random vector. We say that $\boldsymbol{X}$ is a *Gaussian random vector* if we can write

$$\boldsymbol{X} = \boldsymbol{\mu} + \boldsymbol{A}\boldsymbol{Z},$$

where $\boldsymbol{\mu} \in \mathbf{R}^n$, $\boldsymbol{A}$ is an $n \times k$ matrix and $\boldsymbol{Z} := (Z_1\,, \ldots, Z_k)'$ is a k-vector of i.i.d. standard normal random variables.

Proposition 1. *Let $\boldsymbol{X}$ be a Gaussian random vector, as above. Then,*

$$\mathrm{E}\boldsymbol{X} = \boldsymbol{\mu}, \ \mathrm{Var}(\boldsymbol{X}) := \boldsymbol{\Sigma} = \boldsymbol{A}\boldsymbol{A}', \ \textit{and } M_{\boldsymbol{X}}(\boldsymbol{t}) = \mathrm{e}^{\boldsymbol{t}'\boldsymbol{\mu} + \frac{1}{2}\|\boldsymbol{A}'\boldsymbol{t}\|^2} = \mathrm{e}^{\boldsymbol{t}'\boldsymbol{\mu} + \frac{1}{2}\boldsymbol{t}'\boldsymbol{\Sigma}\boldsymbol{t}},$$

for all $\boldsymbol{t} \in \mathbf{R}^n$.

Thanks to the uniqueness theorem of MGF's it follows that the distribution of $\boldsymbol{X}$ is determined by $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$, and the fact that it is multivariate normal. From now on, we sometimes write $\boldsymbol{X} \sim \mathrm{N}_n(\boldsymbol{\mu}\,, \boldsymbol{\Sigma})$, when we mean that $M_{\boldsymbol{X}}(\boldsymbol{t}) = \exp(\boldsymbol{t}'\boldsymbol{\mu} + \frac{1}{2}\boldsymbol{t}'\boldsymbol{\Sigma}\boldsymbol{t})$. Interesetingly enough, the choice of $\boldsymbol{A}$ and $\boldsymbol{Z}$ are typically not unique; only $(\boldsymbol{\mu}\,, \boldsymbol{\Sigma})$ influences the distribution of $\boldsymbol{X}$.

Proof. The expectation of $\boldsymbol{X}$ is $\boldsymbol{\mu}$, since $\mathrm{E}(\boldsymbol{A}\boldsymbol{Z}) = \boldsymbol{A}\mathrm{E}(\boldsymbol{Z}) = \boldsymbol{0}$. Also,

$$\mathrm{E}(\boldsymbol{X}\boldsymbol{X}') = \mathrm{E}\left([\boldsymbol{\mu} + \boldsymbol{A}\boldsymbol{Z}]\,[\boldsymbol{\mu} + \boldsymbol{A}\boldsymbol{Z}]'\right) = \boldsymbol{\mu}\boldsymbol{\mu}' + \boldsymbol{A}\mathrm{E}(\boldsymbol{Z}\boldsymbol{Z}')\boldsymbol{A}'.$$

Since $\mathrm{E}(\boldsymbol{Z}\boldsymbol{Z}') = \boldsymbol{I}$, the variance-covariance of $\boldsymbol{X}$ is $\mathrm{E}(\boldsymbol{X}\boldsymbol{X}') - (\mathrm{E}\boldsymbol{X})(\mathrm{E}\boldsymbol{X})' = \mathrm{E}(\boldsymbol{X}\boldsymbol{X}') - \boldsymbol{\mu}\boldsymbol{\mu}' = \boldsymbol{A}\boldsymbol{A}'$, as desired. Finally, note that $M_{\boldsymbol{X}}(\boldsymbol{t}) = \exp(\boldsymbol{t}'\boldsymbol{\mu}) \cdot M_{\boldsymbol{Z}}(\boldsymbol{A}'\boldsymbol{t})$. This establishes the result on the MGF of $\boldsymbol{X}$, since $M_{\boldsymbol{Z}}(\boldsymbol{s}) = \prod_{j=1}^n \exp(s_j^2/2) = \exp(\frac{1}{2}\|\boldsymbol{s}\|^2)$ for all $\boldsymbol{s} \in \mathbf{R}^n$. □

We say that $\boldsymbol{X}$ has the *multivariate normal distribution* with parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma} := \boldsymbol{A}\boldsymbol{A}'$, and write this as $\boldsymbol{X} \sim \mathrm{N}_n(\boldsymbol{\mu}\,, \boldsymbol{A}\boldsymbol{A}')$.

Theorem 2. *$\boldsymbol{X} := (X_1\,, \ldots, X_n)'$ has a multivariate normal distribution if and only if $\boldsymbol{t}'\boldsymbol{X} = \sum_{i=1}^n t_i X_i$ has a normal distribution on the line for every $\boldsymbol{t} \in \mathbf{R}^n$. That is, $X_1, \ldots, X_n$ are jointly normally distributed if and only if all of their linear combinations are normally distributed.*

Note that the distribution of $\boldsymbol{X}$ depends on $\boldsymbol{A}$ only through the positive semidefinite $n \times n$ matrix $\boldsymbol{\Sigma} := \boldsymbol{A}\boldsymbol{A}'$. Sometimes we say also that $X_1, \ldots, X_n$ are *jointly* normal [or Gaussian] when $\boldsymbol{X} := (X_1\,, \ldots, X_n)'$ has a multivariate normal distribution.

Proof. If $\boldsymbol{X} \in \mathrm{N}_n(\boldsymbol{\mu}\,, \boldsymbol{A}\boldsymbol{A}')$ then we can write it as $\boldsymbol{X} = \boldsymbol{\mu} + \boldsymbol{A}\boldsymbol{Z}$, we as before. In that case, $\boldsymbol{t}'\boldsymbol{X} = \boldsymbol{t}'\boldsymbol{\mu} + \boldsymbol{t}'\boldsymbol{A}\boldsymbol{Z}$ is a linear combination of $Z_1, \ldots, Z_k$, whence has a normal distribution with mean $t_1\mu_1 + \cdots + t_n\mu_n$ and variance $\boldsymbol{t}'\boldsymbol{A}\boldsymbol{A}'\boldsymbol{t} = \|\boldsymbol{A}'\boldsymbol{t}\|^2$.

For the converse, suppose that $\boldsymbol{t}'\boldsymbol{X}$ has a normal distribution for every $\boldsymbol{t} \in \mathbf{R}^n$. Let $\boldsymbol{\mu} := \mathrm{E}\boldsymbol{X}$ and $\boldsymbol{\Sigma} := \mathrm{Var}(\boldsymbol{X})$, and observe that $\boldsymbol{t}'\boldsymbol{X}$ has mean vector $\boldsymbol{t}'\boldsymbol{\mu}$ and variance-covariance matrix $\boldsymbol{t}'\boldsymbol{\Sigma}\boldsymbol{t}$. Therefore, the MGF of the univariate normal $\boldsymbol{t}'\boldsymbol{X}$ is $M_{\boldsymbol{t}'\boldsymbol{X}}(s) = \exp(s\boldsymbol{t}'\boldsymbol{\mu} + \frac{1}{2}s^2\boldsymbol{t}'\boldsymbol{\Sigma}\boldsymbol{t})$ for all $s \in \mathbf{R}$. Note that $M_{\boldsymbol{t}'\boldsymbol{X}}(s) = \mathrm{E}\exp(s\boldsymbol{t}'\boldsymbol{X})$. Therefore, apply this with $s := 1$ to see that $M_{\boldsymbol{t}'\boldsymbol{X}}(1) = M_{\boldsymbol{X}}(\boldsymbol{t})$ is the MGF of a multivariate normal. The uniqueness theorem for MGF's (Theorem 1, p. 27) implies the result. □

2. The nondegenerate case

Suppose $\boldsymbol{X} \sim \mathrm{N}_n(\boldsymbol{\mu}\,, \boldsymbol{\Sigma})$, and recall that $\boldsymbol{\Sigma}$ is always positive semidefinite. We say that $\boldsymbol{X}$ is *nondegenerate* when Σ is positive definite (equivalently, invertible).

Take, in particular, $X \sim \mathrm{N}_1(\mu\,, \Sigma)$; μ can be any real number and Σ is a positive semidefinite 1×1 matrix; i.e., $\Sigma \geq 0$. The distribution of X is defined via its MGF as

$$M_X(t) = \mathrm{e}^{t\mu + \frac{1}{2}t^2\Sigma}.$$

When X is nondegenerate ($\Sigma > 0$), $X \sim \mathrm{N}(\mu\,, \Sigma)$. If $\Sigma = 0$, then $M_X(t) = \exp(t\mu)$; therefore by the uniqueness theorem of MGFs, $\mathrm{P}\{X = \mu\} = 1$. Therefore, $\mathrm{N}_1(\mu\,, \sigma^2)$ is the generalization of $\mathrm{N}(\mu\,, \sigma^2)$ in order to include the case that $\sigma = 0$. We will not write $\mathrm{N}_1(\mu\,, \sigma^2)$; instead we always write $\mathrm{N}(\mu\,, \sigma^2)$ as no confusion should arise.

Theorem 3. *$\boldsymbol{X} \sim \mathrm{N}_n(\boldsymbol{\mu}\,, \boldsymbol{\Sigma})$ has a probability density function if and only if it is nondegenerate. In that case, the pdf of $\boldsymbol{X}$ is*

$$f_{\boldsymbol{X}}(\boldsymbol{a}) = \frac{1}{(2\pi)^{n/2}\,(\det\boldsymbol{\Sigma})^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{a} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{a} - \boldsymbol{\mu})\right)$$

for all $\boldsymbol{a} \in \mathbf{R}^n$.

Proof. First of all let us consider the case that $\boldsymbol{X}$ is degenerate. In that case $\boldsymbol{\Sigma}$ has some number $k < n$ of strictly-positive eigenvalues. The proof of Theorem 2 tells us that we can write $\boldsymbol{X} = \boldsymbol{AZ} + \boldsymbol{\mu}$, where $\boldsymbol{Z}$ is a k-dimensional vector of i.i.d. standard normals and $\boldsymbol{A}$ is an $n \times k$ matrix. Consider the k-dimensional space

$$\boldsymbol{E} := \left\{ \boldsymbol{x} \in \mathbf{R}^n : \ \boldsymbol{x} = \boldsymbol{Az} + \boldsymbol{\mu} \text{ for some } \boldsymbol{z} \in \mathbf{R}^k \right\}.$$

Because $\mathrm{P}\{\boldsymbol{Z} \in \mathbf{R}^k\} = 1$, it follows that $\mathrm{P}\{\boldsymbol{X} \in \boldsymbol{E}\} = 1$. If $\boldsymbol{X}$ had a pdf $f_{\boldsymbol{X}}$, then

$$1 = \mathrm{P}\{\boldsymbol{X} \in \boldsymbol{E}\} = \int_{\boldsymbol{E}} f_{\boldsymbol{X}}(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x}.$$

But the n-dimensional volume of $\boldsymbol{E}$ is zero since the dimension of $\boldsymbol{E}$ is $k < n$. This creates a contradiction [unless $\boldsymbol{X}$ did not have a pdf, that is].

If $\boldsymbol{X}$ is nondegenerate, then we can write $\boldsymbol{X} = \boldsymbol{AZ} + \boldsymbol{\mu}$, where $\boldsymbol{Z}$ is an n-vector of i.i.d. standard normals and $\boldsymbol{\Sigma} = \boldsymbol{AA}'$ is invertible; see the proof of Theorem 2. Recall that the choice of $\boldsymbol{A}$ is not unique; in this case, we can always choose $\boldsymbol{A} := \boldsymbol{\Sigma}^{1/2}$ because $\boldsymbol{\Sigma}^{1/2}\boldsymbol{Z} + \boldsymbol{\mu} \sim \mathrm{N}_n(\boldsymbol{\mu}\,, \boldsymbol{\Sigma})$. In other words,

$$X_i = \sum_{j=1}^n A_{i,j} Z_j + \mu_i = \sum_{j=1}^n \Sigma_{i,j}^{1/2} Z_j + \mu_i := g_i(Z_1\,, \ldots, Z_n) \qquad (1 \le i \le n).$$

If $\boldsymbol{a} = \boldsymbol{\Sigma}^{1/2}\boldsymbol{z} + \boldsymbol{\mu}$, then $\boldsymbol{z} = \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{a} - \boldsymbol{\mu})$. Therefore, the change of variables formula of elementary probability implies that

$$f_{\boldsymbol{X}}(\boldsymbol{a}) = \frac{f_{\boldsymbol{Z}}\left(\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{a} - \boldsymbol{\mu})\right)}{|\det J|},$$

as long as $\det J \neq 0$, where

$$J := \begin{pmatrix} \frac{\partial g_1}{\partial z_1} & \cdots & \frac{\partial g_1}{\partial z_n} \\ \vdots & & \vdots \\ \frac{\partial g_n}{\partial z_1} & \cdots & \frac{\partial g_n}{\partial z_n} \end{pmatrix} = \begin{pmatrix} A_{1,1} & \cdots & A_{1,n} \\ \vdots & & \vdots \\ A_{n,1} & \cdots & A_{n,n} \end{pmatrix} = \boldsymbol{A}.$$

Because $\det(\boldsymbol{\Sigma}) = \det(\boldsymbol{AA}') = (\det \boldsymbol{A})^2$, it follows that $\det \boldsymbol{A} = (\det \boldsymbol{\Sigma})^{1/2}$, and hence

$$f_{\boldsymbol{X}}(\boldsymbol{a}) = \frac{1}{(\det \boldsymbol{\Sigma})^{1/2}}\, f_{\boldsymbol{Z}}\left(\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{a} - \boldsymbol{\mu})\right).$$

Because of the independence of the Z_j's,

$$f_{\boldsymbol{Z}}(\boldsymbol{z}) = \prod_{j=1}^n \frac{\mathrm{e}^{-z_j^2/2}}{\sqrt{2\pi}} = \frac{1}{(2\pi)^{n/2}} \mathrm{e}^{-\boldsymbol{z}'\boldsymbol{z}/2}$$

for all $\boldsymbol{z} \in \mathbf{R}^n$. Therefore,

$$f_{\boldsymbol{Z}}\left(\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{a} - \boldsymbol{\mu})\right) = \frac{1}{(2\pi)^{n/2}} \exp\left(-\frac{1}{2}(\boldsymbol{a} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{a} - \boldsymbol{\mu})\right),$$

and the result follows. □

3. The bivariate normal distribution

A *bivariate normal distribution* has the form $\mathrm{N}_2(\boldsymbol{\mu}\,,\boldsymbol{\Sigma})$, where $\mu_1 = \mathrm{E}X_1$, $\mu_2 = \mathrm{E}X_2$, $\Sigma_{1,1} = \mathrm{Var}(X_1) := \sigma_1^2 > 0$, $\Sigma_{2,2} = \mathrm{Var}(X_2) := \sigma_2^2 > 0$, and $\Sigma_{1,2} = \Sigma_{2,1} = \mathrm{Cov}(X_1\,,X_2)$. Let

$$\rho := \mathrm{Corr}(X_1\,,X_2) := \frac{\mathrm{Cov}(X_1\,,X_2)}{\sqrt{\mathrm{Var}(X_1)\cdot\mathrm{Var}(X_2)}}$$

denote the correlation between X_1 and X_2, and recall that $-1 \le \rho \le 1$. Then, $\Sigma_{1,2} = \Sigma_{2,1} = \rho\sigma_1\sigma_2$, whence

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}.$$

Since $\det\boldsymbol{\Sigma} = \sigma_1^2\sigma_2^2(1-\rho^2)$, it follows immediately that our bivariate normal distribution is non-degenerate if and only if $-1 < \rho < 1$, in which case

$$\boldsymbol{\Sigma}^{-1} = \begin{pmatrix} \dfrac{1}{\sigma_1^2(1-\rho^2)} & -\dfrac{\rho}{1-\rho^2}\cdot\dfrac{1}{\sigma_1\sigma_2} \\[2ex] -\dfrac{\rho}{1-\rho^2}\cdot\dfrac{1}{\sigma_1\sigma_2} & \dfrac{1}{\sigma_2^2(1-\rho^2)} \end{pmatrix}.$$

Because

$$z'\boldsymbol{\Sigma}^{-1}z = \left(\frac{z_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{z_1}{\sigma_1}\right)\left(\frac{z_2}{\sigma_2}\right) + \left(\frac{z_2}{\sigma_2}\right)^2$$

for all $\boldsymbol{z} \in \mathbf{R}^n$, the pdf of $\boldsymbol{X} = (X_1\,,X_2)'$—in the non-degenerate case where there *is* a pdf—is

$$f_{\boldsymbol{X}}(x_1\,,x_2)$$
$$= \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}\exp\left(-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{x_1-\mu_1}{\sigma_1}\right)\left(\frac{x_2-\mu_2}{\sigma_2}\right) + \left(\frac{x_2-\mu_2}{\sigma_2}\right)^2\right]\right).$$

But of course non-degenerate cases are also possible. For instance, suppose $Z \sim \mathrm{N}(0\,,1)$ and define $\boldsymbol{X} := (Z\,,-Z)$. Then $\boldsymbol{X} = \boldsymbol{A}\boldsymbol{Z}$ where $\boldsymbol{A} := (1\,,-1)'$, whence

$$\boldsymbol{\Sigma} = \boldsymbol{A}\boldsymbol{A}' = \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}$$

is singular. In general, if $\boldsymbol{X} \sim \mathrm{N}_n(\boldsymbol{\mu}\,,\boldsymbol{\Sigma})$ and the rank of $\boldsymbol{\Sigma}$ is $k < n$, then $\boldsymbol{X}$ depends only on k [and not n] i.i.d. $\mathrm{N}(0\,,1)$'s. This can be gleaned from the proof of Theorem 2.

4. A few important properties of multivariate normal distributions

Proposition 4. *Let $\boldsymbol{X} \sim \mathrm{N}_n(\boldsymbol{\mu}\,,\Sigma)$. If $\boldsymbol{C}$ is an $m \times n$ matrix and $\boldsymbol{d}$ is an m-vector, then $\boldsymbol{CX}+\boldsymbol{d} \sim \mathrm{N}_m(\boldsymbol{C\mu}+\boldsymbol{d}\,,\boldsymbol{C\Sigma C}')$. In general, $\boldsymbol{C\Sigma C}'$ is positive semidefinite; it is positive definite if and only if it has full rank m.*

In particular, if $\boldsymbol{a}$ is a nonrandom n-vector, then $\boldsymbol{a}'\boldsymbol{X} \sim \mathrm{N}(\boldsymbol{a}'\boldsymbol{\mu}\,,\boldsymbol{a}'\boldsymbol{\Sigma a})$.

Proof. We compute the MGF of $\boldsymbol{CX}+\boldsymbol{d}$ as follows:

$$M_{\boldsymbol{CX}+\boldsymbol{d}}(\boldsymbol{t}) = \mathrm{E}\exp\left(\boldsymbol{t}'\left[\boldsymbol{CX}+\boldsymbol{d}\right]\right) = \mathrm{e}^{\boldsymbol{t}'\boldsymbol{d}} M_{\boldsymbol{X}}(\boldsymbol{s}),$$

where $\boldsymbol{s} := \boldsymbol{C}'\boldsymbol{t}$. Therefore,

$$M_{\boldsymbol{CX}+\boldsymbol{d}}(\boldsymbol{t}) = \exp\left(\boldsymbol{t}'\boldsymbol{d}+\boldsymbol{s}'\boldsymbol{\mu}+\frac{1}{2}\boldsymbol{s}'\boldsymbol{\Sigma s}\right) = \exp\left(\boldsymbol{t}'\boldsymbol{\nu}+\frac{1}{2}\boldsymbol{t}'\boldsymbol{Qt}\right),$$

where $\boldsymbol{\nu} := \boldsymbol{C\mu}+\boldsymbol{d}$ and $\boldsymbol{Q} := \boldsymbol{C\Sigma C}'$. Finally, a general fact about symmetric matrices (Corollary 13, p. 16) implies that the symmetric $m \times m$ matrix $\boldsymbol{C\Sigma C}'$ is nonsingular if and only if it has full rank m. □

Proposition 5. *If $\boldsymbol{X} \in \mathrm{N}_n(\boldsymbol{\mu}\,,\boldsymbol{\Sigma})$, for a nonsingular variance-covariance matrix $\boldsymbol{\Sigma}$, and $\boldsymbol{C}_{m\times n}$ and $\boldsymbol{d}_{n\times 1}$ are nonrandom, then $\boldsymbol{CX}+\boldsymbol{d}$ is nonsingular if and only if* $\mathrm{rank}(\boldsymbol{C}) = m$.

Proof. Recall that the nonsingularity of $\boldsymbol{\Sigma}$ is equivalent to it being positive definite. Now $\boldsymbol{CX}+\boldsymbol{d}$ is multivariate normal by the preceding result. It is nondegenerate if and only if $\boldsymbol{C\Sigma C}'$ is positive definite. But $\boldsymbol{x}'\boldsymbol{C\Sigma C}'\boldsymbol{x} = (\boldsymbol{C}'\boldsymbol{x})'\boldsymbol{\Sigma}(\boldsymbol{C}'\boldsymbol{x}) > 0$ if and only if $(\boldsymbol{C}'\boldsymbol{x}) \neq \boldsymbol{0}$, since $\boldsymbol{\Sigma}$ is positive definite. Therefore, $\boldsymbol{CX}+\boldsymbol{d}$ is nondegenerate if and only if $\boldsymbol{C}'\boldsymbol{x} \neq \boldsymbol{0}$ whenever $\boldsymbol{x} \neq \boldsymbol{0}$. This is equivalent to $\boldsymbol{x}'\boldsymbol{C} \neq \boldsymbol{0}$ for all nonzero vectors $\boldsymbol{x}$; that is, $\boldsymbol{C}$ has row rank—hence rank—m. □

The following is an easy corollary of the previous proposition, and identifies the "standard multivariate normal" distribution as the distribution of i.i.d. standard univariate normal distributions. It also states that we do not change the distribution of a standard multivariate normal if we apply to it an orthogonal matrix.

Corollary 6. *$\boldsymbol{Z} \sim \mathrm{N}_n(\boldsymbol{0}\,,\boldsymbol{I})$ if and only if $Z_1,\ldots,Z_n$ are i.i.d. $\mathrm{N}(0\,,1)$'s. Moreover, if $\boldsymbol{Z} \sim \mathrm{N}_n(\boldsymbol{0}\,,\boldsymbol{I})$ and $\boldsymbol{A}_{n\times n}$ is orthogonal then $\boldsymbol{AZ} \sim \mathrm{N}_n(\boldsymbol{0}\,,\boldsymbol{I})$ also.*

Next we state another elementary fact, derived by looking only at the MGF's. It states that a subset of a multivariate normal vector itself is multivariate normal.

