GENIOMHE

Multivariate Statistics

by Samuel Ortion

Prof.: Cyril Dalmasso

Fall 2023

Contents

1 Introduction

Definition 1: Long Term Nonprocessor (LTNP)

Patient who will remain a long time in good health condition, even with a large viral load (cf. HIV).

Example 1: Genotype: Qualitative or Quantitative?

$$\text{SNP} : \begin{cases} \text{AA} \\ \text{AB} \\ \text{BB} \end{cases} \rightarrow \begin{pmatrix} 0 \\ 1 \\ 2 \end{pmatrix},$$

thus we might consider genotype either as a qualitative variable or quantitative variable.

When the variable are quantitative, we use regression, whereas for qualitative variables, we use an analysis of variance.

2 Linear Model

2.1 Simple Linear Regression

$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$

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

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \varepsilon_n \end{pmatrix}$$

Assumptions

(A_1) ε_i are independent;

(A_2) ε_i are identically distributed;

(A_3) ε_i are i.i.d $\sim \mathcal{N}(0, \sigma^2)$ (homoscedasticity).

2.2 Generalized Linear Model

$g(\mathbb{E}(Y)) = X\beta$

with g being

- Logistic regression: $g(v) = \log\left(\frac{v}{1-v}\right)$, for instance for boolean values,
- Poisson regression: $g(v) = \log(v)$, for instance for discrete variables.

2.2.1 Penalized Regression

When the number of variables is large, e.g, when the number of explanatory variable is above the number of observations, if $p >> n$ (p: the number of explanatory variable, n is the number of observations), we cannot estimate the parameters. In order to estimate the parameters, we can use penalties (additional terms).

Lasso regression, Elastic Net, etc.

2.2.2 Statistical Analysis Workflow

Step 1. Graphical representation;

Step 2. ...

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

is noted equivalently as

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} \\ 1 & x_{21} & x_{22} \\ 1 & x_{31} & x_{32} \\ 1 & x_{41} & x_{42} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \end{pmatrix}.$$

2.3 Parameter Estimation

2.3.1 Simple Linear Regression

2.3.2 General Case

If $\mathbf{X}^T\mathbf{X}$ is invertible, the OLS estimator is:

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} \tag{2.1}$$

2.3.3 Ordinary Least Square Algorithm

We want to minimize the distance between $\mathbf{X}\beta$ and $\mathbf{Y}$:

$$\min \|\mathbf{Y} - \mathbf{X}\beta\|^2$$

(See chapter 3).

$$\begin{aligned}
&\Rightarrow \mathbf{X}\beta = proj^{(1,\mathbf{X})}\mathbf{Y} \\
&\Rightarrow \forall v \in w,\, vy = vproj^{w}(y) \\
&\Rightarrow \forall i : \\
&\quad \mathbf{X}_i\mathbf{Y} = \mathbf{X}_i X\hat{\beta} \qquad \text{where } \hat{\beta} \text{ is the estimator of } \beta \\
&\Rightarrow \mathbf{X}^T\mathbf{Y} = \mathbf{X}^T\mathbf{X}\hat{\beta} \\
&\Rightarrow (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} = (\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{X})\hat{\beta} \\
&\Rightarrow \hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}
\end{aligned}$$

This formula comes from the orthogonal projection of $\mathbf{Y}$ on the vector subspace defined by the explanatory variables $\mathbf{X}$

$\mathbf{X}\hat{\beta}$ is the closest point to $\mathbf{Y}$ in the subspace generated by $\mathbf{X}$.

If H is the projection matrix of the subspace generated by $\mathbf{X}$, $X\mathbf{Y}$ is the projection on $\mathbf{Y}$ on this subspace, that corresponds to $\mathbf{X}\hat{\beta}$.