Proposition 7. *Suppose $\boldsymbol{X} \sim \mathrm{N}_n(\boldsymbol{\mu}\,,\boldsymbol{\Sigma})$ and $1 \le i_1 < i_2 < \cdots < i_k \le n$ is a subsequence of $1\,,\ldots,n$. Then, $(X_{i_1}\,,\ldots,X_{i_k})' \sim \mathrm{N}_k(\boldsymbol{\nu}\,,\boldsymbol{Q})$, where*

$$\boldsymbol{\nu} := \mathrm{E}\begin{pmatrix} X_{i_1} \\ \vdots \\ X_{i_k}\end{pmatrix} = \begin{pmatrix} \mu_{i_1} \\ \vdots \\ \mu_{i_k}\end{pmatrix}, \qquad \boldsymbol{Q} := \mathrm{Var}\begin{pmatrix} X_{i_1} \\ \vdots \\ X_{i_k}\end{pmatrix} = \begin{pmatrix} \Sigma_{i_1,i_1} & \cdots & \Sigma_{i_1,i_k} \\ \vdots & & \vdots \\ \Sigma_{i_k,i_1} & \cdots & \Sigma_{i_k,i_k}\end{pmatrix}.$$

Proposition 8. *Suppose $\boldsymbol{X} \sim \mathrm{N}_n(\boldsymbol{\mu}\,,\boldsymbol{\Sigma})$, and assume that we can divide the X_i's into two groups: $(X_i)_{i\in G}$ and $(X_j)_{j\notin G}$, where G is a subset of the index set $\{1\,,\ldots,n\}$. Suppose in addition that $\mathrm{Cov}(X_i\,,X_j) = 0$ for all $i \in G$ and $j \notin G$. Then, $(X_i)_{i\in G}$ is independent from $(X_j)_{j\notin G}$.*

Thus, for example, if $(X_1\,,X_2\,,X_3)'$ has a trivariate normal distribution and X_1 is uncorrelated from X_2 and X_3, then X_1 is independent of $(X_2\,,X_3)$. For a second example suppose that $(X_1\,,X_2\,,X_3\,,X_4)$ has a multivariate normal distribution and: $\mathrm{E}(X_1X_2) = \mathrm{E}(X_1)\mathrm{E}(X_2)$, $\mathrm{E}(X_1X_3) = \mathrm{E}(X_1)\mathrm{E}(X_3)$, $\mathrm{E}(X_4X_2) = \mathrm{E}(X_4)\mathrm{E}(X_2)$, and $\mathrm{E}(X_4X_3) = \mathrm{E}(X_4)\mathrm{E}(X_3)$, then $(X_1\,,X_4)$ and $(X_2\,,X_3)$ are two independent bivariate normal random vectors.

Proof. I will prove the following special case of the proposition; the general case follows from a similar reasoning, but the notation is messier.

Suppose $(X_1\,,X_2)$ has a bivariate normal distribution and $\mathrm{E}(X_1X_2) = \mathrm{E}(X_1)\mathrm{E}(X_2)$. Then, X_1 and X_2 are independent. In order to prove this we write the MGF of $\boldsymbol{X} := (X_1\,,X_2)'$:

$$\begin{aligned}
M_{\boldsymbol{X}}(\boldsymbol{t}) &= \mathrm{e}^{\boldsymbol{t}'\boldsymbol{\mu}+\frac{1}{2}\boldsymbol{t}'\boldsymbol{\Sigma}\boldsymbol{t}} \\
&= \mathrm{e}^{t_1\mu_1+t_2\mu_2}\cdot\exp\left(\frac{1}{2}(t_1\,,t_2)\begin{pmatrix}\mathrm{Var}(X_1) & 0 \\ 0 & \mathrm{Var}(X_2)\end{pmatrix}\begin{pmatrix}t_1 \\ t_2\end{pmatrix}\right) \\
&= \mathrm{e}^{t_1\mu_1+\frac{1}{2}t_1^2\mathrm{Var}(X_1)}\cdot\mathrm{e}^{t_2\mu_2+\frac{1}{2}t_2^2\mathrm{Var}(X_2)} \\
&= M_{X_1}(t_1)\cdot M_{X_2}(t_2).
\end{aligned}$$

The result follows from the independence theorem for MGF's (Theorem 6, p. 29). □

Remark 9. The previous proposition has generalizations. For instance, suppose we could decompose $\{1\,,\ldots,n\}$ into k disjoint groups $G_1,\ldots,G_k$ [so $G_i \cap G_j = \varnothing$ if $i \neq j$, and $G_1\cup\cdots\cup G_k = \{1\,,\ldots,n\}$] such that $X_{i_1},\ldots,X_{i_k}$ are [pairwise] uncorrelated for all $i_1 \in G_1,\ldots,i_k \in G_k$. Then, $(X_i)_{i\in G_1},\ldots,(X_i)_{i\in G_k}$ are independent multivariate normal random vectors. The proof is the same as in the case $k = 2$. □

Remark 10. It is important that $\boldsymbol{X}$ has a multivariate normal distribution. For instance, we can construct two standard-normal random variables X and Y, on the same probability space, such that X and Y are uncorrelated but dependent. Here is one way to do this: Let $Y \sim \mathrm{N}(0\,,1)$ and $S = \pm 1$ with probability $1/2$

each. Assume that S and Y are independent, and define $X := S|Y|$. Note that

$$\begin{aligned}\mathrm{P}\{X \le a\} &= \mathrm{P}\{X \le a\,,\ S = 1\} + \mathrm{P}\{X \le a\,,\ S = -1\}\\ &= \frac{1}{2}\mathrm{P}\{|Y| \le a\} + \frac{1}{2}\mathrm{P}\{-|Y| \le a\}.\end{aligned}$$

If $a \ge 0$, then $\mathrm{P}\{X \le a\} = \frac{1}{2}\mathrm{P}\{|Y| \le a\} + \frac{1}{2} = \mathrm{P}\{Y \le a\}$. Similarly, $\mathrm{P}\{X \le a\} = \mathrm{P}\{Y \le a\}$ if $a \le 0$. Therefore, $X, Y \sim \mathrm{N}(0\,,1)$. Furthermore, X and Y are uncorrelated because S has mean zero; here is why: $\mathrm{E}(XY) = \mathrm{E}(SY|Y|) = \mathrm{E}(S)\mathrm{E}(Y|Y|) = 0 = \mathrm{E}(X)\mathrm{E}(Y)$. But X and Y are not independent because $|X| = |Y|$: For instance, $\mathrm{P}\{|X| < 1\} > 0$, but $\mathrm{P}\{|X| < 1 \mid |Y| \ge 2\} = 0$. The problem is [and can only be] that $(X\,,Y)'$ is *not* bivariate normal. □

5. Quadratic forms

Given a multivariate-normal random variable $\boldsymbol{X} \sim \mathrm{N}_n(\boldsymbol{0}\,,\boldsymbol{I})$ and an $n \times n$ positive semidefinite matrix $\boldsymbol{A} := (A_{i,j})$, we can consider the random quadratic form

$$Q_{\boldsymbol{A}}(\boldsymbol{X}) := \boldsymbol{X}'\boldsymbol{A}\boldsymbol{X}.$$

We can write $\boldsymbol{A}$, in spectral form, as $\boldsymbol{A} = \boldsymbol{P}\boldsymbol{D}\boldsymbol{P}'$, so that

$$Q_{\boldsymbol{A}}(\boldsymbol{X}) = \boldsymbol{X}'\boldsymbol{P}\boldsymbol{D}\boldsymbol{P}'\boldsymbol{X}.$$

Since $\boldsymbol{P}$ is orthogonal and $\boldsymbol{X} \sim \mathrm{N}_n(\boldsymbol{0}\,,\boldsymbol{I})$, $\boldsymbol{Z} := \boldsymbol{P}'\boldsymbol{X} \sim \mathrm{N}_n(\boldsymbol{0}\,,\boldsymbol{I})$ as well. Therefore,

$$Q_{\boldsymbol{A}}(\boldsymbol{X}) = \boldsymbol{Z}'\boldsymbol{D}\boldsymbol{Z} = \sum_{i=1}^{n} D_{i,i} Z_i^2.$$

If $\boldsymbol{A}$ is a projection matrix, then all of the $D_{i,i}$'s are ones and zeros. In that case, $Q_{\boldsymbol{A}}(\boldsymbol{X}) \sim \chi^2_r$, where $r :=$ the number of eigenvalues of $\boldsymbol{A}$ that are ones; i.e, $r = \mathrm{rank}(\boldsymbol{A})$. Finally, recall that the rank of a projection matrix is equal to its trace (Corollary 16, p. 16). Let us summarize our findings.

Proposition 11. *If $\boldsymbol{X} \sim \mathrm{N}_n(\boldsymbol{0}\,,\boldsymbol{I})$ and $\boldsymbol{A}$ is a projection matrix, then $\boldsymbol{X}'\boldsymbol{A}\boldsymbol{X} \sim \chi^2_{\mathrm{rank}(\boldsymbol{A})} = \chi^2_{\mathrm{tr}(\boldsymbol{A})} = \chi^2_r$, where $r :=$ the total number of nonzero [i.e., one] eigenvalues of $\boldsymbol{A}$.*

Example 12. Let

$$\boldsymbol{A} := \begin{pmatrix} 1 - 1/n & 1/n & -1/n & \cdots & 1/n \\ -1/n & 1 - 1/n & -1/n & \cdots - 1/n & \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -1/n & -1/n & -1/n & \cdots & 1 - 1/n \end{pmatrix}.$$

Then we have seen (Example 5, p. 23) that

$$\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x} = \sum_{i=1}^{n} (x_i - \bar{x})^2 \qquad \text{for all } \boldsymbol{x} \in \mathbf{R}^n.$$

Now let us observe that $\boldsymbol{A}$ has the form

$$\boldsymbol{A} = \boldsymbol{I} - \boldsymbol{B},$$

where $\boldsymbol{B} := (1/n)\mathbf{1}_{n\times n}$. Note that $\boldsymbol{B}$ is symmetric and $\boldsymbol{B}^2 = \boldsymbol{B}$. Therefore, $\boldsymbol{B}$ is a projection, and hence so is $\boldsymbol{A} = \boldsymbol{I} - \boldsymbol{B}$. Clearly, $\mathrm{tr}(\boldsymbol{A}) = n - 1$. Therefore, Proposition 11 implies the familiar fact that if $X_1, \ldots, X_n$ are i.i.d. standard normals, then $\sum_{i=1}^n (X_i - \bar{X})^2 \sim \chi^2_{n-1}$. □

Example 13. If $\boldsymbol{A}$ is an $n \times n$ projection matrix of rank [or trace] r, then $\boldsymbol{I} - \boldsymbol{A}$ is an $n \times n$ projection matrix of rank [or trace] $n - r$. Therefore, $\boldsymbol{X}'(\boldsymbol{I} - \boldsymbol{A})\boldsymbol{X} \sim \chi^2_{n-r}$, whenever $\boldsymbol{X} \sim \mathrm{N}_n(\mathbf{0}\,, \boldsymbol{I})$. □

Example 14. What is the distribution of $\boldsymbol{X}$ is a nonstandard multivariate normal? Suppose $\boldsymbol{X} \sim \mathrm{N}_n(\boldsymbol{\mu}\,, \boldsymbol{\Sigma})$ and $\boldsymbol{A}$ is a projection matrix. If $\boldsymbol{X}$ is nondegenerate, then $\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{X} - \boldsymbol{\mu}) \sim \mathrm{N}_n(\mathbf{0}\,, \boldsymbol{I})$. Therefore,

$$(\boldsymbol{X} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1/2}\boldsymbol{A}\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{X} - \boldsymbol{\mu}) \sim \chi^2_{\mathrm{rank}(\boldsymbol{A})} = \chi^2_{\mathrm{tr}(\boldsymbol{A})},$$

for every $n \times n$ projection matrix $\boldsymbol{A}$. In particular,

$$(\boldsymbol{X} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{X} - \boldsymbol{\mu}) \sim \chi^2_n,$$

which can be seen by specializing the preceding to the projection matrix $\boldsymbol{A} := \boldsymbol{I}$. Specializing further still, we see that if $X_1, \ldots, X_n$ are independent normal random variables, then we obtain the familiar fact that

$$\sum_{i=1}^n \left(\frac{X_i - \mu_i}{\sigma_i}\right)^2 \sim \chi^2_n,$$

where $\mu_i := \mathrm{E}X_i$ and $\sigma_i^2 := \mathrm{Var}(X_i)$. □

Linear Models

1. The basic model

We now study a *linear statistical model*. That is, we study the models where the observations $\boldsymbol{Y} := (Y_1, \ldots, Y_n)'$ has the following assumed property:

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon},$$

where $\boldsymbol{\beta} := (\beta_0, \beta_1, \ldots, \beta_{p-1})$ is a vector of p unknown parameters, and

$$\boldsymbol{X} := \begin{pmatrix} x_{1,0} & \cdots & x_{1,p-1} \\ \vdots & & \vdots \\ x_{n,0} & \cdots & x_{n,p-1} \end{pmatrix}$$

is the socalled "regression matrix," or "design matrix." The elements of the $n \times p$ matrix $\boldsymbol{X}$ are assumed to be known; these are the "descriptive" or "explanatory" variables, and the randomness of the observed values is inherited from the "noise vector," $\boldsymbol{\varepsilon} := (\varepsilon_1, \ldots, \varepsilon_n)'$, which we may think of as being "typically small." Note that we are changing our notation slightly; $\boldsymbol{X}$ is no longer assumed to be a random vector [this is done in order to conform with the historical development of the subject].

Throughout, we assume always that the ε_i's are independent with mean zero and common variance σ^2, where $\sigma > 0$ is possibly [in fact, typically] unknown.

In particular, it follows from this assumption that

$$\mathrm{E}\boldsymbol{\varepsilon} = \boldsymbol{0} \text{ and } \mathrm{Var}(\boldsymbol{\varepsilon}) = \sigma^2\boldsymbol{I}. \tag{1}$$

Let us emphasize that our linear model, once written out coordinatewise, is

$$Y_i = \beta_0 x_{i,0} + \cdots + \beta_{p-1} x_{i,p-1} + \varepsilon_i \qquad (1 \le i \le n).$$

It is intuitively clear that unless $n \geq p$, we cannot hope to effectively use our n observed values in order to estimate the $p+1$ unknowns $\sigma^2, \beta_0, \ldots, \beta_{p-1}$. Therefore, we assume always that

$$n \geq p + 1. \tag{2}$$

This condition guarantees that our linear model is not overspecified.

The best-studied linear models are the *normal models*. Those are linear models for which we assume the more stringent condition that

$$\boldsymbol{\varepsilon} \sim \mathrm{N}_n\left(\mathbf{0}\,, \sigma^2 \boldsymbol{I}\right). \tag{3}$$

Example 1 (A measurement-error model)**.** Here we study a socalled measurement-error model: Suppose the observations $Y_1, \ldots, Y_n$ satisfy

$$Y_i = \mu + \varepsilon_i \qquad (1 \leq i \leq n)$$

for an unknown parameter μ. This is a simplest example of a linear model, where $\boldsymbol{\beta} = \mu$ is 1×1, and $\boldsymbol{X} := \mathbf{1}_{n \times 1}$ is a vector of n ones. □

Example 2 (Simple linear regression)**.** In simple linear regression we assume that the observed values have the form

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \qquad (1 \leq i \leq n),$$

where x_i is the predictive variable the corresponds to observation i, and β_0, β_1 are unknown. Simple linear regression fits into our theory of linear models, once we set the design matrix as

$$\boldsymbol{X} := \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}.$$

Example 3 (Polynomial regression)**.** Consider a nonlinear regression model

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_{p-1} x_i^{p-1} + \varepsilon_i \qquad (1 \leq i \leq n),$$

where p is a known integer ≥ 1 [$p-1$ denotes the degree of the polynomial approximation to the observed y's]. Then polynomial regression models are linear models with design matrices of the form

$$\boldsymbol{X} := \begin{pmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^{p-1} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_n & x_n^2 & \cdots & x_n^{p-1} \end{pmatrix}.$$

Example 4 (One-way layout with equal observations)**.** In the simplest case of one-way layout, in analysis of variance, our observations are indexed by vectors themselves as follows:

$$Y_{i,j} = \mu_i + \varepsilon_{i,j} \qquad (1 \leq i \leq I\ ,\ 1 \leq j \leq J).$$

For instance, suppose we are interested in the effect of I different fertilizers. We apply these fertilizers to J different blocks, independently, and "measure the effect." Then, $Y_{i,j}$ is the effect of fertilizer i in block j. The preceding model is assuming that, up to sampling error, the effect of fertilizer i is μ_i. This is a linear model. Indeed, we can create a new random vector $\boldsymbol{Y}$ of IJ observations by simply "vectorizing" the $Y_{i,j}$'s:

$$\boldsymbol{Y} := (Y_{1,1}\,, \ldots, Y_{1,J}\,, Y_{2,1}\,, \ldots, Y_{2,J}\,, \ldots, Y_{I,1}\,, \ldots, Y_{I,J})'\,.$$

The vector $\boldsymbol{\beta}$ of unknowns is $\boldsymbol{\beta} := (\mu_1\,, \ldots, \mu_I)'$, and the design matrix is the following $IJ \times I$ matrix:

$$\boldsymbol{X} := \begin{pmatrix} \mathbf{1}_{J\times 1} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{1}_{J\times 1} & \mathbf{0} & \cdots & \mathbf{0} \\ \vdots & \vdots & \vdots & & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{1}_{J\times 1} \end{pmatrix},$$

where $\mathbf{1}_{J\times 1} := (1\,, \ldots, 1)'$ is a J-vector of all ones. It is possible to show that *one-way layout with unequal number of observations* is also a linear model. That is the case where $Y_{i,j} = \mu_i + \varepsilon_{i,j}$, where $1 \le i \le I$ and $1 \le j \le J_i$ [the number of observed values might differ from block to block].

2. The least-squares estimator of $\boldsymbol{\theta} := \boldsymbol{X\beta}$

Let us return to our general linear model

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

Ultimately, our goal is to first find and then analyze the least-squares estimator $\widehat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$. But first, let us find the least-squares estimate for $\boldsymbol{\theta} := \boldsymbol{X\beta}$. In other words, we wish to perform the following optimization problem:

$$\min_{\boldsymbol{\beta}\in\mathrm{R}^p} \|\boldsymbol{Y} - \boldsymbol{X\beta}\| = \min_{\boldsymbol{\theta}\in\mathcal{C}(\boldsymbol{X})} \|\boldsymbol{Y} - \boldsymbol{\theta}\|.$$

Abstractly speaking, the minimizer solves

$$\widehat{\boldsymbol{\theta}} = P_{\mathcal{C}(\boldsymbol{X})}\boldsymbol{Y}.$$

But is there an optimal $\boldsymbol{\beta}$? As we shall see next, there certainly is a unique $\widehat{\boldsymbol{\beta}}$ when $\boldsymbol{X}$ has full rank.

3. The least-squares estimator of $\boldsymbol{\beta}$

Our *least-squares estimator* $\widehat{\boldsymbol{\beta}}$ of the vector parameter $\boldsymbol{\beta}$ is defined via

$$\min_{\boldsymbol{\beta}\in\mathrm{R}^p} \|\boldsymbol{Y} - \boldsymbol{X\beta}\| = \left\|\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\right\|.$$

We aim to solve this minimization problem under natural conditions on the design matrix $\boldsymbol{X}$. But first, let us introduce some notation. The vector

$$\widehat{\boldsymbol{Y}} := \widehat{\boldsymbol{\theta}} := \boldsymbol{X}\widehat{\boldsymbol{\beta}}$$

is called the vector of *fitted values*, and the coordinates of

$$\boldsymbol{e} := \boldsymbol{Y} - \widehat{\boldsymbol{Y}}$$

are the socalled *residuals*.

Now we write our minimization problem in the following form: First find the minimizing $\widehat{\boldsymbol{\beta}} \in \mathbf{R}^p$ that solves

$$\min_{z \in \mathcal{C}(\boldsymbol{X})} \|\boldsymbol{Y} - \boldsymbol{z}\| = \left\|\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\right\|.$$

Now we know from Proposition 18 (p. 19) that the vector $\boldsymbol{z}$ that achieves this minimum does so uniquely, and is given by $\boldsymbol{P}_{\mathcal{C}(\boldsymbol{X})}\boldsymbol{Y}$, where we recall $\boldsymbol{P}_{\mathcal{C}(\boldsymbol{X})} := \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ is projection onto the column space of $\boldsymbol{X}$; this of course is valid provided that $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ exists. Now the matrix $\boldsymbol{P}_{\mathcal{C}(\boldsymbol{X})}$ plays such an important role that it has its own name: It is called the *hat matrix*, and is denoted as

$$\boldsymbol{H} := \boldsymbol{P}_{\mathcal{C}(\boldsymbol{X})} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'.$$

$\boldsymbol{H}$ is called the *hat matrix* because it maps the observations $\boldsymbol{Y}$ to the fitted values $\widehat{\boldsymbol{Y}}$ [informally, it puts a "hat" over $\boldsymbol{Y}$]. More precisely, the defining feature of $\boldsymbol{H}$ is that

$$\widehat{\boldsymbol{Y}} = \boldsymbol{H}\boldsymbol{Y},$$

once again provided that $\boldsymbol{X}'\boldsymbol{X}$ is nonsingular. The following gives us a natural method for checking this nonsingularity condition in terms of $\boldsymbol{X}$ directly.

Lemma 5. *$\boldsymbol{X}'\boldsymbol{X}$ is nonsingular if and only if* $\operatorname{rank}(\boldsymbol{X}) = p$.

Proof. Basic linear algebra tells us that the positive semidefinite $\boldsymbol{X}'\boldsymbol{X}$ is nonsingular if and only if it is positive definite; i.e., if and only if it has full rank. Since the rank of $\boldsymbol{X}'\boldsymbol{X}$ is the same as the rank of $\boldsymbol{X}$, and since $n > p$, $\boldsymbol{X}'\boldsymbol{X}$ has full rank if and only if its rank, which is the same as $\operatorname{rank}(\boldsymbol{X})$, is p. □

From now on we always assume [unless we state explicitly otherwise] that $\operatorname{rank}(\boldsymbol{X}) = p$. We have shown that, under this condition, if there is a $\widehat{\boldsymbol{\beta}}$, then certainly $\widehat{\boldsymbol{Y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}} = \boldsymbol{H}\boldsymbol{Y} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$. In order to find $\widehat{\boldsymbol{\beta}}$ from this, multiply both sides by $(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ to see that $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$.

The quantity $\text{RSS} := \boldsymbol{e}'\boldsymbol{e} = \|\boldsymbol{e}\|^2 := \|\boldsymbol{Y} - \widehat{\boldsymbol{Y}}\|^2$ is called the *sum of squared residuals*, also known as the *residual sum of squared errors,* and is given by $\|(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}\|^2 = \|\boldsymbol{P}_{\mathcal{C}(\boldsymbol{X})^\perp}\boldsymbol{Y}\|^2$. In particular, we have the following:

Proposition 6. *If* $\operatorname{rank}(\boldsymbol{X}) = p$, *then the least-squares estimator of* $\boldsymbol{\beta}$ *is*

$$\widehat{\boldsymbol{\beta}} := (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}, \tag{4}$$

and $\text{RSS} = \|(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}\|^2$.

Let us make some elementary computations with the least squares estimator of $\boldsymbol{\beta}$.

Lemma 7. *$\widehat{\boldsymbol{\beta}}$ is an unbiased estimator of $\boldsymbol{\beta}$, and* $\text{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$.

Proof. Because $\text{E}\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\text{E}\boldsymbol{Y} = \boldsymbol{\beta}$, it follows that $\widehat{\boldsymbol{\beta}}$ is unbiased. Also, $\text{Var}(\widehat{\boldsymbol{\beta}}) = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\text{Var}(\boldsymbol{Y})\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$, because of (1). □

4. Optimality

Theorem 8. *Let $\boldsymbol{\theta} := \boldsymbol{X\beta}$ be estimated, via least squares, by $\widehat{\boldsymbol{\theta}} := \boldsymbol{HY}$. Then, for all linear unbiased estimates of $\boldsymbol{c}'\boldsymbol{\theta}$, the estimator $\boldsymbol{c}'\widehat{\boldsymbol{\theta}}$ uniquely minimizes the variance.*

By a "linear unbiased estimator of $\boldsymbol{c}'\boldsymbol{\theta}$" we mean an estimator of the form $\sum_{j=1}^n a_jY_j$ whose expectation is $\boldsymbol{c}'\boldsymbol{\theta}$. In this sense, $\boldsymbol{c}'\widehat{\boldsymbol{\theta}}$ is the *best linear unbiased estimator* [or "BLUE"] of $\boldsymbol{c}'\boldsymbol{\theta}$. The preceding can be improved upon as follows, though we will not prove it:

Theorem 9 (Rao)**.** *Under the normal model* (3)*, $\boldsymbol{c}'\widehat{\boldsymbol{\theta}}$ is the unique UMVUE of $\boldsymbol{c}'\boldsymbol{\theta}$, for every nonrandom vector $\boldsymbol{c} \in \mathbf{R}^n$.*

Let us consider Theorem 8 next.

Proof of Theorem 8. We saw on page 41 that $\widehat{\boldsymbol{\theta}} := \boldsymbol{HY}$ irrespective of whether or not $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ exists. Therefore,

$$\text{E}\left(\boldsymbol{c}'\widehat{\boldsymbol{\theta}}\right) = \boldsymbol{c}'\text{E}\widehat{\boldsymbol{\theta}} = \boldsymbol{c}'\boldsymbol{\theta}, \quad \text{Var}\left(\boldsymbol{c}'\widehat{\boldsymbol{\theta}}\right) = \boldsymbol{c}'\text{Var}(\boldsymbol{HY})\boldsymbol{c} = \boldsymbol{c}'\boldsymbol{H}'\text{Var}(\boldsymbol{Y})\boldsymbol{Hc}$$
$$= \sigma^2\|\boldsymbol{Hc}\|^2.$$

Any other linear estimator has the form $\boldsymbol{a}'\boldsymbol{Y}$, and satisfies

$$\text{E}\left(\boldsymbol{a}'\boldsymbol{Y}\right) = \boldsymbol{a}'\text{E}\boldsymbol{Y} = \boldsymbol{a}'\boldsymbol{\theta}, \quad \text{Var}\left(\boldsymbol{a}'\boldsymbol{Y}\right) = \boldsymbol{a}'\text{Var}(\boldsymbol{Y})\boldsymbol{a} = \sigma^2\|\boldsymbol{a}\|^2.$$

If, in addition, $\boldsymbol{a}'\boldsymbol{Y}$ is unbiased, then it follows that $\boldsymbol{a}'\boldsymbol{\theta} = \boldsymbol{c}'\boldsymbol{\theta}$; i.e., $\boldsymbol{a} - \boldsymbol{c}$ is orthogonal to $\boldsymbol{\theta}$. This should hold no matter what value $\boldsymbol{\beta}$ [and hence $\boldsymbol{\theta}$] takes. Since $\mathcal{C}(\boldsymbol{X})$ is the collection of all possible values of $\boldsymbol{\theta}$, it follows that $\boldsymbol{a} - \boldsymbol{c}$ is orthogonal to $\mathcal{C}(\boldsymbol{X})$. Because $\boldsymbol{H}$ is projection onto $\mathcal{C}(\boldsymbol{X})$, it follows that $\boldsymbol{H}(\boldsymbol{a} - \boldsymbol{c}) = \boldsymbol{a}$; equivalently, $\boldsymbol{Hc} = \boldsymbol{Ha}$. Therefore, $\text{Var}(\boldsymbol{c}'\widehat{\boldsymbol{\theta}}) = \sigma^2\|\boldsymbol{Ha}\|^2$ and

$$\text{Var}\left(\boldsymbol{a}'\boldsymbol{Y}\right) - \text{Var}\left(\boldsymbol{c}'\widehat{\boldsymbol{\theta}}\right) = \sigma^2\left\{\|\boldsymbol{a}\|^2 - \|\boldsymbol{Ha}\|^2\right\} = \sigma^2\|(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{a}\|^2 \geq 0,$$

thanks to the Pythagorean property. □

5. Regression and prediction

Now that we have the least-squares estimate for $\boldsymbol{\beta}$, let us use it in order to make prediction.

Recall that our model is $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$. In applications, Y_i is the ith observation for the y variable, and the linear model is really saying that given an explanatory variable $\boldsymbol{x} = (x_0\,, \ldots, x_{p-1})'$,

$$y = \beta_0 x_0 + \cdots + \beta_{p-1} x_{p-1} + \text{“noise.”}$$

Therefore, our prediction, for a given $\boldsymbol{x}$, is

$$\text{[predicted value] } y = \widehat{\beta}_0 x_0 + \cdots + \widehat{\beta}_{p-1} x_{p-1}, \tag{5}$$

where $\widehat{\beta} = (\widehat{\beta}_0\,, \ldots, \widehat{\beta}_{p-1})'$ is the least-squares estimate of $\boldsymbol{\beta}$. We may view the right-hand side of (5) as a function of $\boldsymbol{x}$, and call (5) the equation for the “regression line.”

6. Estimation of $\boldsymbol{\sigma}^2$

We wish to also estimate σ^2. The estimator of interest to us turns out to be the following:

$$S^2 := \frac{1}{n-p}\text{RSS}, \tag{6}$$

as the next lemma suggests.

Lemma 10. *S^2 is an unbiased estimator of σ^2.*

Proof. Recall that RSS $= \boldsymbol{e}'\boldsymbol{e} = \|\boldsymbol{Y} - \boldsymbol{HY}\|^2$. We can write the RSS as $\|(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}\|^2 = \boldsymbol{Y}'(\boldsymbol{I} - \boldsymbol{H})'(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y} = \boldsymbol{Y}'(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}$. In other words, the RSS is a random quadratic form for the matrix $\boldsymbol{A} := \boldsymbol{I} - \boldsymbol{H}$, and hence

$$\begin{aligned} \mathrm{E}(\text{RSS}) &= \mathrm{tr}((\boldsymbol{I} - \boldsymbol{H})\mathrm{Var}(\boldsymbol{Y})) + (\mathrm{E}\boldsymbol{Y})'(\boldsymbol{I} - \boldsymbol{H})(\mathrm{E}\boldsymbol{Y}) \\ &= \sigma^2 \mathrm{tr}(\boldsymbol{I} - \boldsymbol{H}) + (\boldsymbol{X\beta})'(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{X\beta}. \end{aligned}$$

Because $\boldsymbol{X\beta} \in \mathcal{C}(\boldsymbol{X})$, $\boldsymbol{I} - \boldsymbol{H}$ projects onto the orthogonal subspace of where it is, therefore $(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{X\beta} = \boldsymbol{0}$. And the trace of the projection matrix $(\boldsymbol{I} - \boldsymbol{H})$ is its rank, which is $n - \mathrm{tr}(\boldsymbol{H}) = n - p$, since $\boldsymbol{X}$ has full rank p. It follows that $\mathrm{E}(\text{RSS}) = \sigma^2(n-p)$, and therefore $\mathrm{E}(S^2) = \sigma^2$. □

7. The normal model

In the important special case of the normal model,

$$\boldsymbol{Y} \sim \mathrm{N}_n\left(\boldsymbol{X\beta}\,, \sigma^2 \boldsymbol{I}\right).$$

Therefore, $\widehat{\boldsymbol{\beta}} \sim \mathrm{N}_p(\boldsymbol{\beta}\,, \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1})$. And the vector $(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y} = \boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}$ of *residual errors* is also a multivariate normal:

$$(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y} \sim \mathrm{N}_{n-p}\left((\boldsymbol{I} - \boldsymbol{H})\boldsymbol{X\beta}\,, \sigma^2(\boldsymbol{I} - \boldsymbol{H})(\boldsymbol{I} - \boldsymbol{H})'\right) = \mathrm{N}_{n-p}\left(\boldsymbol{0}\,, \sigma^2(\boldsymbol{I} - \boldsymbol{H})\right).$$

Therefore, in the normal model,

$$S^2 = \frac{1}{n-p}\,\|(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}\|^2 \sim \sigma^2 \frac{\chi^2_{n-p}}{n-p}.$$

Finally, we note that $\boldsymbol{t}'\widehat{\boldsymbol{\beta}} + \boldsymbol{s}'(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}$ is a matrix times $\boldsymbol{Y}$ for all $\boldsymbol{t} \in \mathbf{R}^p$ and $\boldsymbol{s} \in \mathbf{R}^{n-p}$. Therefore, $(\widehat{\boldsymbol{\beta}}\,, (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y})$ is also a multivariate normal. But

$$\mathrm{Cov}\left(\widehat{\boldsymbol{\beta}}\ ,\ (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}\right) = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\mathrm{Var}(\boldsymbol{Y})(\boldsymbol{I}-\boldsymbol{H})' = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{I}-\boldsymbol{H}) = \mathbf{0},$$

since the columns of $\boldsymbol{X}$ are obviously orthogonal to every element in $\mathcal{C}(\boldsymbol{X})^{\perp}$ and $\boldsymbol{I} - \boldsymbol{H} = \boldsymbol{P}_{\mathcal{C}(\boldsymbol{X})^{\perp}}$. This shows that $\widehat{\boldsymbol{\beta}}$ and $(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}$ are independent, and hence $\widehat{\boldsymbol{\beta}}$ and S^2 are also independent. Thus, we summarize our findings.

Theorem 11. *The least-squares estimator $\widehat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ is given by* (4)*; it is always unbiased. Moreover, S^2 is an unbiased estimator of σ^2. Under the normal model, S and $\widehat{\boldsymbol{\beta}}$ are independent, and*

$$\widehat{\boldsymbol{\beta}} \sim \mathrm{N}_p\left(\boldsymbol{\beta}\,, \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}\right), \quad S^2 \sim \sigma^2 \frac{\chi^2_{n-p}}{n-p}.$$

Recall that $\boldsymbol{Y}$ has the nondegenerate multivariate normal distribution $N_n(\boldsymbol{X\beta}\,, \sigma^2\boldsymbol{I})$. Therefore, its pdf is

$$f_{\boldsymbol{Y}}(\boldsymbol{y}) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}\|\boldsymbol{y} - \boldsymbol{X\beta}\|^2\right).$$

This shows readily the following.

Lemma 12. *In the normal model, $\widehat{\boldsymbol{\beta}}$ is the MLE of $\boldsymbol{\beta}$ and $\left(\frac{n-p}{n}\right) S^2$ is the MLE for σ^2.*

Proof. Clearly, maximizing the likelihood function, over all $\boldsymbol{\beta}$, is the same as minimizing $\|\boldsymbol{Y} - \boldsymbol{X\beta}\|^2$. Therefore, MLE = least squares for $\boldsymbol{\beta}$. As for σ^2, we write the log likelihood function:

$$L(\sigma) = -\frac{n}{2}\ln(2\pi) - n\ln\sigma - \frac{1}{2\sigma^2}\|\boldsymbol{Y} - \boldsymbol{X\beta}\|^2.$$

Then,

$$L'(\sigma) = -\frac{n}{\sigma} + \frac{1}{\sigma^3}\|\boldsymbol{Y} - \boldsymbol{X\beta}\|^2.$$

Set $L'(\sigma) = 0$ and solve to see that the MLE of σ^2 is $\frac{1}{n}\|\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\|^2 = \left(\frac{n-p}{n}\right) S^2$, thanks to the MLE principle and the already-proven fact that the MLE of $\boldsymbol{\beta}$ is $\widehat{\boldsymbol{\beta}}$. □

8. Some examples

1. A measurement-error model. Recall the measurement-error model

$$Y_i = \mu + \varepsilon_i \qquad (1 \le i \le n).$$

We have seen that this is a linear model with $p = 1$, $\boldsymbol{\beta} = \mu$, and $\boldsymbol{X} := \mathbf{1}_{n\times 1}$. Since $\boldsymbol{X}'\boldsymbol{X} = n$ and $\boldsymbol{X}'\boldsymbol{Y} = \sum_{i=1}^n Y_i$, we have

$$\widehat{\boldsymbol{\beta}} = \widehat{\mu} := \bar{Y},$$

and

$$\frac{1}{n-1}S^2 = \frac{1}{n-1}\left\|\boldsymbol{Y} - \bar{Y}\mathbf{1}_{n\times 1}\right\|^2 = \frac{1}{n-1}\sum_{i=1}^n (Y_i - \bar{Y})^2 := S^2.$$

These are unbiased estimators for μ and σ^2, respectively. Under the normal model, $\bar{Y}$ and S^2 are independent, $\bar{Y} \sim \mathrm{N}(\mu\,,\sigma^2)$ and $(n-1)S^2 \sim \sigma^2\chi^2_{n-1}$. These are some of the highlights of Math. 5080.

2. Simple linear regression. Recall that simple linear regression is our linear model in the special case that $p = 2$, $\boldsymbol{\beta} = (\beta_0\,,\beta_1)'$, and

$$\boldsymbol{X} := \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}.$$

We have

$$\boldsymbol{X}'\boldsymbol{X} = \begin{pmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{pmatrix}, \quad \boldsymbol{X}'\boldsymbol{Y} = \begin{pmatrix} \sum_{i=1}^n Y_i \\ \sum_{i=1}^n x_i Y_i \end{pmatrix},$$

and

$$\begin{aligned}(\boldsymbol{X}'\boldsymbol{X})^{-1} &= \frac{1}{n\sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2}\begin{pmatrix} \sum_{i=1}^n x_i^2 & -\sum_{i=1}^n x_i \\ -\sum_{i=1}^n x_i & n \end{pmatrix} \\ &= \frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2}\begin{pmatrix} \frac{1}{n}\sum_{i=1}^n x_i^2 & -\bar{x} \\ -\bar{x} & 1 \end{pmatrix}.\end{aligned}$$

Therefore,

$$\begin{pmatrix} \widehat{\beta}_0 \\ \widehat{\beta}_1 \end{pmatrix} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y},$$

which leads to

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})Y_i}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \text{ and } \widehat{\beta}_0 = \bar{Y} - \widehat{\beta}_1\bar{x}.$$

We have derived these formulas by direct computation already. In this way we find that the fitted values are

$$\widehat{\boldsymbol{Y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}} = \begin{pmatrix} \widehat{\beta}_0 + \widehat{\beta}_1 x_1 \\ \vdots \\ \widehat{\beta}_0 + \widehat{\beta}_1 x_n \end{pmatrix}.$$

Also,

$$S^2 = \frac{1}{n-2}\left\|\boldsymbol{Y} - \widehat{\boldsymbol{Y}}\right\|^2 = \frac{1}{n-2}\sum_{i=1}^n \left(Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i\right)^2,$$

and this is independent of $(\widehat{\beta}_0\,,\widehat{\beta}_1)$ under the normal model.

Recall that our linear model is, at the moment, the simple regression model,

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i.$$

Perhaps the most important first question that we can ask in this context is $\beta_1 = 0$; that is, we wish to know whether the x variables are [linearly] independent of the Y's. Let us try to find a confidence interval for β_1 in order to answer this question. From now on, we work under the normal model. Recall that under the normal model,

$$\widehat{\boldsymbol{\beta}} \sim \mathrm{N}_2\left(\boldsymbol{\beta}\,,\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}\right) \;\Rightarrow\; \widehat{\beta}_1 \sim \mathrm{N}\left(\beta_1\,,\sigma^2\left[(\boldsymbol{X}'\boldsymbol{X})^{-1}\right]_{2,2}\right) = \mathrm{N}\left(\beta_1\,,\frac{\sigma^2}{\sum_{i=1}^n (x_i-\bar{x})^2}\right).$$

Equivalently,

$$Z := \frac{\sqrt{\sum_{i=1}^n (x_i-\bar{x})^2}}{\sigma}\left(\widehat{\beta}_1 - \beta_1\right) \sim \mathrm{N}(0\,,1).$$

Now, S^2/σ^2 is independent of Z and is distributed as $\chi^2_{n-2}/(n-2)$. Therefore,

$$\sqrt{\sum_{i=1}^n (x_i-\bar{x})^2}\left(\frac{\widehat{\beta}_1-\beta_1}{S}\right) = \frac{Z}{\sqrt{S^2/\sigma^2}} \sim t_{n-2}.$$

Therefore, a $(1-\alpha)\times 100\%$ confidence interval for β_1 is

$$\widehat{\beta}_1 \pm \frac{S}{\sqrt{\sum_{i=1}^n (x_i-\bar{x})^2}}\, t_{n-2}^{\alpha/2},$$

where t_ν^r is the point whose right area, under the t_ν pdf, is r. If zero is not in this confidence interval, then our statistical prediction is that β_1 is not zero [at the confidence level α].

3. A remark on polynomial regression. Recall that, in polynomial regression, we have $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$, where $\boldsymbol{X}$ is the following design matrix:

$$\boldsymbol{X} := \begin{pmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^{p-1} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_n & x_n^2 & \cdots & x_n^{p-1} \end{pmatrix}.$$

If $p = 2$, then this is simple linear regression. Next consider the "quadratic regression" model where $p = 3$. That is,

$$\boldsymbol{X} := \begin{pmatrix} 1 & x_1 & x_1^2 \\ \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 \end{pmatrix} \quad\Rightarrow\quad \boldsymbol{X}'\boldsymbol{X} = \begin{pmatrix} n & \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 & \sum_{i=1}^n x_i^3 \\ \sum_{i=1}^n x_i^2 & \sum_{i=1}^n x_i^3 & \sum_{i=1}^n x_i^4 \end{pmatrix}.$$

Because $n \geq 4$, $\boldsymbol{X}$ is nonsingular if and only if $\boldsymbol{x}$ is not a vector of constants [a natural condition]. But you see that already it is painful to invert $\boldsymbol{X}'\boldsymbol{X}$. This example shows the importance of using a computer in linear models: Even fairly simple models are hard to work with, using only direct calculations.

Lecture 8

Assessing Normality

1. Visual data exploration

A big part of our theory of linear models has been under the normal model; that is the most successful part of the theory applies when $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$ where we assumed that $\boldsymbol{\varepsilon} \sim \mathrm{N}_n(\boldsymbol{0}\,, \sigma^2\boldsymbol{I})$; equivalently, that $\varepsilon_1, \ldots, \varepsilon_n$ are independent $\mathrm{N}(0\,, \sigma^2)$'s.

A natural question, for a given data set, is to ask, "is the noise coming as i.i.d. $\mathrm{N}(0\,, \sigma^2)$'s"?

Since the noise is not observable in our model, it seems natural that we "estimate" it using the residuals $\boldsymbol{e} := \boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}$. It stands to reason that if $\boldsymbol{\varepsilon}$ is a vector of n i.i.d. $\mathrm{N}(0\,, \sigma^2)$'s, then the histogram of $e_1\,, \ldots, e_n$ should look like a normal density. One should not underestimate the utility of this simple idea; for instance, we should see, roughly, that:

– approximately 68.3% of the e_i's should fall in $[-S^2\,, S^2]$,

– approximately 95.4% of the e_i's should fall in $[-2S^2\,, 2S^2]$, etc.

This is very useful as a first attempt to assess the normality of the noise in our problem. But it is not conclusive since we cannot assign confidence levels [the method is a little heuristic]. It turns out that this method can be improved upon in several directions, and with only a little more effort.

2. General remarks

We can use the Pearson's χ^2-test in order to test whether a certain data comes from the $N(\mu_0\,, \sigma_0^2)$ distribution, where μ_0 and σ_0 are known. Now we wish to address the same problem, but in the more interesting case that μ_0 and σ_0 are *unknown*. [For instance, you may wish to know whether or not you are allowed to use the usual homoskedasticity assumption in the usual measurement-error model of linear models.]

Here we discuss briefly some solutions to this important problem. Although we write our approach specifically in the context of linear models, these ideas can be developed more generally to test for normality of data in other settings.

3. Histograms

Consider the linear model $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$. The pressing question is, "is it true that $\boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \sigma^2\mathbf{I}_n)$"?

To answer this, consider the "residuals,"

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

If $\boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \sigma^2\mathbf{I}_n)$ then one would like to think that the histogram of the e_i's should look like a normal pdf with mean 0 and variance σ^2 (why?). How close is close? It helps to think more generally.

Consider a sample $U_1, \ldots, U_n$ (e.g., $U_i = e_i$). We wish to know where the U_i's are coming from a normal distribution. The first thing to do is to plot the histogram. In R you type,

```
hist(u,nclass=n)
```

where u denotes the vector of the samples $U_1, \ldots, U_n$ and n denotes the number of bins in the histogram.

For instance, consider the following exam data:

```
16.8 9.2 0.0 17.6 15.2 0.0 0.0 10.4 10.4 14.0 11.2 13.6
12.4 14.8 13.2 17.6 9.2 7.6 9.2 14.4 14.8 15.6 14.4 4.4
14.0 14.4 0.0 0.0 10.8 16.8 0.0 15.2 12.8 14.4 14.0 17.2
0.0 14.4 17.2 0.0 0.0 0.0 14.0 5.6 0.0 0.0 13.2 17.6 16.0
16.0 0.0 12.0 0.0 13.6 16.0 8.4 11.6 0.0 10.4 0.0 14.4 0.0
18.4 17.2 14.8 16.0 16.0 0.0 10.0 13.6 12.0 15.2
```

The command hist(f1.dat,nclass=15) produces Figure 8.1(a).[1]

Try this for different values of nclass to see what types of histograms you can obtain. You should always ask, "which one represents the truth the best"? Is there a unique answer?

Now the data $U_1, \ldots, U_n$ is probably not coming from a normal distribution if the histogram does not have the "right" shape. Ideally, it would be symmetric, and the tails of the distribution taper off rapidly.

In Figure 8.1(a), there were many students who did not take the exam in question. They received a '0' but this grade should probably not contribute to our knowledge of the distribution of all such grades. Figure 8.1(b) shows the

[1] You can obtain this data freely from the website below:
http://www.math.utah.edu/~davar/math6010/2011/Notes/f1.dat.

histogram of the same data set when the zeros are removed. [This histogram appears to be closer to a normal density.]

4. QQ-Plots

QQ-plots are a better way to assess how closely a sample follows a certain distribution.

To understand the basic idea note that if $U_1, \ldots, U_n$ is a sample from a normal distribution with mean μ and variance σ^2, then about 68.3% of the sample points should fall in $[\mu-\sigma, \mu+\sigma]$, 95.4% should fall in $[\mu-2\sigma, \mu+2\sigma]$, etc.

Now let us be more careful still. Let $U_{1:n} \le \cdots \le U_{n:n}$ denote the order statistics of $U_1, \ldots, U_n$. Then no matter how you make things precise, the fraction of data "below" $U_{j:n}$ is $(j \pm 1)/n$. So we make a continuity correction and *define* the fraction of the data below $U_{j:n}$ to be $(j - \frac{1}{2})/n$.

Consider the normal "quantiles," $q_1, q_2, \ldots, q_n$:

$$\Phi(q_j) := \int_{-\infty}^{q_j} \frac{\mathrm{e}^{-x^2/2}}{\sqrt{2\pi}}\,\mathrm{d}x = \frac{j-\frac{1}{2}}{n}; \qquad \text{i.e., } q_j := \Phi^{-1}\left(\frac{j-\frac{1}{2}}{n}\right).$$

Now suppose once again that $U_1, \ldots, U_n \sim \mathrm{N}(\mu\,, \sigma^2)$ is a random [i.i.d.] sample. Let $Z_j := (U_j - \mu)/\sigma$, so that $Z_1, \ldots, Z_n \sim \mathrm{N}(0\,,1)$. The Z's are standardized data, and we expect the fraction of the standardized data that fall below q_j to be about $(j - \frac{1}{2})/n$. In other words, we can put together our observations to deduce that $Z_{j:n} \approx q_j$. Because $U_{j:n} = \sigma Z_{j:n} + \mu$, it follows that $U_{j:n} \approx \sigma q_j + \mu$. In other words, we expect the sample order statistics $U_{1:n}, \ldots, U_{n:n}$ to be very close to some linear function of the normal quantiles $q_1, \ldots, q_n$. In other words, if $U_1, \ldots, U_n$ is a random sample from some normal distribution, then we expect the scatterplot of the pairs $(q_1\,, U_{1:n}), \ldots, (q_n\,; U_{n:n})$ to follow closely a line. [The slope and intercept are σ and μ, respectively.]

QQ-plots are simply the plots of the $\mathrm{N}(0\,,1)$-quantiles $q_1, \ldots, q_n$ versus the order statistics $U_{1:n}, \ldots, U_{n:n}$. To draw the qqplot of a vector $\boldsymbol{u}$ in R, you simply type

$$\mathrm{qqnorm}(u).$$

Figure 8.2(a) contains the qq-plot of the exam data we have been studying here.

5. The Correlation Coefficient of the QQ-Plot

In its complete form, the R-command qqnorm has the following syntax:

$$\mathrm{qqnorm(u, datax = FALSE, plot = TRUE)}.$$

The parameter u denotes the data; datax is "FALSE" if the data values are drawn on the y-axis (default). It is "TRUE" if you wish to plot $(U_{j:n}, q_j)$ instead of the more traditional $(q_j, U_{j:n})$. The option plot=TRUE (default) tells R to plot the qq-plot, whereas plot=FALSE produces a vector. So for instance, try

$$\mathrm{V} = \mathrm{qqnorm}(\mathrm{u}, \mathrm{plot} = \mathrm{FALSE}).$$

This creates two vectors: V\$$x$ and V\$$y$. The first contains the values of all q_j's, and the second all of the $U_{j:n}$'s. So now you can compute the correlation coefficient of the qq-plot by typing:

$$\begin{aligned}&\mathrm{V} = \mathrm{qqnorm}(\mathrm{u}, \mathrm{plot} = \mathrm{FALSE})\\ &\mathrm{cor}(\mathrm{V}\$x, \mathrm{V}\$y).\end{aligned}$$

If you do this for the qq-plot of the grade data, then you will find a correlation of ≈ 0.910. After censoring out the no-show exams, we obtain a correlation of ≈ 0.971. This produces a noticeable difference, and shows that the grades are indeed normal.

In fact, one can analyse this procedure statistically ["is the sample correlation coefficient corresponding to the line sufficiently close to ± 1"?].

6. Some benchmarks

Figures 8.3 and 8.4 contain four distinct examples. I have used qq-plot in the prorgam environment "R." The image on the left-hand side of Figure 8.3 shows a simulation of 10000 standard normal random variables (in R, you type `x=rnorm(10000,0,1))`, and its qq-plot is drawn on typing `qqnorm(x)`. In a very strong sense, **this figure is the benchmark**.

The image on the right-hand side of Figure 8.3 shows a simulation of 10000 standard Cauchy random variables. That is, the density function is $f(x) = (1/\pi)(1+x^2)^{-1}$. This is done by typing `y=rcauchy(10000,0,1)`, and the resulting qq-plot is produced upon typing `qqnorm(y)`. We know that the Cauchy has much fatter tails than normals. For instance,

$$\mathrm{P}\{\mathrm{Cauchy} > a\} = \frac{1}{\pi}\int_a^\infty \frac{dx}{1+x^2} \sim \frac{1}{\pi a} \quad (\text{as } a \to \infty),$$

whereas $\mathrm{P}\{\mathrm{N}(0\,,1) > a\}$ decays faster than exponentially.[2] Therefore, for a large,

$$\mathrm{P}\{\mathrm{N}(0\,,1) > a\} \ll \mathrm{P}\{\mathrm{Cauchy} > a\}.$$

This heavy-tailedness can be read off in Figure 8.3(b): The Cauchy qq-plot grows faster than linearly on the right-hand side. *And this means that the*

[2] In fact it can be shown that $\bar\Phi(a) := \int_a^\infty \varphi(x)\,dx \approx a^{-1}\varphi(a)$ as $a \to \infty$, where φ denotes the N$(0\,,1)$ pdf. Here is why: Let $G(a) := a^{-1}\varphi(a)$. We know from the fundamental theorem of calculus that $\bar\Phi'(a) = -\varphi(a) = -aG(a)$. Also, $G'(a) = -a^{-1}G(a) - aG(a) \approx -aG(a)$ as $a \to \infty$. In summary: $\bar\Phi(a), G(a) \approx$ and $\bar\Phi'(a) \approx G'(a)$. Therefore, $\bar\Phi(a) \approx G(a)$, thanks to the L'Hôpital's rule of calculus.

standard Cauchy distribution has fatter right-tails. Similar remarks apply to the left tails.

Figure 8.4(a) shows the result of the qq-plot of a simulation of 10000 iid uniform-$(0\,,1)$ random variables. [To generate these uniform random variables you type, `runif(10000,0,1)`.]

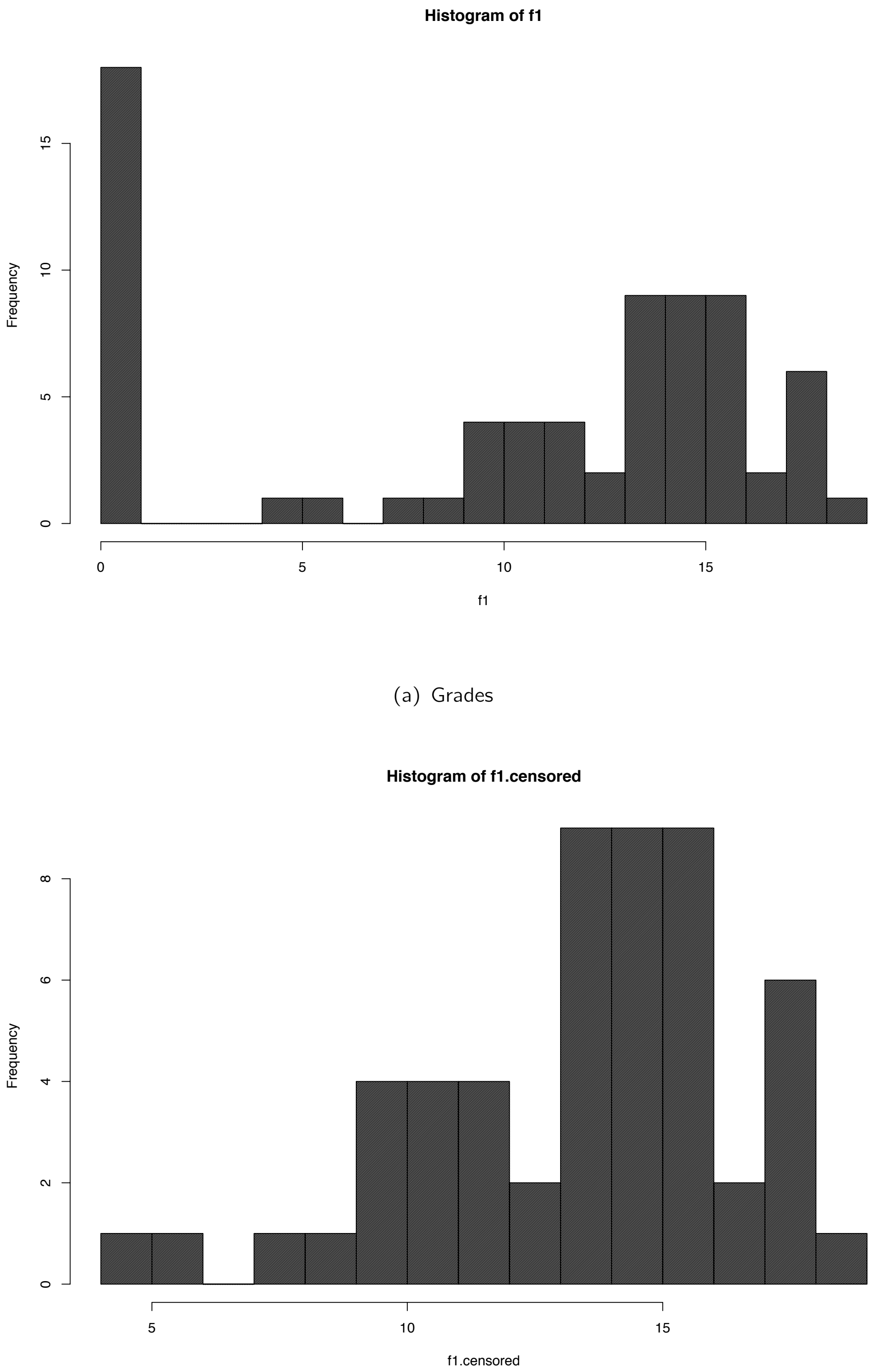
Now uniform-$(0\,,1)$ random variables have much smaller tails than normals because uniforms are in fact *bounded*. This fact manifests itself in Figure 8.4(a). For instance, we can see that the right-tail of the qq-plot for uniform-$(0\,,1)$ grows less rapidly than linearly. And this shows that the right-tail of a uniform is much smaller than that of a normal. Similar remarks apply to the left tails.

A comparison of the figures mentioned so far should give you a feeling for how sensitive qq-plots are to the effects of tails. [All are from distributions that are symmetric about their median.] Finally, let us consider Figure 8.4(a), which shows an example of 10000 Gamma random variables with $\alpha = \beta = 1$. You generate them in R by typing `x=rgamma(10000,1,1)`. Gamma distributions are inherently *asymmetric*. You can see this immediately in the qq-plot for Gammas; see Figure 8.4(b). Because Gamma random variables are nonnegative, the left tail is much smaller than that of a normal. Hence, the left tail of the qq-plot grows more slowly than linearly. The right tail however is fatter. [This is always the case. However, for the sake of simplicity consider the special case where Gamma=Exponential.] This translates to the faster-than-linear growth of the right-tail of the corresponding qq-plot (Figure 8.4(b)).

I have shown you Figures 8.3 and 8.4 in order to high-light the basic features of qq-plots in ideal settings. By “ideal” I mean “simulated data,” of course.

Real data does not generally lead to such sleek plots. Nevertheless one learns a lot from simulated data, mainly because simulated data helps identify key issues without forcing us to have to deal with imperfections and other flaws.

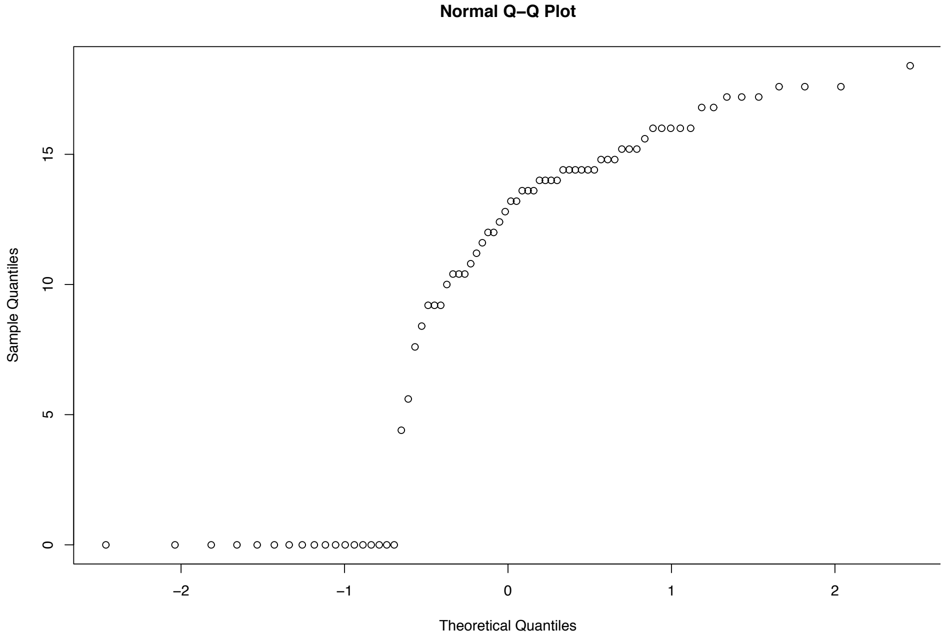
But it is important to keep in mind that it is real data that we are ultimately after. And so the histogram and qq-plot of a certain real data set are depicted in Figure 8.5. Have a careful look and ask yourself a number of questions: Is the data normally distributed? Can you see how the shape of the histogram manifests itself in the shape and gaps of the qq-plot? Do the tails look like those of a normal distribution? To what extent is the ”gap” in the histogram “real”? By this I mean to ask what do you think might happen if we change the bin-size of the histogram in Figure 8.5?



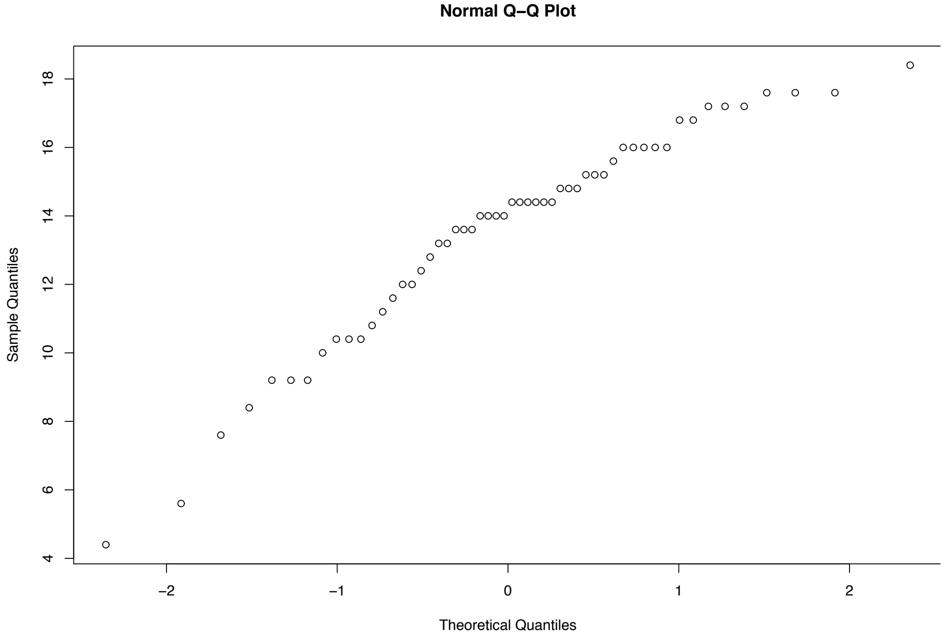
(a) Grades

(b) Censored Grades

Figure 8.1. Histogram of grades and censored grades



(a) QQ-plot of grades



(b) QQ-plot of censored grades

Figure 8.2. QQ-plots of grades and censored grades

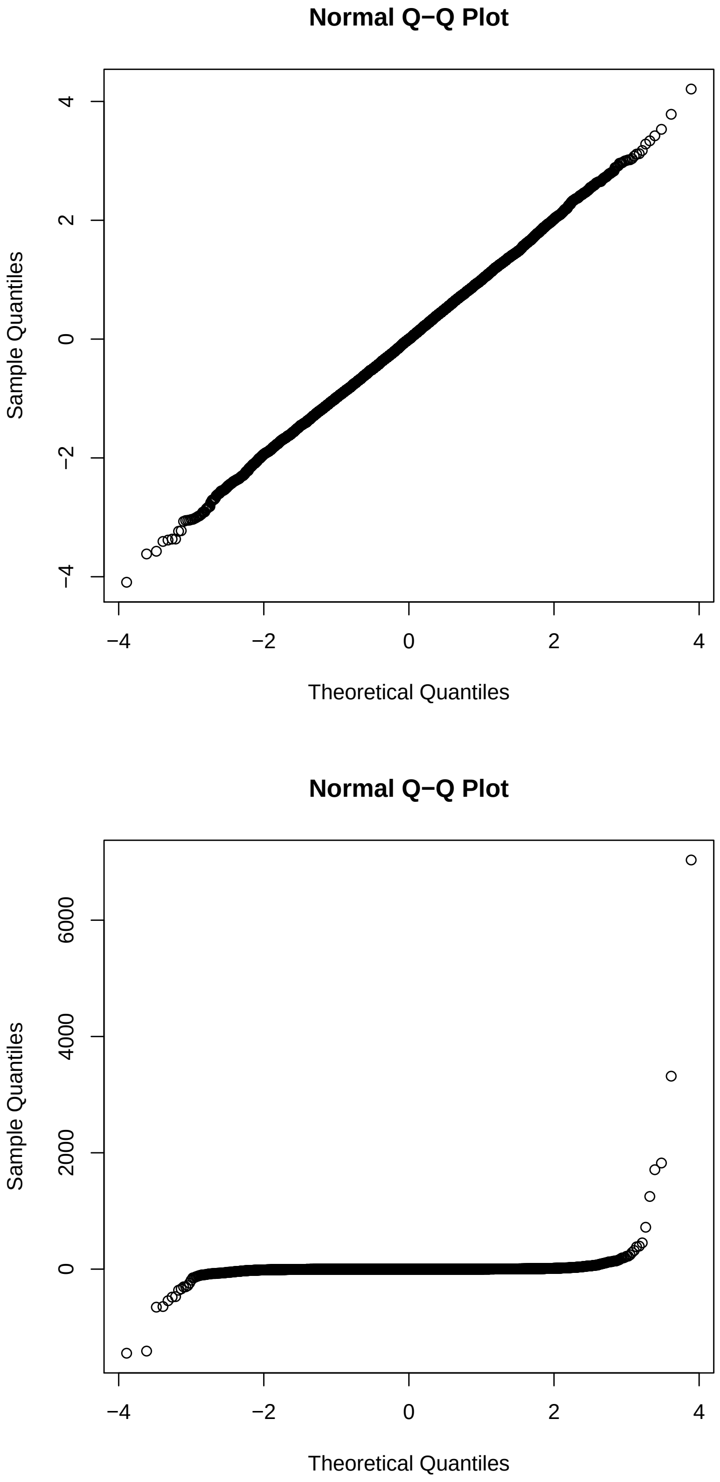


Figure 8.3. (a) is N(0 , 1) data; (b) is Cauchy data

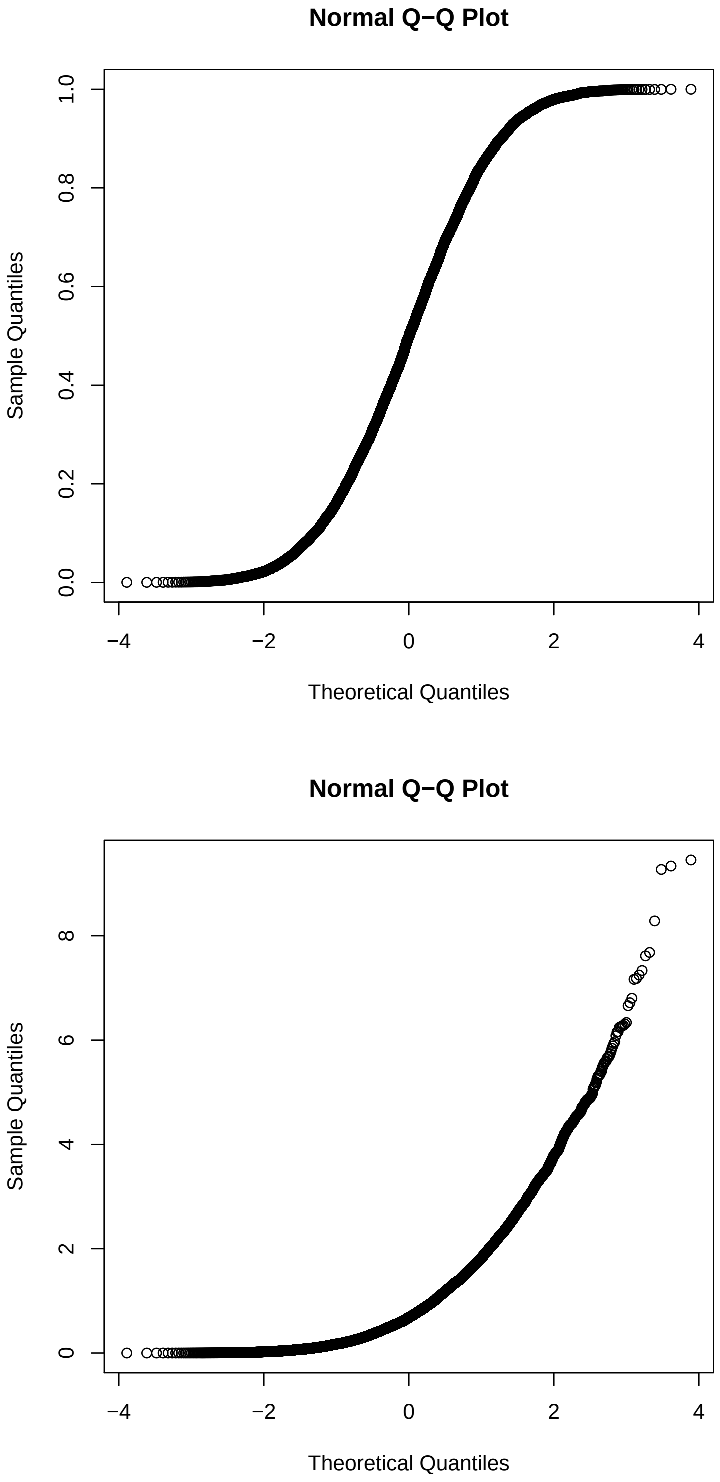


Figure 8.4. (a) is the qq-plot of unif-$(0\,,1)$; (b) is the qq-plot of a Gamma$(1\,,1)$.

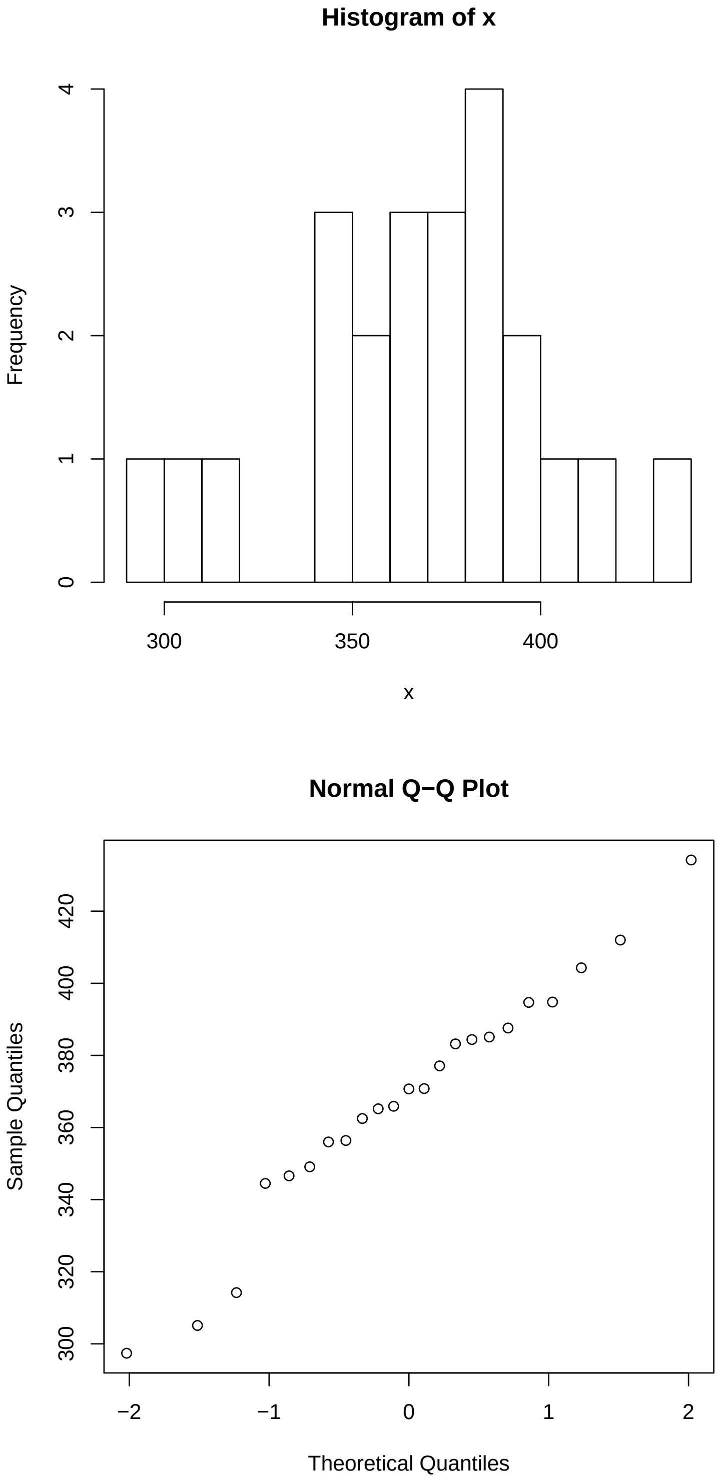


Figure 8.5. Histogram and qq-plot of the data

Lecture 9

Hypothesis Testing

Throughout, we assume the normal-error linear model that is based on the model

$$y = \beta_1 x_1 + \cdots + \beta_p x_p + \text{noise}.$$

[Note the slight change in the notation.]

1. A test for one parameter

Suppose we want to test to see whether or not the $(\ell + 1)$st x-variable has a [linear] effect on the y variable. Of course, $1 \le \ell \le p$, so we are really testing the statistical hypothesis

$$H_0 : \ \beta_\ell = 0.$$

Since $\widehat{\boldsymbol{\beta}} \sim \mathrm{N}_p \left(\boldsymbol{\beta} \, , \sigma^2 (\boldsymbol{X}'\boldsymbol{X})^{-1} \right)$, it follows that

$$\widehat{\beta}_\ell \sim \mathrm{N} \left(\beta_\ell \, , \sigma^2 \left[(\boldsymbol{X}'\boldsymbol{X})^{-1} \right]_{\ell,\ell} \right).$$

Because S is independent of $\widehat{\boldsymbol{\beta}}$ and hence $\widehat{\beta}_\ell$, and since $S^2/\sigma^2 \sim \chi^2_{n-p}/(n-p)$,

$$\frac{\widehat{\beta}_\ell - \beta_\ell}{S\sqrt{[(\boldsymbol{X}'\boldsymbol{X})^{-1}]_{\ell,\ell}}} = \frac{\sigma}{S} \cdot \frac{\widehat{\beta}_\ell - \beta_\ell}{\sigma\sqrt{[(\boldsymbol{X}'\boldsymbol{X})^{-1}]_{\ell,\ell}}} \sim t_{n-p}.$$

Therefore, it is now a routine matter to set up a t-test for $H_0 : \ \beta_\ell = 0$. As usual, testing has implications that are unattractive; it is much better to present a confidence interval [which you can then use for a test if you want, any way]: A $(1 - \alpha) \times 100\%$ confidence interval for β_ℓ is

$$\widehat{\beta}_\ell \pm S t^{\alpha/2}_{n-p} \sqrt{[(\boldsymbol{X}'\boldsymbol{X})^{-1}]_{\ell,\ell}}.$$

If you really insist on performing a level-α test for β_ℓ, it suffices to check to see if this confidence interval contains 0. If 0 is not in the confidence interval then you reject. Otherwise, you do nothing.

2. Least-squares estimates for contrasts

We wish to study a more general problem. Recall that our model has the form

$$y = \beta_1 x_1 + \beta_1 x_1 + \cdots + \beta_p x_p + \text{noise}.$$

If we sample then the preceding becomes $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$, as before. As part of model verification, we might ask to see if $(x_i)_{i\in J}$ should be excised from the model, where $J := \{\ell\,, \ell+1\,, \ldots\,, r\}$ is a subset of the index $\{1\,, \ldots\,, n\}$. In other words, we ask

$$H_0 : \ \beta_\ell = \cdots = \beta_r = 0.$$

Note that we can translate the preceding, using the language of matrix analysis, as $H_0 : \ \boldsymbol{A\beta} = \boldsymbol{0}$, where

$$\boldsymbol{A} := \begin{pmatrix} \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} \end{pmatrix},$$

where the identity matrix in the middle is $(r-\ell+1) \times (r-\ell+1)$; it starts on position $(\ell\,,\ell)$ and runs $r-\ell$ units in rows and in columns.

Now we ask a slightly more general question [it pays to do this, as it turns out]: Suppose $\boldsymbol{A}$ is a $q \times p$ matrix of full rank $q \le p$, and we are interested in testing the hypothesis,

$$H_0 : \ \boldsymbol{A\beta} = \boldsymbol{0}. \tag{1}$$

The first question to ask is, "how can we estimate $\boldsymbol{\beta}$"? The answer is given to us by the principle of least squares: We write [as before]

$$\boldsymbol{Y} = \boldsymbol{\theta} + \boldsymbol{\varepsilon}, \qquad \text{where } \boldsymbol{\theta} := \boldsymbol{X\beta},$$

and $\boldsymbol{\varepsilon} = (\varepsilon_1\,, \ldots\,, \varepsilon_n)'$ are mean-zero random variables, and first find the least-squares estimate $\widehat{\boldsymbol{\theta}}_{H_0}$ of $\boldsymbol{\theta}$, under the assumption that H_0 is valid. That is, we seek to minimize $\|\boldsymbol{Y} - \boldsymbol{Xb}\|^2$ over all p-vectors $\boldsymbol{b}$ such that $\boldsymbol{Ab} = \boldsymbol{0}$. The optimal value yields $\widehat{\boldsymbol{\theta}}_{H_0} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}_{H_0}$.Then we obtain $\widehat{\boldsymbol{\beta}}_{H_0}$ by noticing that if $\boldsymbol{X}$ has full rank, then $\widehat{\boldsymbol{\beta}}_{H_0} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\widehat{\boldsymbol{\theta}}_{H_0}$.

Now it follows by differentiation [or just geometrically] that $\widehat{\boldsymbol{\theta}}_{H_0}$ is the projection of $\boldsymbol{Y}$ onto the subspace $\mathcal{G}$ of all vectors of the form $\boldsymbol{\vartheta} = \boldsymbol{Xb}$ that satisfy $\boldsymbol{Ab} = \boldsymbol{0}$, where $\boldsymbol{b}$ is a p-vector. We can simplify this description a little when $\boldsymbol{X}$ has full rank. Note that whenever $\boldsymbol{\vartheta} = \boldsymbol{Xb}$, we can solve to get $\boldsymbol{b} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\vartheta}$. Therefore, it follows that—when $\boldsymbol{X}$ has full rank— $\widehat{\boldsymbol{\theta}}_{H_0}$ is the projection of the observations vector $\boldsymbol{Y}$ onto the subspace $\mathcal{G}$ of all vectors of the form $\boldsymbol{\vartheta}$ that satisfy

$$\boldsymbol{A}_1\boldsymbol{\vartheta} = \boldsymbol{0}, \quad \text{where} \quad \boldsymbol{A}_1 := \boldsymbol{A}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'.$$

In other words, $\mathcal{G}$ is the subspace of $\mathcal{C}(\boldsymbol{X})$, whose every element $\boldsymbol{\vartheta}$ is orthogonal to every row of $\boldsymbol{A}_1$. In symbols,

$$\mathcal{G} = \mathcal{C}(\boldsymbol{X}) \cap \left[\mathcal{C}(\boldsymbol{A}_1')\right]^{\perp}.$$

Because $\boldsymbol{A}_1'\boldsymbol{b} = \boldsymbol{X}\boldsymbol{c}$ for $\boldsymbol{c} := (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{A}\boldsymbol{b}$, it follows that $\mathcal{C}(\boldsymbol{A}_1')$ is a subspace of $\mathcal{C}(\boldsymbol{X})$. Therefore, we can apply the Pytheagorean property to see that

$$\begin{aligned}\widehat{\boldsymbol{\theta}}_{H_0} &= \boldsymbol{P}_{\mathcal{G}}\boldsymbol{Y} = \boldsymbol{P}_{\mathcal{C}(\boldsymbol{X})\cap[\mathcal{C}(\boldsymbol{A}_1')]^{\perp}}\boldsymbol{Y}\\ &= \boldsymbol{P}_{\mathcal{C}(\boldsymbol{X})}\boldsymbol{Y} - \boldsymbol{P}_{\mathcal{C}(\boldsymbol{A}_1')}\boldsymbol{Y}\\ &= \widehat{\boldsymbol{\theta}} - \boldsymbol{A}_1'(\boldsymbol{A}_1\boldsymbol{A}_1')^{-1}\boldsymbol{A}_1\boldsymbol{Y}.\end{aligned}$$

Now

$$\boldsymbol{A}_1\boldsymbol{A}_1' = \boldsymbol{A}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{A}' = \boldsymbol{A}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{A}'.$$

Therefore,

$$\widehat{\boldsymbol{\theta}}_{H_0} = \widehat{\boldsymbol{\theta}} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{A}'\left[\boldsymbol{A}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{A}'\right]^{-1}\boldsymbol{A}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}.$$

Aside: How do we know that $\boldsymbol{A}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{A}'$ is nonsingular? Note that $\boldsymbol{A}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{A}'$ is positive semidefinite. Now $\boldsymbol{X}'\boldsymbol{X}$ is positive definite; therefore, so is its inverse. Therefore, we can write $(\boldsymbol{X}'\boldsymbol{X})^{-1} = \boldsymbol{B}^2 = \boldsymbol{B}\boldsymbol{B}'$, where $\boldsymbol{B} := (\boldsymbol{X}'\boldsymbol{X})^{-1/2}$ is the square root of $(\boldsymbol{X}'\boldsymbol{X})^{-1}$. In this way we find that the rank of $\boldsymbol{A}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{A}'$ is the same as the rank of $\boldsymbol{A}\boldsymbol{A}'$. Since $\boldsymbol{A}$ has full rank, $\boldsymbol{A}\boldsymbol{A}'$ is invertible. Equivalently, full rank. Equivalently, $\boldsymbol{A}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{A}'$ is a full-rank positive definite matrix; hence nonsingular.

The vector $\widehat{\boldsymbol{\theta}}_{H_0} := \widehat{\boldsymbol{Y}}_{H_0} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}_{H_0}$ is the vector of fitted values, assuming that H_0 is correct. Therefore, the least-squares estimate for $\boldsymbol{\beta}$—under H_0—is

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_{H_0} &:= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\widehat{\boldsymbol{\theta}}_{H_0} = \widehat{\boldsymbol{\beta}} - (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{A}'\left[\boldsymbol{A}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{A}'\right]^{-1}\boldsymbol{A}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}\\ &= \widehat{\boldsymbol{\beta}} - (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{A}'\left[\boldsymbol{A}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{A}'\right]^{-1}\boldsymbol{A}\widehat{\boldsymbol{\beta}}\\ &= \left(\boldsymbol{I} - (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{A}'\left[\boldsymbol{A}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{A}'\right]^{-1}\boldsymbol{A}\right)\widehat{\boldsymbol{\beta}}.\end{aligned}$$

This can be generalized further as follows: Suppose we wish to test

$$H_0: \ \boldsymbol{A}\boldsymbol{\beta} = \boldsymbol{c},$$

where $\boldsymbol{c}$ is a known q-vector [we just studied this in the case that $\boldsymbol{c} = \boldsymbol{0}$]. Then we reduce the problem to the previous one as follows: First find a known p-vector $\boldsymbol{\beta}_0$ such that $\boldsymbol{A}\boldsymbol{\beta}_0 = \boldsymbol{c}$. Then, create a new parametrization of our problem by setting

$$\boldsymbol{\gamma} := \boldsymbol{\beta} - \boldsymbol{\beta}_0,$$

and

$$\widetilde{\boldsymbol{Y}} := \boldsymbol{X}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}, \quad \text{equivalently} \quad \widetilde{\boldsymbol{Y}} := \boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}_0.$$

Since $\boldsymbol{A\gamma} = \mathbf{0}$, we know the least-squares estimate $\widehat{\boldsymbol{\gamma}}_{H_0}$ is given by

$$\widehat{\boldsymbol{\gamma}}_{H_0} = \left(\boldsymbol{I} - (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{A}'\left[\boldsymbol{A}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{A}'\right]^{-1}\boldsymbol{A}\right)\widehat{\boldsymbol{\gamma}},$$

where

$$\widehat{\boldsymbol{\gamma}} := (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\widetilde{\boldsymbol{Y}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y} - \boldsymbol{\beta}_0 = \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0.$$

In other words,

$$\begin{aligned}
\widehat{\boldsymbol{\beta}}_{H_0} - \boldsymbol{\beta}_0 &= \left(\boldsymbol{I} - (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{A}'\left[\boldsymbol{A}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{A}'\right]^{-1}\boldsymbol{A}\right)\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right)\\
&= \left(\boldsymbol{I} - (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{A}'\left[\boldsymbol{A}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{A}'\right]^{-1}\boldsymbol{A}\right)\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{A}'\left[\boldsymbol{A}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{A}'\right]^{-1}\boldsymbol{A\beta}_0\\
&= \left(\boldsymbol{I} - (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{A}'\left[\boldsymbol{A}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{A}'\right]^{-1}\boldsymbol{A}\right)\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{A}\left[\boldsymbol{A}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{A}'\right]^{-1}\boldsymbol{c}.
\end{aligned}$$

In this way, we have discovered the following:

Theorem 1. *Consider once again the general linear model $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$. If $\boldsymbol{A}_{q\times p}$ and $\boldsymbol{c}_{q\times 1}$ are known, and $\boldsymbol{A}$ has full rank $q \le p$, then the least-squares estimate for $\boldsymbol{\beta}$—under the null hypothesis H_0 : $\boldsymbol{A\beta} = \boldsymbol{c}$—is*

$$\widehat{\boldsymbol{\beta}}_{H_0} = \boldsymbol{\Theta}\widehat{\boldsymbol{\beta}} + \boldsymbol{\mu},$$

where

$$\boldsymbol{\Theta} := \left(\boldsymbol{I} - (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{A}'\left[\boldsymbol{A}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{A}'\right]^{-1}\boldsymbol{A}\right), \tag{2}$$

and

$$\boldsymbol{\mu} = \boldsymbol{\mu}(\boldsymbol{c}) := (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{A}'\left[\boldsymbol{A}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{A}'\right]^{-1}\boldsymbol{c}, \tag{3}$$

provided that $\boldsymbol{X}$ has full rank.

3. The normal model

Now consider the same problem under the normal model. That is, we consider H_0 : $\boldsymbol{A\beta} = \boldsymbol{c}$ under the assumption that $\boldsymbol{\varepsilon} \sim \mathrm{N}_p(\mathbf{0}\,, \sigma^2\boldsymbol{I})$.

Theorem 2. *Consider the normal-error linear model $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$. If $\boldsymbol{A}_{q\times p}$ and $\boldsymbol{c}_{q\times 1}$ are known, and $\boldsymbol{A}$ has full rank $q \le p$, then the least-squares estimate for $\boldsymbol{\beta}$—under the null hypothesis H_0 : $\boldsymbol{A\beta} = \boldsymbol{c}$—satisfies*

$$\widehat{\boldsymbol{\beta}}_{H_0} \sim \mathrm{N}_p\left(\boldsymbol{\beta}\,, \sigma^2\boldsymbol{\Theta}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{\Theta}'\right),$$

provided that $\boldsymbol{X}$ has full rank.

Indeed, since

$$\widehat{\boldsymbol{\beta}} \sim \mathrm{N}_p\left(\boldsymbol{\beta}\,, \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}\right) \quad \text{and} \quad \widehat{\boldsymbol{\beta}}_{H_0} = \boldsymbol{\Theta}\widehat{\boldsymbol{\beta}} + \boldsymbol{\mu},$$

it follows that

$$\widehat{\boldsymbol{\beta}}_{H_0} \sim \mathrm{N}_p\left(\boldsymbol{\Theta\beta} + \boldsymbol{\mu}\,, \sigma^2\boldsymbol{\Theta}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{\Theta}'\right).$$

Therefore, it remains to check that $\boldsymbol{\Theta\beta} + \boldsymbol{\mu} = \boldsymbol{\beta}$ when $\boldsymbol{A\beta} = \boldsymbol{c}$. But this is easy to see directly.

Next we look into inference for σ^2. Recall that our estimation of σ^2 was based on RSS $:= \|\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\|^2$. Under H_0, we do the natural thing and estimate σ^2 instead by

$$\begin{aligned}\text{RSS}_{H_0} &:= \left\|\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_{H_0}\right\|^2\\ &= \left\|\underbrace{\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}}_{\mathcal{T}_1} - \underbrace{\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{A}'\left[\boldsymbol{A}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{A}'\right]^{-1}\left[\boldsymbol{c} - \boldsymbol{A}\widehat{\boldsymbol{\beta}}\right]}_{\mathcal{T}_2}\right\|^2.\end{aligned}$$

I claim that $\mathcal{T}_2$ is orthogonal to $\mathcal{T}_1$; indeed,

$$\begin{aligned}\mathcal{T}_2'\mathcal{T}_1 = &\left[\boldsymbol{c} - \boldsymbol{A}\widehat{\boldsymbol{\beta}}\right]'\left[\boldsymbol{A}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{A}'\right]^{-1}\boldsymbol{A}\overbrace{(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}}^{\widehat{\boldsymbol{\beta}}}\\ &\qquad - \left[\boldsymbol{c} - \boldsymbol{A}\widehat{\boldsymbol{\beta}}\right]'\left[\boldsymbol{A}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{A}'\right]^{-1}\boldsymbol{A}\underbrace{(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}}_{\boldsymbol{I}}\widehat{\boldsymbol{\beta}}\\ = 0.\end{aligned}$$

Therefore, the Pythagorean property tells us that

$$\begin{aligned}\text{RSS}_{H_0} &= \left\|\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\right\|^2 + \|\mathcal{T}_2\|^2\\ &= \text{RSS} + \|\mathcal{T}_2\|^2.\end{aligned}$$

Next we compute

$$\begin{aligned}\|\mathcal{T}_2\|^2 &= \mathcal{T}_2'\mathcal{T}_2\\ &= \left[\boldsymbol{c} - \boldsymbol{A}\widehat{\boldsymbol{\beta}}\right]'\left[\boldsymbol{A}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{A}'\right]^{-1}\underbrace{\underbrace{\boldsymbol{A}(\boldsymbol{X}'\boldsymbol{X})^{-1}\underbrace{\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}}_{\boldsymbol{I}}\boldsymbol{A}'}_{\boldsymbol{A}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{A}'}\left[\boldsymbol{A}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{A}'\right]^{-1}}_{\boldsymbol{I}}\left[\boldsymbol{c} - \boldsymbol{A}\widehat{\boldsymbol{\beta}}\right]\\ &= \left[\boldsymbol{c} - \boldsymbol{A}\widehat{\boldsymbol{\beta}}\right]'\left[\boldsymbol{A}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{A}'\right]^{-1}\left[\boldsymbol{c} - \boldsymbol{A}\widehat{\boldsymbol{\beta}}\right].\end{aligned}$$

In other words,

$$\text{RSS}_{H_0} = \text{RSS} + \left[\boldsymbol{c} - \boldsymbol{A}\widehat{\boldsymbol{\beta}}\right]'\left[\boldsymbol{A}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{A}'\right]^{-1}\left[\boldsymbol{c} - \boldsymbol{A}\widehat{\boldsymbol{\beta}}\right]. \qquad (4)$$

Moreover, the two terms on the right-hand side are independent because $\widehat{\boldsymbol{\beta}}$ and $\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}$—hence $\widehat{\boldsymbol{\beta}}$ and RSS $= \|\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\|^2$—are independent. Now we know the distribution of RSS $:= (n-p)S^2 \sim \sigma^2\chi^2_{n-p}$. Therefore, it remains to find the distribution of the second term on the right-hand side of (4). But

$$\boldsymbol{A}\widehat{\boldsymbol{\beta}} \sim \mathrm{N}_q\left(\boldsymbol{A\beta}\,,\sigma^2\boldsymbol{A}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{A}'\right) \overset{H_0}{=} \mathrm{N}_q\left(\boldsymbol{c}\,,\sigma^2\underbrace{\boldsymbol{A}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{A}'}_{:=\boldsymbol{\Sigma}}\right).$$

Therefore, $\boldsymbol{Z} := \sigma^{-1}\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{A}\widehat{\boldsymbol{\beta}} - \boldsymbol{c}) \sim \mathrm{N}_q(\boldsymbol{0}\,,\boldsymbol{I}_{q\times q})$. Also, we can write the second term on the right-hand side of (4) as

$$\left[\boldsymbol{A}\widehat{\boldsymbol{\beta}} - \boldsymbol{c}\right]' \left[\boldsymbol{A}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{A}'\right]^{-1} \left[\boldsymbol{A}\widehat{\boldsymbol{\beta}} - \boldsymbol{c}\right] = \sigma^2\boldsymbol{Z}'\boldsymbol{Z} = \sigma^2\|\boldsymbol{Z}\|^2 \sim \sigma^2\chi^2_q.$$

Let us summarize our efforts.

Theorem 3. *Consider normal–error linear model $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$. Suppose $\boldsymbol{A}_{q\times p}$ and $\boldsymbol{c}_{q\times 1}$ are known, and $\boldsymbol{A}$ has full rank $q \le p$. Then under the null hypothesis $H_0:\ \boldsymbol{A\beta} = \boldsymbol{c}$, we can write*

$$\mathrm{RSS}_{H_0} = \mathrm{RSS} + W,$$

provided that $\boldsymbol{X}$ has full rank, where RSS *and W are independent, we recall that* $\mathrm{RSS} \sim \sigma^2(n-p)\chi^2_{n-p}$, *and $W \sim \sigma^2\chi^2_q$. In particular,*

$$\begin{aligned}\frac{(\mathrm{RSS}_{H_0} - \mathrm{RSS})/q}{\mathrm{RSS}/(n-p)} &\overset{H_0}{\sim} \frac{\chi^2_q/q}{\chi^2_{n-p}/(n-p)} \qquad \text{[the two } \chi^2\text{'s are independent]}\\ &= F_{q,n-p}.\end{aligned}$$

See your textbook for the distribution of this test statistic under the alternative [this is useful for power computations]. The end result is a "noncentral F distribution."

4. Examples

1. A measurement-error model. For our first example, consider a random sample $Y_1,\ldots,Y_n \sim \mathrm{N}(\mu\,,\sigma^2)$; equivalently,

$$Y_i = \mu + \varepsilon_i \qquad (1 \le i \le n),$$

where $\boldsymbol{\varepsilon} \sim \mathrm{N}_n(\boldsymbol{0}\,,\sigma^2\boldsymbol{I})$. This is a linear model with $p = 1$, $\boldsymbol{X} := \boldsymbol{1}_{n\times 1}$, and $\boldsymbol{\beta} := \mu$. Recall that $(\boldsymbol{X}'\boldsymbol{X})^{-1} = 1/n$ and hence $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y} = \bar{Y}$.

If we test $H_0:\ \mu = \mu_0$ for a μ_0 the is known, then $\boldsymbol{A} = 1$ is a 1×1 matrix $(q = 1)$ and $\boldsymbol{A\beta} = \boldsymbol{c}$ with $\boldsymbol{c} = \mu_0$.

Given that H_0 is true, the least-squares estimator of μ $[\hat{\boldsymbol{\beta}}_{H_0}]$ is

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{H_0} := \hat{\mu}_{H_0} &= \hat{\boldsymbol{\beta}} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{A}'\left[\boldsymbol{A}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{A}'\right]^{-1}\left(\boldsymbol{c} - \boldsymbol{A}\hat{\boldsymbol{\beta}}\right)\\ &= \bar{Y} + \frac{1}{n}\cdot n\cdot(\mu_0 - \bar{Y}) = \mu_0.\end{aligned}$$

[Is this sensible?] And

$$\text{RSS} = \|\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\|^2 = \sum_{i=1}^{n}(Y_i - \bar{Y})^2 = ns_y^2.$$

Therefore,

$$\begin{aligned}\text{RSS}_{H_0} - \text{RSS} &= (\boldsymbol{A}\hat{\boldsymbol{\beta}} - \boldsymbol{c})' \left[\boldsymbol{A}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{A}'\right]^{-1} \left(\boldsymbol{A}\hat{\boldsymbol{\beta}} - \boldsymbol{c}\right) \\ &= n(\bar{Y} - \mu_0)^2.\end{aligned}$$

And

$$\frac{(\text{RSS}_{H_0} - \text{RSS})/q}{\text{RSS}/(n-p)} = \frac{(\bar{Y} - \mu_0)^2}{s_y^2/(n-1)} \overset{H_0}{\sim} F_{1,n-1}.$$

But

$$\frac{\bar{Y} - \mu_0}{s_y/\sqrt{n-1}} \overset{H_0}{\sim} t_{n-1}.$$

Therefore, in particular, $t_k^2 = F_{1,k}$.

2. Simple linear regression. Here,

$$Y_i = \alpha + \beta x_i + \varepsilon_i \qquad (1 \le i \le n).$$

Therefore, $p = 2$,

$$\boldsymbol{\beta} = \begin{pmatrix}\alpha \\ \beta\end{pmatrix}, \qquad \text{and} \qquad \boldsymbol{X} = \begin{pmatrix}1 & x_1 \\ \vdots & \vdots \\ 1 & x_n\end{pmatrix}.$$

Recall that the least-squares estimates of α and β are

$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{x}, \qquad \hat{\beta} = \frac{rs_y}{s_x}.$$

Now consider testing the hypothesis,

$$H_0: \ \beta = 0, \ \alpha = \mu_0,$$

where μ_0 is known.

Let $\boldsymbol{c} = (\mu_0\,, 0)'$ and $\boldsymbol{A} = \boldsymbol{I}_2$, so that $q = 2$. Then, H_0 is the same as $H_0: \ \boldsymbol{A}\boldsymbol{\beta} = \boldsymbol{c}$. We have

$$\hat{\boldsymbol{\beta}}_{H_0} = \hat{\boldsymbol{\beta}} + (\boldsymbol{c} - \boldsymbol{A}\hat{\boldsymbol{\beta}}) = \boldsymbol{c} = \begin{pmatrix}\mu_0 \\ 0\end{pmatrix}.$$

[Is this sensible?]

Now,

$$\text{RSS}_{H_0} - \text{RSS} = (\boldsymbol{A}\hat{\boldsymbol{\beta}} - \boldsymbol{c})' \left[\boldsymbol{A}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{A}'\right]^{-1} (\boldsymbol{A}\hat{\boldsymbol{\beta}} - \boldsymbol{c}) = (\hat{\boldsymbol{\beta}} - \boldsymbol{c})'(\boldsymbol{X}'\boldsymbol{X})(\hat{\boldsymbol{\beta}} - \boldsymbol{c}).$$

Now,

$$\boldsymbol{X}'\boldsymbol{X} = n\begin{pmatrix}1 & \bar{x} \\ \bar{x} & \overline{x^2}\end{pmatrix} \quad \Rightarrow \quad \hat{\boldsymbol{\beta}} - \boldsymbol{c} = \begin{pmatrix}\bar{Y} - \hat{\beta}\bar{x} - \mu_0 \\ \hat{\beta}\end{pmatrix}.$$

Therefore,

$$(\hat{\boldsymbol{\beta}} - \boldsymbol{c})'(\boldsymbol{X}'\boldsymbol{X}) = n(\bar{Y} - \mu_0)(1\,, \bar{x}),$$

whence

$$\mathrm{RSS}_{H_0} - \mathrm{RSS} = n(\bar{Y} - \mu_0)^2.$$

Next we compute

$$\mathrm{RSS} = \|\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\|^2 = \sum_{i=1}^n \left[Y_i - (\boldsymbol{X}\hat{\boldsymbol{\beta}})_i\right]^2.$$

Since

$$\boldsymbol{X}\hat{\boldsymbol{\beta}} = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \begin{pmatrix} \bar{Y} - \hat{\beta}\bar{x} \\ \hat{\beta} \end{pmatrix} = \left(\bar{Y} + \hat{\beta}(x_i - \bar{x})\right)_{i=1}^n = \left(\bar{Y} + \frac{rs_y}{s_x}(x_i - \bar{x})\right)_{i=1}^n,$$

it follows that

$$\begin{aligned} \mathrm{RSS} &= \sum_{i=1}^n \left[(Y_i - \bar{Y})^2 + \frac{r^2 s_y^2}{s_x^2}(x_i - \bar{x})^2 - 2\frac{rs_y}{s_x}(Y_i - \bar{Y})(x_i - \bar{x})\right] \\ &= ns_y^2 + nr^2 s_y^2 - \frac{2nrs_y}{s_x}\sum_{i=1}^n (Y_i - \bar{Y})(x_i - \bar{x}) \\ &= ns_y^2 + nr^2 s_y^2 - 2nr^2 s_y^2 \\ &= ns_y^2(1 - r^2). \end{aligned}$$

Therefore,

$$\frac{(\bar{Y} - \mu_0)^2}{s_y^2(1 - r^2)} \overset{H_0}{\sim} F_{2,n-2}.$$

3. Two-sample mean. Consider two populations: $\mathrm{N}(\mu_1\,, \sigma^2)$ and $\mathrm{N}(\mu_2\,, \sigma^2)$ with equal variances. We wish to know if $\mu_1 = \mu_2$. Take two independent random samples,

$$\begin{aligned} y_{1,1}, \ldots, y_{1,n_1} &\sim \mathrm{N}(\mu_1\,, \sigma^2), \\ y_{2,1}, \ldots, y_{2,n_2} &\sim \mathrm{N}(\mu_2\,, \sigma^2). \end{aligned}$$

We have a linear model: $p = 2$, $n = n_1 + n_2$,

$$\boldsymbol{Y} = \begin{pmatrix} y_{1,1} \\ \vdots \\ y_{1,n_1} \\ y_{2,1} \\ \vdots \\ y_{2,n_2} \end{pmatrix}, \qquad \boldsymbol{X} = \begin{pmatrix} 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 0 & 1 \\ \vdots & \vdots \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} \mathbf{1}_{n_1\times 1} & \mathbf{0}_{n_1\times 1} \\ \mathbf{0}_{n_2\times 1} & \mathbf{1}_{n_2\times 1} \end{pmatrix}.$$

In particular,

$$\boldsymbol{X}'\boldsymbol{X} = \begin{pmatrix} n_1 & 0 \\ 0 & n_2 \end{pmatrix} \quad \Rightarrow \quad (\boldsymbol{X}'\boldsymbol{X})^{-1} = \begin{pmatrix} n_1^{-1} & 0 \\ 0 & n_2^{-1} \end{pmatrix}.$$

So now consider

$$H_0 :\ \mu_1 = \mu_2 \qquad \Longleftrightarrow \qquad H_0 :\ \mu_1 = \mu_2 \qquad \Longleftrightarrow \qquad H_0 :\ \underbrace{(1\,,-1)}_{\boldsymbol{A}} \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} = \mathbf{0}.$$

That is, $q = 1$, $\boldsymbol{A} := (1\,,-1)$, and $\boldsymbol{c} = 0$. In this way we find that

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y} = \begin{pmatrix} \bar{y}_{1,\bullet} \\ \bar{y}_{2,\bullet} \end{pmatrix}.$$

[Does this make intuitive sense?]

In order to find $\hat{\boldsymbol{\beta}}_{H_0}$, we first compute

$$\boldsymbol{A}\hat{\boldsymbol{\beta}} = (1\,,-1) \begin{pmatrix} \bar{y}_{1,\bullet} \\ \bar{y}_{2,\bullet} \end{pmatrix} = \bar{y}_{1,\bullet} - \bar{y}_{2,\bullet}.$$

Also,

$$(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{A}' = \begin{pmatrix} n_1^{-1} & 0 \\ 0 & n_2^{-1} \end{pmatrix} \begin{pmatrix} 1 \\ -1 \end{pmatrix} = \begin{pmatrix} n_1^{-1} \\ -n_2^{-1} \end{pmatrix},$$

so that

$$\boldsymbol{A}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{A}' = \frac{1}{n_2} + \frac{1}{n_2} = \frac{n}{n_1 n_2}.$$

Now we put things together:

$$\begin{aligned} \hat{\boldsymbol{\beta}}_{H_0} &= \hat{\boldsymbol{\beta}} + \begin{pmatrix} n_1^{-1} \\ -n_2^{-1} \end{pmatrix} \frac{n_1 n_2}{n} \left(\bar{y}_{1,\bullet} - \bar{y}_{2,\bullet}\right) \\ &= \hat{\boldsymbol{\beta}} + \begin{pmatrix} \dfrac{n_2}{2} \left(\bar{y}_{1,\bullet} - \bar{y}_{2,\bullet}\right) \\ \\ -\dfrac{n_1}{n} \left(\bar{y}_{1,\bullet} - \bar{y}_{2,\bullet}\right) \end{pmatrix} \\ &= \begin{pmatrix} \dfrac{n_1}{n}\bar{y}_{1,\bullet} + \dfrac{n_2}{n}\bar{y}_{2,\bullet} \\ \\ \dfrac{n_1}{n}\bar{y}_{1,\bullet} + \dfrac{n_2}{n}\bar{y}_{2,\bullet} \end{pmatrix}. \end{aligned}$$

Since $n_2\bar{y}_{2,\bullet} = \sum_{j=1}^{n_2} y_{j2,j}$ and $n_1\bar{y}_{1,\bullet} = \sum_{i=1}^{n_1} y_{1,j}$, it follows that

$$\hat{\boldsymbol{\beta}}_{H_0} = \begin{pmatrix} \bar{y}_{\bullet,\bullet} \\ \bar{y}_{\bullet,\bullet} \end{pmatrix}.$$

[Does this make sense?] Since

$$\boldsymbol{X}\hat{\boldsymbol{\beta}} = \begin{pmatrix} 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 0 & 1 \\ \vdots & \vdots \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \bar{y}_{1,\bullet} \\ \bar{y}_{2,\bullet} \end{pmatrix} = \begin{pmatrix} \bar{y}_{1,\bullet} \\ \vdots \\ \bar{y}_{1,\bullet} \\ \bar{y}_{2\bullet} \\ \vdots \\ \bar{y}_{2,\bullet} \end{pmatrix} = \begin{pmatrix} \bar{y}_{1\bullet}\mathbf{1}_{n_1\times 1} \\ \bar{y}_{2,\bullet}\mathbf{1}_{n_2\times 1} \end{pmatrix},$$

we have

$$\begin{aligned} \text{RSS} &= \sum_{j=1}^{n_1} \left(y_{1,j} - \bar{y}_{1\bullet}\right)^2 + \sum_{j=1}^{n_2} \left(y_{2,j} - \bar{y}_{2\bullet}\right)^2 \\ &= n_1 s_1^2 + n_2 s_2^2. \end{aligned}$$

I particular,

$$\frac{\text{RSS}}{n-p} = \frac{n_1}{n-2} s_1^2 + \frac{n_2}{n-2} s_2^2 := s_p^2$$

is the socalled “pooled variance.”

Similarly,

$$\text{RSS}_{H_0} - \text{RSS} = \frac{n_1 n_2}{n} \left(\bar{y}_{1,\bullet} - \bar{y}_{2,\bullet}\right)^2.$$

Therefore,

$$\frac{\frac{n_1 n_2}{n} \left(\bar{y}_{1,\bullet} - \bar{y}_{2,\bullet}\right)^2}{s_p^2} \overset{H_0}{\sim} F_{1,n-2} \quad \Rightarrow \quad \frac{\bar{y}_{1,\bullet} - \bar{y}_{2,\bullet}}{s_p\sqrt{\frac{n}{n_1 n_2}}} \overset{H_0}{\sim} t_{n-2}.$$

4. ANOVA: One-way layout. Consider p populations that are respectively distributed as $\mathrm{N}(\mu_1\,,\sigma^2), \ldots, \mathrm{N}(\mu_p\,,\sigma^2)$. We wish to test

$$H_0: \ \mu_1 = \cdots = \mu_p.$$

We have seen that we are in the setting of linear models, so we can compute $\hat{\boldsymbol{\beta}}_{H_0}$ etc. that way. I will leave this up to you and compute directly instead: Sample $y_{j,1}\,, \ldots y_{j,n_k}$ i.i.d. $\mathrm{N}(\mu_j\,,\sigma^2)$ [independent also as j varies]. Then we vectorize:

$$\boldsymbol{Y} := \begin{pmatrix} y_{1,1} \\ \vdots \\ y_{1,n_1} \\ \vdots \\ y_{p,1} \\ \vdots \\ y_{p,n_p} \end{pmatrix}; \qquad \text{etc.}$$

Instead we now find $\hat{\boldsymbol{\beta}}$ directly by solving

$$\min_{\boldsymbol{\mu}} \sum_{i=1}^{p} \sum_{j=1}^{n_i} \left(y_{i,j} - \mu_i\right)^2 .$$

That is, compute

$$\frac{\partial}{\partial \mu_i} \sum_{i=1}^{p} \sum_{j=1}^{n_i} \left(y_{i,j} - \mu_i\right)^2 = -\sum_{j=1}^{n_i} 2\left(y_{i,j} - \mu_i\right) \equiv 0 \implies \hat{\mu}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{i,j} = \bar{y}_{i,\bullet}.$$

This yields

$$\hat{\boldsymbol{\beta}} = \begin{pmatrix} \bar{y}_{1,\bullet} \\ \vdots \\ \bar{y}_{n,\bullet} \end{pmatrix} .$$

What about $\hat{\boldsymbol{\beta}}_{H_0}$? Under H_0, $\mu_1 = \cdots = \mu_p \equiv \mu$ and so $q = p - 1$. So we have

$$\min_{\mu} \sum_{i=1}^{p} \sum_{j=1}^{n_i} \left(y_{i,j} - \mu\right)^2 \implies \hat{\boldsymbol{\beta}}_{H_0} = \begin{pmatrix} \bar{y}_{\bullet,\bullet} \\ \vdots \\ \bar{y}_{\bullet,\bullet} \end{pmatrix} .$$

Also,

$$\mathrm{RSS} = \sum_{i=1}^{p} \sum_{j=1}^{n_i} \left(y_{i,j} - \bar{y}_{i,\bullet}\right)^2 ,$$

and

$$\begin{aligned}
\mathrm{RSS}_{H_0} &= \sum_{i=1}^{p} \sum_{j=1}^{n_i} \left(y_{i,j} - \bar{y}_{\bullet,\bullet}\right)^2 = \sum_{i=1}^{p} \sum_{j=1}^{n_i} \left(y_{i,j} - \bar{y}_{i,\bullet} + \bar{y}_{i,\bullet} - \bar{y}_{\bullet,\bullet}\right)^2 \\
&= \sum_{i=1}^{p} \sum_{j=1}^{n_i} \left(y_{i,j} - \bar{y}_{i,\bullet}\right)^2 + 2 \sum_{i=1}^{p} \sum_{j=1}^{n_i} \left(y_{i,j} - \bar{y}_{i,\bullet}\right) \left(\bar{y}_{i,\bullet} - \bar{y}_{\bullet,\bullet}\right) \\
&\qquad\qquad + \sum_{i=1}^{p} \sum_{j=1}^{n_i} \left(\bar{y}_{i,\bullet} - \bar{y}_{\bullet,\bullet}\right)^2 \\
&= \sum_{i=1}^{p} \sum_{j=1}^{n_i} \left(y_{i,j} - \bar{y}_{i,\bullet}\right)^2 + \sum_{i=1}^{p} \sum_{j=1}^{n_i} \left(\bar{y}_{i,\bullet} - \bar{y}_{\bullet,\bullet}\right)^2 \\
&= \mathrm{RSS} + \sum_{i=1}^{p} n_i \left(\bar{y}_{i,\bullet} - \bar{y}_{\bullet,\bullet}\right)^2 .
\end{aligned}$$

It follows from the general theory that

$$\frac{\sum_{i=1}^{p} n_i \left(\bar{y}_{i,\bullet} - \bar{y}_{\bullet,\bullet}\right)^2 / (p-1)}{\sum_{i=1}^{p} \sum_{j=1}^{n_i} \left(y_{i,j} - \bar{y}_{i,\bullet}\right)^2 / (n-p)} \overset{H_0}{\sim} F_{p-1,n-p}.$$

“Statistical interpretation”:

$$\frac{\sum_{i=1}^{p} n_i \left(\bar{y}_{i,\bullet} - \bar{y}_{\bullet,\bullet}\right)^2}{p-1} = \text{The variation between the samples;}$$

whereas

$$\frac{\sum_{i=1}^{p} \sum_{j=1}^{n_i} \left(y_{i,j} - \bar{y}_{i,\bullet}\right)^2}{n-p} = \text{The variation within the samples.}$$

Therefore,

$$\mathrm{RSS}_{H_0} = \text{Variation between} + \text{Variation within} = \text{Total variation.}$$

Confidence Intervals and Sets

Throughout we adopt the normal-error model, and wish to say some things about the construction of confidence intervals [and sets] for the parameters $\beta_0, \ldots, \beta_{p-1}$.

1. Confidence intervals for one parameter

Suppose we want a confidence intervals for the ith parameter β_i, where $1 \le i \le p$ is fixed. Recall that

$$\frac{\hat\beta_i - \beta_i}{\sqrt{\sigma^2\left[(\boldsymbol{X}'\boldsymbol{X})^{-1}\right]_{i,i}}} \sim \mathrm{N}(0\,,1),$$

and

$$S^2 := \frac{\mathrm{RSS}}{n-p} = \frac{1}{n-p}\sum_{i=1}^n \left(Y_i - \left(\boldsymbol{X}\hat{\boldsymbol{\beta}}\right)_i\right)^2 \sim \frac{\sigma^2\chi^2_{n-p}}{n-p},$$

and is independent of $\hat{\boldsymbol{\beta}}$. Therefore,

$$\frac{\hat\beta_i - \beta_i}{\sqrt{S^2\left[(\boldsymbol{X}'\boldsymbol{X})^{-1}\right]_{i,i}}} \sim \frac{\mathrm{N}(0\,,1)}{\sqrt{\chi^2_{n-p}/(n-p)}} = t_{n-p},$$

where the normal and χ^2 random variables are independent. Therefore,

$$\hat\beta_i \pm t^{(\alpha/2)}_{n-p} S\sqrt{\left[(\boldsymbol{X}'\boldsymbol{X})^{-1}\right]_{i,i}}$$

is a $(1-\alpha)\times 100\%$ confidence interval for β_i. This yields a complete analysis of confidence intervals for a univariate parameter; these confidence intervals can also be used for testing, of course.

2. Confidence ellipsoids

The situation is more interesting if we wish to say something about more than one parameter at the same time. For example, suppose we want to know about $(\beta_0\,,\beta_1)$ in the overly-simplified case that $\sigma = 1$. The general philosophy of confidence intervals [for univariate parameters] suggests that we look for a random set Ω such that

$$\mathrm{P}\left\{\begin{pmatrix}\hat\beta_1\\ \hat\beta_2\end{pmatrix} - \begin{pmatrix}\beta_1\\ \beta_2\end{pmatrix} \in \Omega\right\} = 1-\alpha.$$

And we know that

$$\begin{pmatrix}\hat\beta_1\\ \hat\beta_2\end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix}\beta_1\\ \beta_2\end{pmatrix}\,, \boldsymbol{\Sigma}\right),$$

where $\boldsymbol{\Sigma}$ is a 2×2 matrix with

$$\Sigma_{i,j} = \left[(\boldsymbol{X}'\boldsymbol{X})^{-1}\right]_{i,j}.$$

Here is a possible method: Let

$$\hat{\boldsymbol{\gamma}} := \begin{pmatrix}\hat\beta_1 - \beta_1\\ \hat\beta_2 - \beta_2\end{pmatrix}, \quad \text{so that} \quad \hat{\boldsymbol{\gamma}} \sim \mathrm{N}_2\left(\mathbf{0}\,, \boldsymbol{\Sigma}\right).$$

Also, recall that

$$\hat{\boldsymbol{\gamma}}'\boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{\gamma}} \sim \chi^2_2.$$

Therefore, one natural choice for Ω is

$$\Omega := \left\{\boldsymbol{x} \in \mathbf{R}^2 :\ \boldsymbol{x}'\boldsymbol{\Sigma}^{-1}\boldsymbol{x} \le \chi^2_2(\alpha/2)\right\}.$$

What does Ω look like? In order to answer this, let us apply the spectral theorem:

$$\boldsymbol{\Sigma} = \boldsymbol{PDP}' \qquad \Longrightarrow \qquad \boldsymbol{\Sigma}^{-1} = \boldsymbol{PD}^{-1}\boldsymbol{P}'.$$

Then we can represent Ω as follows:

$$\begin{aligned}\Omega &= \left\{\boldsymbol{x} \in \mathbf{R}^2 :\ \boldsymbol{x}'\boldsymbol{PD}^{-1}\boldsymbol{P}'\boldsymbol{x} \le \chi^2_2(\alpha/2)\right\}\\ &= \left\{\boldsymbol{x} \in \mathbf{R}^2 :\ (\boldsymbol{P}'\boldsymbol{x})\boldsymbol{D}^{-1}(\boldsymbol{P}'\boldsymbol{x}) \le \chi^2_2(\alpha/2)\right\}\\ &= \left\{\boldsymbol{Py} \in \mathbf{R}^2 :\ \boldsymbol{y}'\boldsymbol{D}^{-1}\boldsymbol{y} \le \chi^2_2(\alpha/2)\right\}\\ &= \left\{\boldsymbol{Py} \in \mathbf{R}^2 :\ \frac{y_1^2}{\lambda_1} + \frac{y_2^2}{\lambda_2} \le \chi^2_2(\alpha/2)\right\}.\end{aligned}$$

Consider

$$\mathcal{E} := \left\{\boldsymbol{y} \in \mathbf{R}^2 :\ \frac{y_1^2}{\lambda_1} + \frac{y_2^2}{\lambda_2} \le \chi^2_2(\alpha/2)\right\}. \tag{1}$$

This is the interior of an ellipsoid, and $\Omega = \boldsymbol{P}\mathcal{E}$ is the image of the ellipsoid under the "linear orthogonal map" $\boldsymbol{P}$. Such sets are called "generalized ellipsoids," and we have found a $(1-\alpha)\times 100\%$ confidence [generalized] ellipsoid for $(\beta_1\,,\beta_2)'$.

The preceding can be generalized to any number of the parameters $\beta_{i_1},\ldots,\beta_{i_k}$, but is hard to work with, as the geometry of Ω can be complicated [particularly if $k \gg 2$]. Therefore, instead we might wish to look for approximate confidence sets that are easier to work with. Before we move on though, let me mention that if you want to know whether or not $H_0:\ \beta_1 = \beta_{1,0},\ \beta_2 = \beta_{2,0}$, then we can use these confidence bounds fairly easily, since it is not hard to check whether or not $(\beta_{1,0}\,,\beta_{2,0})'$ is in Ω: You simply compute the scalar quantity

$$(\beta_{1,0}\,,\beta_{2,0})\,\boldsymbol{\Sigma}^{-1}\begin{pmatrix}\beta_{1,0}\\ \beta_{2,0}\end{pmatrix},$$

and check to see if it is $\le \chi^2_2(\alpha/2)$! But if you really need to imagine or see the confidence set[s], then this exact method can be unwieldy [particularly in higher dimensions than 2].

3. Bonferonni bounds

Our approximate confidence intervals are based on a fact from general probability theory.

Proposition 1 (Bonferonni's inequality)**.** *Let $E_1,\ldots,E_k$ be k events. Then,*

$$\mathrm{P}\left(E_1\cap\cdots\cap E_k\right) \ge 1-\sum_{j=1}^k \mathrm{P}(E_j^c) = 1-\sum_{j=1}^k\left(1-\mathrm{P}(E_j)\right).$$

Proof. The event $E_1^c\cup\cdots\cup E_k^c$ is the complement of $E_1\cap\cdots\cap E_k$. Therefore,

$$\mathrm{P}\left(E_1\cap\cdots\cap E_k\right) = 1-\mathrm{P}\left(E_1^c\cup\cdots\cup E_k^c\right),$$

and this is $\ge 1-\sum_{j=1}^k \mathrm{P}(E_j^c)$ because the probability of a union is at most the sum of the individual probabilities. □

Here is how we can use Bonferonni's inequality. Define

$$C_j := \hat\beta_j \pm t_{n-p}^{(\alpha/4)} S\sqrt{[(\boldsymbol{X}'\boldsymbol{X})^{-1}]_{j,j}} \qquad (j=1,2).$$

We have seen already that

$$\mathrm{P}\{\beta_j \in C_j\} = 1-\frac{\alpha}{2}.$$

[This is why we used $\alpha/4$ in the definition of C_j.] Therefore, Bonferonni's inequality implies that

$$\mathrm{P}\{\beta_1\in C_1\ ,\ \beta_2\in C_2\} \ge 1-\left[\frac{\alpha}{2}+\frac{\alpha}{2}\right] = 1-\alpha.$$

In other words, $(C_1\,, C_2)$ forms a "conservative" $(1-\alpha)\times 100\%$ simultaneous confidence interval for $(\beta_1\,,\beta_2)$. This method becomes very inaccurate quickly as the number of parameters of interest grows. For instance, if you want Bonferonni confidence sets for $(\beta_1\,,\beta_2\,,\beta_3)$, then you need to use individual confidence intervals with confidence level $1-(\alpha/3)$ each. And for k parameters you need individual confidence level $1-(\alpha/k)$, which can yield bad performance when k is large.

This method is easy to implement, but usually **very** conservative.[1]

4. Scheffé's simultaneous conservative confidence bounds

There is a lovely method, due to Scheffé, that works in a similar fashion to the Bonferonni method; but has also the advantage of being often [far] less conservative! [We are now talking strictly about our linear model.] The starting point of this discussion is a general fact from matrix analysis.

Proposition 2 (The Rayleigh—Ritz inequality)**.** *If $\boldsymbol{L}_{p\times p}$ is positive definite, then for every p-vector $\boldsymbol{b}$,*

$$\boldsymbol{b}'\boldsymbol{L}^{-1}\boldsymbol{b} = \max_{\boldsymbol{h}\neq\boldsymbol{0}}\left[\frac{(\boldsymbol{h}'\boldsymbol{b})^2}{\boldsymbol{h}'\boldsymbol{L}\boldsymbol{h}}\right].$$

The preceding is called an inequality because it says that

$$\boldsymbol{b}'\boldsymbol{L}^{-1}\boldsymbol{b} \le \frac{(\boldsymbol{h}'\boldsymbol{b})^2}{\boldsymbol{h}'\boldsymbol{L}\boldsymbol{h}} \qquad \text{for all } \boldsymbol{h}_{p\times 1},$$

and it also tells us that the inequality is achieved for some $\boldsymbol{h}\in\mathbf{R}^p$.

Proof of Rayleigh–Ritz inequality. Recall the Cauchy–Schwarz inequality from your linear algebra course: $(\boldsymbol{u}'\boldsymbol{v})^2 \le \|\boldsymbol{u}\|^2\cdot\|\boldsymbol{v}\|^2$ with equality if and only if $\boldsymbol{v} = a\boldsymbol{u}$ for some $a\in\mathbf{R}$. It follows then that

$$\max_{\boldsymbol{v}\neq\boldsymbol{0}}\left[\frac{(\boldsymbol{u}'\boldsymbol{v})^2}{\boldsymbol{v}'\boldsymbol{v}}\right] = \|\boldsymbol{u}\|^2.$$

Now write $\boldsymbol{L} := \boldsymbol{P}\boldsymbol{D}\boldsymbol{P}'$, in spectral form, and change variables:

$$\boldsymbol{v} := \left(\boldsymbol{L}^{1/2}\right)'\boldsymbol{h},$$

in order to see that

$$\boldsymbol{u}'\boldsymbol{v} = \boldsymbol{u}'\left(\boldsymbol{L}^{1/2}\right)'\boldsymbol{h}, \qquad \boldsymbol{v}'\boldsymbol{v} = \boldsymbol{h}'\boldsymbol{L}\boldsymbol{h},$$

[1] On the other hand, the Bonferonni method can be applied to a wide range of statistical problems that involve simultaneous confidence intervals [and is not limited to the theory of linear models. So it is well worth your while to understand this important method.

and so

$$\|\boldsymbol{u}\|^2 = \max_{\boldsymbol{h}\neq\boldsymbol{0}} \left[\frac{\left(\boldsymbol{u}' \left(\boldsymbol{L}^{1/2} \right)' \boldsymbol{h} \right)^2}{\boldsymbol{h}'\boldsymbol{L}\boldsymbol{h}} \right].$$

Change variables again: $\boldsymbol{b} := \boldsymbol{L}^{1/2}\boldsymbol{u}$ to see that

$$\|\boldsymbol{u}\|^2 = \left\| \boldsymbol{L}^{-1/2}\boldsymbol{b} \right\|^2 = \boldsymbol{b}'\boldsymbol{L}^{-1}\boldsymbol{b}.$$

This does the job. □

Now let us return to Scheffé's method. Recall that

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \sim \mathrm{N}_p \left(\boldsymbol{0}\,, \sigma^2 (\boldsymbol{X}'\boldsymbol{X})^{-1} \right),$$

so that

$$\frac{1}{\sigma^2} \left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \right)' (\boldsymbol{X}'\boldsymbol{X}) \left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \right) \sim \chi^2_p,$$

and hence

$$\frac{\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \right)' (\boldsymbol{X}'\boldsymbol{X}) \left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \right)}{pS^2} = \frac{\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \right)' (\boldsymbol{X}'\boldsymbol{X}) \left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \right)/p}{S^2} \sim F_{p,n-p}.$$

We apply the Rayleigh–Ritz inequality with $\boldsymbol{b} := \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}$ and $\boldsymbol{L} := (\boldsymbol{X}'\boldsymbol{X})^{-1}$ in order to find that

$$\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \right)' (\boldsymbol{X}'\boldsymbol{X}) \left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \right) = \max_{\boldsymbol{h}\neq\boldsymbol{0}} \left[\frac{\left\{ \boldsymbol{h}' \left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \right) \right\}^2}{\boldsymbol{h}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{h}} \right].$$

Therefore,

$$\mathrm{P}\left\{ \frac{1}{pS^2} \max_{\boldsymbol{h}\neq\boldsymbol{0}} \left[\frac{\left\{ \boldsymbol{h}' \left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \right) \right\}^2}{\boldsymbol{h}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{h}} \right] \leq F_{p,n-p}(\alpha) \right\} = 1 - \alpha.$$

Equivalently,

$$\mathrm{P}\left\{ \frac{\left| \boldsymbol{h}'\hat{\boldsymbol{\beta}} - \boldsymbol{h}'\boldsymbol{\beta} \right|}{\sqrt{\boldsymbol{h}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{h}}} \leq \sqrt{pS^2\, F_{p,n-p}(\alpha)} \text{ for all } \boldsymbol{h} \in \mathbf{R}^p \right\} = 1 - \alpha.$$

If we restrict attention to a subcollection of $\boldsymbol{h}$'s then the probability is even more. In particular, consider only $\boldsymbol{h}$'s that are the standard basis vectors of $\mathbf{R}^p$, in order to deduce from the preceding that

$$\mathrm{P}\left\{ \frac{\left| \hat{\beta}_j - \beta_j \right|}{\sqrt{[(\boldsymbol{X}'\boldsymbol{X})]_{j,j}}} \leq \sqrt{pS^2\, F_{p,n-p}(\alpha)} \text{ for all } 1 \leq j \leq p \right\} \geq 1 - \alpha.$$

In other words, we have demonstrated the following.

Theorem 3 (Scheffé)**.** *The following are conservative* $(1-\alpha)\times 100\%$ *simultaneous confidence bounds for* $(\beta_1\,,\ldots,\beta_p)'$*:*

$$\hat\beta_j \pm \sqrt{pS^2\,[(\boldsymbol{X}'\boldsymbol{X})^{-1}]_{j,j}\;F_{p,n-p}(\alpha)} \qquad (1\le j\le p).$$

When the sample size n is very large, the preceding yields an asymptotic simplification. Recall that $\chi^2_{n-p}/(n-p)\to 1$ as $n\to\infty$. Therefore, $F_{p,n-p}(\alpha)\approx\chi^2_p(\alpha)/p$ for $n\gg 1$. Therefore, the following are asymptotically-conservative simultaneous confidence bounds:

$$\hat\beta_j \pm S\sqrt{[(\boldsymbol{X}'\boldsymbol{X})^{-1}]_{j,j}\;\chi^2_p(\alpha)} \qquad (1\le j\le p) \text{ for } n\gg 1.$$

5. Confidence bounds for the regression surface

Given a vector $\boldsymbol{x}\in\mathbf{R}^p$ of predictor variables, our linear model yields $\mathrm{E}\boldsymbol{y}=\boldsymbol{x}'\boldsymbol{\beta}$. In other words, we can view our efforts as one about trying to understand the unknown function

$$f(\boldsymbol{x}) := \boldsymbol{x}'\boldsymbol{\beta}.$$

And among other things, we have found the following estimator for f:

$$\hat f(\boldsymbol{x}) := \boldsymbol{x}'\hat{\boldsymbol{\beta}}.$$

Note that if $\boldsymbol{x}\in\mathbf{R}^p$ is held fixed, then

$$\hat f(\boldsymbol{x}) \sim \mathrm{N}\left(f(\boldsymbol{x})\,,\sigma^2\boldsymbol{x}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}\right).$$

Therefore, a $(1-\alpha)\times 100\%$ confidence interval for $f(\boldsymbol{x})$ [for a fixed predictor variable $\boldsymbol{x}$] is

$$\hat f(\boldsymbol{x}) \pm S\sqrt{\boldsymbol{x}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}}\;t^{(\alpha/2)}_{n-p}.$$

That is, if we are interested in a confidence interval for $f(\boldsymbol{x})$ for a fixed $\boldsymbol{x}$, then we have

$$\mathrm{P}\left\{f(\boldsymbol{x}) = \hat f(\boldsymbol{x}) \pm S\sqrt{\boldsymbol{x}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}}\;t^{(\alpha/2)}_{n-p}\right\} = 1-\alpha.$$

On the other hand, we can also apply Scheffé's method and obtain the following simultaneous $(1-\alpha)\times 100\%$ confidence set:

$$\mathrm{P}\left\{f(\boldsymbol{x}) = \hat f(\boldsymbol{x}) \pm S\sqrt{p\boldsymbol{x}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}\,F_{p,n-p}(\alpha)} \text{ for all } \boldsymbol{x}\in\mathbf{R}^p\right\} \ge 1-\alpha.$$

Example 4 (Simple linear regression)**.** Consider the basic regression model

$$Y_i = \alpha+\beta x_i+\varepsilon_i \qquad (1\le i\le n).$$

If $\boldsymbol{w}=(1\,,w)'$, then a quick computation yields

$$\boldsymbol{w}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{w} = \frac{\left[\overline{x^2}-2w\bar x+w^2\right]}{ns_x^2},$$

where we recall $s_x^2 := \sum_{i=1}^p (x_i - \bar{x})^2$. Therefore, a simultaneous confidence interval for all of $\alpha + \beta w$'s, as w ranges over $\mathbf{R}$, is

$$\hat{\alpha} + \hat{\beta} w \pm S\sqrt{\frac{2}{ns_x^2}\left[\overline{x^2} - 2w\bar{x} + w^2\right] F_{2,n-2}(\alpha)}.$$

This expression can be simplified further because:

$$\begin{aligned}\overline{x^2} - 2w\bar{x} + w^2 &= \overline{x^2} - (\bar{x})^2 + (\bar{x})^2 - 2w\bar{x} + w^2\\ &= s_x^2 + (\bar{x} - w)^2.\end{aligned}$$

Therefore, with probability $1 - \alpha$, we have

$$\alpha + \beta w = \hat{\alpha} + \hat{\beta} w \pm S\sqrt{2\left(\frac{1}{n} + \frac{(\bar{x} - w)^2}{ns_x^2}\right) F_{2,n-2}(\alpha)} \qquad \text{for all } w.$$

On the other hand, if we want a confidence interval for $\alpha + \beta w$ for a fixed w, then we can do better using our t-test:

$$\alpha + \beta w = \hat{\alpha} + \hat{\beta} w \pm S\sqrt{\left(\frac{1}{n} + \frac{(\bar{x} - w)^2}{ns_x^2}\right)} t_{n-2}^{(\alpha/2)}.$$

You should check the details of this computation. □

6. Prediction intervals

The difference between confidence and prediction intervals is this: For a confidence interval we try to find an interval around the **parameter** $\boldsymbol{x}'\boldsymbol{\beta}$. For a prediction interval we do so for the **random variable** $y_0 := \boldsymbol{x}_0'\boldsymbol{\beta} + \varepsilon_0$, where $\boldsymbol{x}_0$ is known and fixed and ε_0 is the "noise," which is hitherto unobserved (i.e., independent of the vector $\boldsymbol{Y}$ of observations).

It is not hard to construct a good prediction interval of this type: Note that

$$\hat{y}_0 := \boldsymbol{x}_0'\hat{\boldsymbol{\beta}}$$

satisfies

$$\hat{y}_0 - \boldsymbol{y}_0 = \boldsymbol{x}_0'\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) - \varepsilon_0 \sim \mathrm{N}\left(0\,, \sigma^2\left[\boldsymbol{x}_0'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_0 + 1\right]\right).$$

Therefore, a prediction interval is

$$\hat{y}_0 \pm S\, t_{n-p}^{(\alpha/2)}\sqrt{\boldsymbol{x}_0'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_0 + 1}.$$

This means that

$$\mathrm{P}\left\{y_0 \in \hat{y}_0 \pm S\, t_{n-p}^{(\alpha/2)}\sqrt{\boldsymbol{x}_0'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_0 + 1}\right\} = 1 - \alpha,$$

but note that both y_0 and the prediction interval are now random.

Lecture 11

Polynomial Regression

1. Introduction

Here is a careful statement of the Weierstrass approximation theorem, which states in loose terms that every continuous function on a bounded set can be approximated arbitrarily well by a polynomial.

Theorem 1 (Weierstrass Approximation Theorem)**.** *If $f:[a\,,b]\to\mathbf{R}$ is continuous, then for all $\epsilon>0$ there exists a polynomial P such that*

$$\max_{a\le x\le b}|f(x)-P(x)|<\epsilon.$$

Therefore, if we can trying to estimate a variable y using only a variable x, then it makes good sense to try to fit polynomials to our regression fit. The general linear model becomes, in this case,

$$Y_i=\beta_0+\beta_1x_i+\beta_2x_i^2+\cdots+\beta_kx_i^k+\varepsilon_i \qquad (1\le i\le n).$$

This is indeed a linear model, as we have seen, with design matrix

$$\boldsymbol{X}=\begin{pmatrix}1 & x_1 & x_1^2 & \cdots & x_1^k\\ \vdots & \vdots & \vdots & & \vdots\\ 1 & x_n & x_n^2 & \cdots & x_n^k\end{pmatrix}$$

A practical problem that arises here is this: If k is large then $\boldsymbol{X}'\boldsymbol{X}$ is "ill-conditioned" and therefore very hard to invert accurately. For instance, suppose $x_1,\ldots,x_n$ are spaced uniformly in $(0\,,1)$, so that $x_1={}^1\!/n$, $x_2={}^2\!/n$, $\ldots$, $x_n={}^n\!/n=1$. In that case,

$$(\boldsymbol{X}'\boldsymbol{X})_{i,j}=\sum_{k=1}^n X_{k,i}X_{k,j}=\sum_{k=1}^n x_k^ix_k^j=\sum_{k=1}^n\left(\frac{k}{n}\right)^i\left(\frac{k}{n}\right)^j=n\times\frac{1}{n}\sum_{k=1}^n\left(\frac{k}{n}\right)^{i+j}.$$

Riemann's theory of integral tells us that

$$\frac{1}{n}\sum_{k=1}^{n}\left(\frac{k}{n}\right)^{i+j} \simeq \int_0^1 x^{i+j}\,\mathrm{d}x = \frac{1}{i+j+1},$$

when n is large. Therefore, the $k \times k$ matrix $\boldsymbol{X}'\boldsymbol{X}$ satisfies

$$\boldsymbol{X}'\boldsymbol{X} \simeq n\begin{pmatrix} \frac{1}{3} & \frac{1}{4} & \cdots & \frac{1}{k+2} \\ \frac{1}{4} & \frac{1}{5} & \cdots & \frac{1}{k+3} \\ \vdots & \vdots & & \vdots \\ \frac{1}{k+2} & \frac{1}{k+3} & \cdots & \frac{1}{2k+1} \end{pmatrix} := n\boldsymbol{M},$$

for large values of n. For example,

$$\boldsymbol{M} = \begin{pmatrix} 1/3 & 1/4 & 1/5 & 1/6 & 1/7 & 1/8 & 1/9 \\ 1/4 & 1/5 & 1/6 & 1/7 & 1/8 & 1/9 & 1/10 \\ 1/5 & 1/6 & 1/7 & 1/8 & 1/9 & 1/10 & 1/11 \\ 1/6 & 1/7 & 1/8 & 1/9 & 1/10 & 1/11 & 1/12 \\ 1/7 & 1/8 & 1/9 & 1/10 & 1/11 & 1/12 & 1/13 \\ 1/8 & 1/9 & 1/10 & 1/11 & 1/12 & 1/13 & 1/14 \\ 1/9 & 1/10 & 1/11 & 1/12 & 1/13 & 1/14 & 1/15 \end{pmatrix}.$$

in the $k = 7$ case. Already k is only modestly sized, but you see that almost half of the entries in $\boldsymbol{M}$ are below 0.01. Therefore, a small error in the computation $\boldsymbol{M}^{-1}$ leads to a massive error in the computation of $(\boldsymbol{X}'\boldsymbol{X})^{-1} \simeq (1/n)\boldsymbol{M}^{-1}$ for modest-to-large values of n. Your textbook recommends that you keep polynomial regression to below polynomials of degree 6.

2. Orthogonal polynomials

A more robust approach is to use *orthogonal polynomials*. That is, reparametrize our polynomial regression model such as

$$Y_i = \gamma_0\phi_0(x_i) + \gamma_1\phi_1(x_i) + \cdots + \gamma_k\phi_k(x_i) + \varepsilon_i,$$

where the γ_i's are unknown parameters, ϕ_j is a jth-order polynomial, and

$$\sum_{i=1}^{n}\phi_m(x_i)\phi_l(x_i) = \begin{cases} 1 & \text{whenever } m = l, \\ 0 & \text{whenever } m \neq l \end{cases} = I_{m,l}.$$

[In other words, $\phi_0, \ldots, \phi_k$ are *orthonormal over* $\{x_i\}_{i=1}^n$.]

Note that this is a linear model with $p := k + 1$. Indeed,

$$\boldsymbol{Y} = \boldsymbol{X\gamma} + \boldsymbol{\varepsilon},$$

where

$$\boldsymbol{X} = \begin{pmatrix} \phi_0(x_1) & \cdots & \phi_k(x_1) \\ \vdots & & \vdots \\ \phi_0(x_n) & \cdots & \phi_k(x_n) \end{pmatrix}.$$

Before we discuss how we can find these ϕ_i's, let us see what we have gained by using them. Notice that $X_{i,j} = \phi_j(x_i)$ $(1 \le i \le n,\ 0 \le j \le k)$ [note that the matrix is now indexed in a slightly funny way, but this is of course ok]. Therefore,

$$(\boldsymbol{X}'\boldsymbol{X})_{i,j} = \sum_k X_{k,i}X_{k,j} = \sum_k \phi_i(x_k)\phi_j(x_k) = I_{i,j} \Longrightarrow \boldsymbol{X}'\boldsymbol{X} = \boldsymbol{I}.$$

In particular, it is now a trivial task to invert $\boldsymbol{X}'\boldsymbol{X} = \boldsymbol{I}$! Our parameter estimates are also simplified: $\hat{\boldsymbol{\gamma}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y} = \boldsymbol{X}'\boldsymbol{Y}$. Therefore,

$$\hat{\gamma}_i = \sum_j X_{j,i}Y_j = \sum_{j=1}^n \phi_i(x_j)Y_j,$$

and

$$\hat{\boldsymbol{\gamma}} \sim \mathrm{N}_{k+1}\left(\boldsymbol{\gamma}\,, \sigma^2\boldsymbol{I}\right) \quad \text{under the normal-error model.}$$

In other words, under the normal model, the $\hat{\gamma}_i$'s are independent!

Let us compute the fitted values next: $\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\gamma}} = \boldsymbol{X}\boldsymbol{X}'\boldsymbol{Y}$; therefore,

$$\begin{aligned}\mathrm{RSS} &= \left\|\boldsymbol{Y} - \hat{\boldsymbol{Y}}\right\|^2 = \|\boldsymbol{Y}\|^2 - 2\boldsymbol{Y}'\hat{\boldsymbol{Y}} + \|\hat{\boldsymbol{Y}}\|^2 \\ &= \|\boldsymbol{Y}\|^2 - 2(\boldsymbol{X}'\boldsymbol{Y})^2 + (\boldsymbol{X}'\boldsymbol{Y})^2 = \|\boldsymbol{Y}\|^2 - (\boldsymbol{X}'\boldsymbol{Y})^2 \\ &= \|\boldsymbol{Y}\|^2 - \|\hat{\boldsymbol{\gamma}}\|^2.\end{aligned}$$

Now suppose we are testing for contrasts:

$$H_0:\ \boldsymbol{A\gamma} = \boldsymbol{c},$$

where $\boldsymbol{A}_{q\times(k+1)}$ is full rank with $q \le k$, and $\boldsymbol{c}_{q\times 1}$ is a known vector. Then,

$$\begin{aligned}\hat{\boldsymbol{\gamma}}_{H_0} &= \hat{\boldsymbol{\gamma}} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{A}'\left[\boldsymbol{A}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{A}'\right]^{-1}(\boldsymbol{c} - \boldsymbol{A}\hat{\boldsymbol{\gamma}}) \\ &= \hat{\boldsymbol{\gamma}} + \boldsymbol{A}'(\boldsymbol{A}\boldsymbol{A}')^{-1}(\boldsymbol{c} - \boldsymbol{A}\hat{\boldsymbol{\gamma}})\,,\end{aligned}$$

and

$$\begin{aligned}\mathrm{RSS}_{H_0} - \mathrm{RSS} &= (\boldsymbol{A}\hat{\boldsymbol{\gamma}} - \boldsymbol{c})'\left[\boldsymbol{A}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{A}'\right]^{-1}(\boldsymbol{A}\hat{\boldsymbol{\gamma}} - \boldsymbol{c}) \\ &= (\boldsymbol{A}\hat{\boldsymbol{\gamma}} - \boldsymbol{c})'(\boldsymbol{A}\boldsymbol{A}')^{-1}(\boldsymbol{A}\hat{\boldsymbol{\gamma}} - \boldsymbol{c})\,.\end{aligned}$$

In particular,

$$\frac{(\boldsymbol{A}\hat{\boldsymbol{\gamma}} - \boldsymbol{c})'(\boldsymbol{A}\boldsymbol{A}')^{-1}(\boldsymbol{A}\hat{\boldsymbol{\gamma}} - \boldsymbol{c})/q}{\left(\|\boldsymbol{Y}\|^2 - \|\hat{\boldsymbol{\gamma}}\|^2\right)/(n-k-1)} \overset{H_0}{\sim} F_{q,n-p}.$$

One can work out, using similar means, Scheffé-type conservative confidence intervals, etc.

3. The Forsythe–Hayes 2-step method

It remains to introduce a way of finding our orthonormal polynomials $\phi_0, \ldots, \phi_k$. Remember that our goal is to find these so that, given $x_1, \ldots, x_n$,

$$\sum_{i=1}^{n} \phi_m(x_i)\phi_\ell(x_i) = I_{m,\ell}.$$

Forsythe (1957) and Hayes (1969) adapted a classical method from analysis (Christoffel, 1858; Darboux, 1878) to the present setting in order to find an algorithm for finding such polynomials. We follow their method next.

Suppose we could find polynomials $\psi_0, \ldots, \psi_k$ such that

$$\sum_{i=1}^{n} \psi_m(x_i)\psi_\ell(x_i) = 0 \qquad \text{whenever } m \neq \ell.$$

That is, ψ's are merely orthogonal polynomials. Then,

$$\phi_m(x) := \frac{\psi_m(x)}{\sqrt{\sum_{i=1}^{n}[\psi_m(x_i)]^2}} \qquad (0 \le m \le k,\ x \in \mathbf{R})$$

defines our orthonormal polynomials. It remains to compute the ψ_j's.

We start with

$$\psi_0(x) := 1,$$

and ψ_1 has the form $\psi_1(x) = x - a_1$. [The slightly more general form $\psi_1(x) = c(a - x)$ does not gain us any further insight; why?] The constant a_1 has to be chosen so that $\sum_{i=1}^{n} \psi_1(x_i)\psi_0(x_i) = 0 \Longrightarrow \sum_{i=1}^{n} x_i - na_1 = 0 \Longrightarrow a_1 = \bar{x}$. That is,

$$\psi_1(x) = x - \bar{x}.$$

From here on we can describe things algorithmically. Suppose we have defined $\psi_0, \ldots, \psi_r$ so that they are orthogonal to each other [*the induction hypothesis*]. Then for all $r \ge 1$,

$$\psi_{r+1}(x) := (x - a_{r+1})\psi_r(x) - b_r\psi_{r-1}(x),$$

where a_{r+1} and b_r are chosen so that ψ_{r+1} is orthogonal to $\psi_0, \ldots, \psi_r$. In particular, we must have

$$\begin{aligned} 0 &= \sum_{i=1}^{n} \psi_{r+1}(x_i)\psi_r(x_i) \\ &= \sum_{i=1}^{n}(x_i - a_{r+1})[\psi_r(x_i)]^2 - b_r \overbrace{\sum_{i=1}^{n} \psi_{r-1}(x_i)\psi_r(x_i)}^{=0 \text{ by induction}} \\ &= \sum_{i=1}^{n} x_i[\psi_r(x_i)]^2 - a_{r+1}\sum_{i=1}^{n}[\psi_r(x_i)]^2 \qquad \text{(by induction).} \end{aligned}$$

Therefore,if we choose

$$a_{r+1} = \frac{\sum_{i=1}^{n} x_i [\psi_r(x_i)]^2}{\sum_{i=1}^{n} [\psi_r(x_i)]^2},$$

then ψ_{r+1} is orthogonal to ψ_r.

In order to find b_r we note that we must set things up—at the very least—so that ψ_{r+1} is orthogonal to ψ_{r-1}. That is,

$$\begin{aligned}
0 &= \sum_{i=1}^{n} \psi_{r+1}(x_i)\psi_{r-1}(x_i)\\
&= \sum_{i=1}^{n} (x_i - a_{r+1})\psi_r(x_i)\psi_{r-1}(x_i) - b_r \sum_{i=1}^{n} [\psi_{r-1}(x_i)]^2\\
&= \sum_{i=1}^{n} x_i\psi_r(x_i)\psi_{r-1}(x_i) - a_{r+1}\sum_{i=1}^{n} \psi_r(x_i)\psi_{r-1}(x_i) - b_r \sum_{i=1}^{n} [\psi_{r-1}(x_i)]^2\\
&= \sum_{i=1}^{n} x_i\psi_r(x_i)\psi_{r-1}(x_i) - b_r \sum_{i=1}^{n} [\psi_{r-1}(x_i)]^2;
\end{aligned}$$

thanks to induction. We can simplify this a little more still, mainly because $\psi_r(x) = (x - a_r)\psi_{r-1}(x) - b_r\psi_{r-2}(x)$. In this way we find that

$$\begin{aligned}
\sum_{i=1}^{n} x_i\psi_r(x_i)\psi_{r-1}(x_i) &= \sum_{i=1}^{n} (x_i - a_r)\psi_r(x_i)\psi_{r-1}(x_i) \qquad \text{(induction)}\\
&= \sum_{i=1}^{n} [\psi_r(x_i)]^2 + b_r \sum_{i=1}^{n} \psi_{r-2}(x_i)\psi_r(x_i) = \sum_{i=1}^{n} [\psi_r(x_i)]^2.
\end{aligned}$$

Therefore, our candidate for b_r is

$$b_r = \frac{\sum_{i=1}^{n} [\psi_r(x_i)]^2}{\sum_{i=1}^{n} [\psi_{r-1}(x_i)]^2}.$$

With the present choice of a_{r+1} and b_r, ψ_{r+1} is orthogonal to both ψ_r and ψ_{r-1}. Now it remains to prove that this choice of a_{r+1} and b_r actually works. That is, we proceed to show that, for our construction, ψ_{r+1} is orthogonal to ψ_ℓ for every $\ell \le r-2$. Note that

$$\begin{aligned}
\sum_{i=1}^{n} \psi_{r+1}(x_i)\psi_\ell(x_i) &= \sum_{i=1}^{n} (x_i - a_{r+1})\psi_r(x_i)\psi_\ell(x_i) - b_r \sum_{i=1}^{n} \psi_{r-1}(x_i)\psi_\ell(x_i)\\
&= \sum_{i=1}^{n} x_i\psi_r(x_i)\psi_\ell(x_i)\\
&= \sum_{i=1}^{n} (x_i - a_{\ell+1})\psi_r(x_i)\psi_\ell(x_i) - b_\ell \overbrace{\sum_{i=1}^{n} \psi_r(x_i)\psi_{\ell+1}(x_i)}^{=0 \text{ by induction}}.
\end{aligned}$$

The defining property of the ψ's tells us that for $\ell \le r-2$,

$$\sum_{i=1}^{n} \psi_{r+1}(x_i)\psi_{\ell}(x_i) = \sum_{i=1}^{n} \psi_r(x_i)\psi_{\ell+1}(x_i) = 0,$$

thanks to induction.

Splines

1. Introduction

One way to think of linear regression is this: We have data of pairs of points $(x_1\,,y_1),\ldots,(x_n\,,y_n)$, and try to fit a line through this. We have added to this procedure by studying what happens if the error in this approximation $[y=\beta_0+\beta_1x+\text{noise}]$ has a normal distribution etc. Here we study a different angle on these problems: Are there better curves than lines that can be fitted through the points $(x_1\,,y_1),\ldots,(x_n\,,y_n)$? One answer is polynomial regression. Another is splines, which I will say a few things about.

Suppose we have a fixed interval $[a\,,b]$ in which we are seeing numbers, and also have a subdivision of this interval,

$$a:=\xi_0\le\xi_1<\cdots<\xi_k\le\xi_{k+1}:=b.$$

We say that a function $S:[a\,,b]\to\mathbf{R}$ is a *spline* with *knots* $\xi_1<\cdots<\xi_k$ if there exists an integer $M\ge1$—called the *order of S*—such that: (i) S is a polynomial of degree $M-1$ in $[\xi_i\,,\xi_{i+1}]$ for every $1\le i\le k-1$; and (ii) S and its first $M-2$ derivatives all exist and are continuous everywhere in $[a\,,b]$. When $a=-\infty$ and $b=\infty$, we think of ξ_0 and ξ_{k+1} also as knots of S.

In other words, S is a spline if it is smooth function that is piecewise polynomial.

2. Linear splines

Suppose our data is $(\xi_1\,,y_1),\ldots,(\xi_k\,,y_k)$ and we wish to do a piecewise linear fit. The following is an order $M=2$ spline that does the job:

$$S(x)=y_i+(y_{i+1}-y_i)\left(\frac{x-\xi_i}{\xi_{i+1}-\xi_i}\right)\qquad\text{for every }\xi_i\le x<\xi_{i+1}.$$

This does ths job. (Draw a picture!) But we need to understand how it was derived.

First you reparametrize the interval $[\xi_i\,,\xi_{i+1})$ by $[0\,,1)$. That is,

$$t := \frac{x-\xi_i}{\xi_{i+1}-\xi_i} \qquad \text{for } \xi_i \le x < \xi_{i+1}. \tag{1}$$

As the variable x increases from ξ_i to ξ_{i+1}, the variable t increases from 0 to 1. That is, with this parametrization,

$$S(x) = Y_i(t) \qquad \text{when } \xi_i \le x < \xi_{i+1}. \tag{2}$$

In parameter t, the ith piece of the spline should look like a line:

$$Y_i(t) = a_i + b_i t \qquad (0 \le t < 1),$$

subject to $Y_i(0) = y_i$ and $Y_i(1) = y_{i+1}$. Therefore, $a_i = y_i$ and $a_i + b_i = y_{i+1}$. This leads us to

$$Y_i(t) = y_i + (y_{i+1} - y_i)t \qquad (0 \le t < 1).$$

Use this together with (1) and (2) in order to compute the spline at any point x.

3. Quadratic splines

Here, $M = 3$. We apply the same reparametrization (1) and (2) as before, but now use the following quadratic interpolation formula in place of the previous linear one:

$$Y_i(t) = a_i + b_i t + c_i t^2 \qquad (0 \le t < 1);$$

subject to $Y_i(0) = y_i$, $Y_i(1) := \lim_{t\uparrow 1} Y_i(t) = y_{i+1}$. This yields,

$$a_i = y_i, \quad a_i + b_i + c_i = y_{i+1} \Longrightarrow \begin{cases} y_i = a_i, \\ y_{i+1} - y_i = b_i + c_i. \end{cases}$$

This identifies the a_i's and shows also that if we can compute all of the b_i's then we will have explicit formulas for the c_i's as well.

We also have a continuity condition on the derivative of order $M - 2 = 1$. To see how that manifests itself note that

$$Y_i'(t) = b_i + 2c_i t \qquad (0 \le t < 1).$$

In particular, $Y_i'(0) = b_i$ and $Y_i'(1) = b_i + 2c_i$. If $\xi_i \le x < \xi_{i+1}$, then

$$S'(x) = \frac{\mathrm{d}S(x)}{\mathrm{d}x} = \frac{\mathrm{d}Y_i(t)}{\mathrm{d}t}\cdot\frac{\mathrm{d}t}{\mathrm{d}x} = \frac{Y_i'(t)}{\xi_{i+1}-\xi_i} = \frac{b_i + 2c_i t}{\xi_{i+1}-\xi_i},$$

thanks to the chain rule. Since we want S' to be continuous everywhere, it is enough to check that $S'(\xi_{i+1}-) = S'(\xi_{i+1}+)$. But

$$S'(\xi_{i+1}-) = \frac{Y_i'(1)}{\xi_{i+1}-\xi_i} = \frac{b_i + 2c_i}{\xi_{i+1}-\xi_i},$$

and

$$S'(\xi_{i+1}+) = \frac{Y'_{i+1}(0)}{\xi_{i+2}-\xi_{i+1}} = \frac{b_{i+1}}{\xi_{i+2}-\xi_{i+1}}.$$

Therefore, we also want

$$\frac{b_i + 2c_i}{\xi_{i+1}-\xi_i} = \frac{b_{i+1}}{\xi_{i+2}-\xi_{i+1}}.$$

Equivalently,

$$c_i = \frac{1}{2}\left[\left(\frac{\xi_{i+1}-\xi_i}{\xi_{i+2}-\xi_{i+1}}\right) b_{i+1} - b_i\right].$$

Since $y_{i+1} - y_i = b_i + c_i$, the preceeding shows that

$$\begin{aligned} y_{i+1} - y_i &= b_i + \frac{1}{2}\left[\left(\frac{\xi_{i+1}-\xi_i}{\xi_{i+2}-\xi_{i+1}}\right) b_{i+1} - b_i\right] \\ &= \frac{b_{i+1}}{2}\left(\frac{\xi_{i+1}-\xi_i}{\xi_{i+2}-\xi_{i+1}}\right) + \frac{b_i}{2}. \end{aligned}$$

Equivalently, that

$$b_{i+1} = [2(y_{i+1} - y_i) - b_i]\left(\frac{\xi_{i+2}-\xi_{i+1}}{\xi_{i+1}-\xi_i}\right).$$

That is, if we knew b_0 then we could iteratively compute all of the b_i's, and be finished.

In order to find b_0 we need one more condition. Usually, people assume a zero derivative condition at the left-most knot:

$$S'(a) = 0 \Longrightarrow Y'_0(0) = 0 \Longrightarrow b_0 = 0.$$

This completes the algorithm for computing our spline. Any other choice of $S'(a)$, than zero, would yield a different [perfectly ok] spline as well.

In practice, it is not enough to use quadratic splines, since splines start to look smooth at the cubic level. This requires a great deal more computation, but is still a managable problem.

4. Cubic splines

Now we consider the still-more interesting case that $M = 4$ [cubic splines]. Recall that we want a piece-wise cubic fit S of our data such that S, S', and S'' are continuous [not piecewise continuous] functions. We will adopt the boundary conditions

$$S''(a) = S''(b) = 0,$$

as it is customary. As before, we work on each piece $[\xi_i\,, \xi_{i+1}]$ separately by first changing variables for the ith piece:

$$t := \frac{x-\xi_i}{\xi_{i+1}-\xi_i} \qquad (\xi_i \le x < \xi_{i+1}).$$

Then in t coordinates, our spline model is

$$Y_i(t) = a_i + b_i t + c_i t^2 + d_i t^3 \qquad (0 \le t < 1),$$

in the ith piece. First of all, our spline function has to go through the data points. That is, $Y_i(0) = y_i$ for all $0 \le i \le k$. In other words,

$$a_i = y_i \qquad (0 \le i \le k). \tag{3}$$

This shows that the a_i's are determined. We need to compute the b's, c's, and d's.

The statement that S is continuous is equivalent to $Y_{i+1}(0) = Y_i(1)$ for all i. That is,

$$y_{i+1} = y_i + b_i + c_i + d_i \qquad (0 \le i \le k-1). \tag{4}$$

We want also that S' to be continuous. Now, if $\xi_i \le x < \xi_{i+1}$, then

$$S'(x) = \frac{\mathrm{d}S(x)}{\mathrm{d}x} = \frac{\mathrm{d}Y_i(t)}{\mathrm{d}t} \cdot \frac{\mathrm{d}t}{\mathrm{d}x} = \frac{Y_i'(t)}{\xi_{i+1} - \xi_i} = \frac{b_i + 2c_i t + 3d_i t^2}{\xi_{i+1} - \xi_i}.$$

The continuity of S' is equivalent to the statement that $S'(\xi_{i+1}-) = S'(\xi_{i+1}+)$ for all i. Put in yet another way, we want

$$S'(\xi_{i+1}-) = \frac{Y_i'(1)}{\xi_{i+1} - \xi_i} = S'(\xi_{i+1}+) = \frac{Y_{i+1}'(0)}{\xi_{i+2} - \xi_{i+1}}.$$

Plug in the value of $Y_i'(1)$ and $Y_{i+1}'(0)$ to see that

$$\frac{b_i + 2c_i + 3d_i}{\xi_{i+1} - \xi_i} = \frac{b_{i+1}}{\xi_{i+2} - \xi_{i+1}}. \tag{5}$$

This shows that if we could solve the c_i's and the d_i's, as well as b_0, then we could also find all of the b_i's. In fact, the way we do so is as follows: Set

$$B_i := \frac{b_i}{\xi_{i+1} - \xi_i}, \; C_i := \frac{c_i}{\xi_{i+1} - \xi_i}, \; D_i := \frac{d_i}{\xi_{i+1} - \xi_i}.$$

Then (5) says that

$$B_i + 2C_i + 3D_i = B_{i+1} \quad \Longleftrightarrow \quad B_{i+1} - B_i = 2C_i + 3D_i.$$

Add from $i = 0$ to $i = j$ in order to see that for all j,

$$B_j = B_0 + 2\sum_{i=0}^{j} C_i + 3\sum_{i=0}^{j} D_i. \tag{6}$$

Our second-derivative requirement is that S'' is continuous. If $\xi_i \le x < \xi_{i+1}$, then the chain rule tells us that

$$S''(x) = Y_i''(t) \cdot \frac{\mathrm{d}t}{\mathrm{d}x} = \frac{2c_i + 6d_i t}{\xi_{i+1} - \xi_i},$$

since $\mathrm{d}^2t/\mathrm{d}x^2 = 0$. Thus, we set $S''(\xi_{i+1}-) = S''(\xi_{i+1}+)$ in order to see that

$$\frac{Y_i''(1)}{\xi_{i+1} - \xi_i} = \frac{Y_{i+1}''(0)}{\xi_{i+2} - \xi_{i+1}},$$

which is to say that

$$\frac{2c_i + 6d_i}{\xi_{i+1} - \xi_i} = \frac{2c_{i+1}}{\xi_{i+2} - \xi_{i+1}}.$$

I.e.,

$$D_i = \frac{d_i}{\xi_{i+1} - \xi_i} = \frac{1}{3}\left[\frac{c_{i+1}}{\xi_{i+2} - \xi_{i+1}} - \frac{c_i}{\xi_{i+1} - \xi_i}\right] = \frac{C_{i+1} - C_i}{3}. \tag{7}$$

Thus, if we could compute all of the c_i's, then we will know all of the d_i's as well, and thereby all b_i's, which ends our task. In fact, by (7),

$$\sum_{\ell=0}^{i} D_\ell = \frac{1}{3}\sum_{\ell=0}^{i}(C_{\ell+1} - C_\ell) = \frac{C_{i+1} - C_0}{3}.$$

Plug this into (6) and change variables $[j \leftrightarrow i]$ to obtain

$$\begin{aligned} B_i &= B_0 + 2\sum_{\ell=0}^{i} C_\ell + 3[C_{i+1} - C_0] \\ &= B_0 - C_0 + 2\sum_{\ell=1}^{i} C_\ell + 3C_{i+1}. \end{aligned}$$

Now, by (4),

$$\begin{aligned} E_i := \frac{y_{i+1} - y_i}{\xi_{i+1} - \xi_i} &= B_i + C_i + D_i \\ &= B_0 - C_0 + 2\sum_{\ell=1}^{i} C_\ell + 3C_{i+1} + C_i + \frac{C_{i+1} - C_i}{3} \\ &= B_0 - C_0 + 2\sum_{\ell=1}^{i-1} C_\ell + \frac{8}{3}C_i + \frac{10}{3}C_{i+1}, \end{aligned}$$

where $\sum_{\ell=1}^{0} C_\ell := 0$. In other words, given B_0 and C_0 [oftentimes, people choose them to be zero], the vector $\boldsymbol{C} := (C_1, \ldots, C_{k+1})'$ solves

$$\boldsymbol{E} = (B_0 - C_0)\mathbf{1}_{(k+1)\times 1} + \boldsymbol{MC},$$

where $\boldsymbol{M}$ is the following $(k+1) \times (k+1)$ matrix:

$$\boldsymbol{M} := \begin{pmatrix} {}^{8}\!/_{3} & {}^{10}\!/_{3} & 0 & 0 & \cdots & 0 & 0 & 0 & 0 \\ 2 & {}^{8}\!/_{3} & {}^{10}\!/_{3} & 0 & \cdots & 0 & 0 & 0 & 0 \\ 2 & 2 & {}^{8}\!/_{3} & {}^{10}\!/_{3} & \cdots & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \ddots & \ddots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \ddots & \ddots & \ddots & \vdots & \vdots \\ 2 & 2 & 2 & 2 & \cdots & 2 & {}^{8}\!/_{3} & {}^{10}\!/_{3} & 0 \\ 2 & 2 & 2 & 2 & \cdots & 2 & 2 & {}^{8}\!/_{3} & {}^{10}\!/_{3} \\ 2 & 2 & 2 & 2 & \cdots & 2 & 2 & 2 & {}^{8}\!/_{3} \end{pmatrix}.$$

The matrix $\boldsymbol{M}$ turns out to be invertible. Therefore, we obtain $\boldsymbol{C}$ by

$$\boldsymbol{C} = \boldsymbol{M}^{-1} \left[\boldsymbol{E} - (B_0 - C_0)\mathbf{1}_{(k+1)\times 1} \right].$$

Now, (7) yields $\boldsymbol{D} := (D_1, \ldots, D_{k+1})'$, and (6) allows us to finally compute $\boldsymbol{B} := (B_1, \ldots, B_{k+1})'$. Now backsolve for the b_i's, c_i's, and the d_i's to finish [$b_i = (\xi_{i+1} - \xi_i)B_i$, etc.]