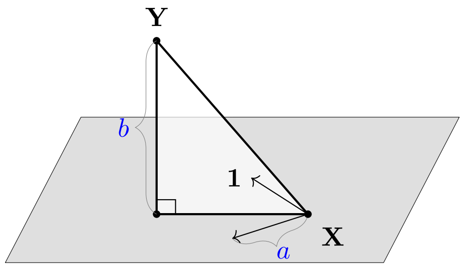

Figure 2.1 Orthogonal projection of $\mathbf{Y}$ on plan generated by the base described by $\mathbf{X}$. a corresponds to $\|\mathbf{X}\hat{\beta} - \bar{\mathbf{Y}}\|^2$ and b corresponds to $\|\mathbf{Y} - \hat{\beta}\mathbf{X}\|^2$

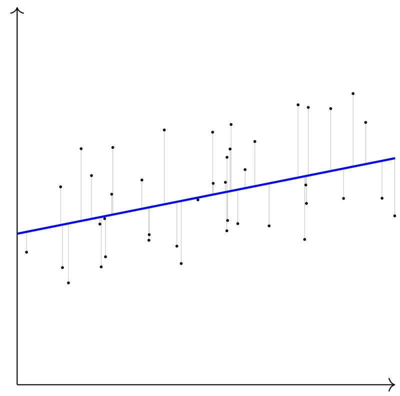

Figure 2.2 Ordinary least squares and regression line with simulated data.

2.4 Coefficient of Determination: R^2

Definition 2: R^2

$$0 \leq R^2 = \frac{\|\mathbf{X}\hat{\beta} - \bar{\mathbf{Y}}\,\mathbf{1}\|^2}{\|\mathbf{Y} - \bar{\mathbf{Y}}\,\mathbf{1}\|^2} = 1 - \frac{\|\mathbf{Y} - \mathbf{X}\hat{\beta}\|^2}{\|\mathbf{Y} - \bar{\mathbf{Y}}\,\mathbf{1}\|^2} \leq 1$$

proportion of variation of $\mathbf{Y}$ explained by the model.

3 Elements of Linear Algebra

Remark 1: vector

Let u a vector, we will use interchangeably the following notations: u and $\vec{u}$

Let $u = \begin{pmatrix} u_1 \\ \vdots \\ u_n \end{pmatrix}$ and $v = \begin{pmatrix} v_1 \\ \vdots \\ v_n \end{pmatrix}$

Definition 3: Scalar Product (Dot Product)

$$\langle u, v \rangle = \begin{pmatrix} u_1, \ldots, u_v \end{pmatrix} \begin{pmatrix} v_1 \\ \vdots \\ v_n \end{pmatrix}$$
$$= u_1 v_1 + u_2 v_2 + \ldots + u_n v_n$$

We may use $\langle u, v \rangle$ or $u \cdot v$ notations.

Dot product properties

Commutative $\langle u, v \rangle = \langle v, u \rangle$

Distributive $\langle (u + v), w \rangle = \langle u, w \rangle + \langle v, w \rangle$

$\langle u, v \rangle = \|u\| \times \|v\| \times \cos(\widehat{u, v})$

$\langle a, a \rangle = \|a\|^2$

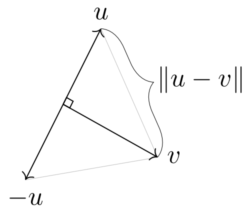

Figure 3.1 Scalar product of two orthogonal vectors.

Definition 4: Norm

Length of the vector.

$$\|u\| = \sqrt{\langle u, v \rangle}$$

$\|u, v\| > 0$

Definition 5: Distance

$$dist(u, v) = \|u - v\|$$

Definition 6: Orthogonality

Remark 2

$$(dist(u, v))^2 = \|u - v\|^2,$$

and

$$\langle v - u, v - u \rangle$$

$$\begin{aligned}\langle v - u, v - u \rangle &= \langle v, v \rangle + \langle u, u \rangle - 2\langle u, v \rangle \\ &= \|v\|^2 + \|u\|^2 \\ &= -2\langle u, v \rangle\end{aligned}$$

$$\begin{aligned}\|u - v\|^2 &= \|u\|^2 + \|v\|^2 - 2\langle u, v \rangle \\ \|u + v\|^2 &= \|u\|^2 + \|v\|^2 + 2\langle u, v \rangle\end{aligned}$$

Proposition 1: Scalar product of orthogonal vectors

$$u \perp v \Leftrightarrow \langle u, v \rangle = 0$$

Indeed. $\|u-v\|^2 = \|u+v\|^2$, as illustrated in Figure 3.1.

$$\begin{aligned}&\Leftrightarrow -2\langle u,v\rangle = 2\langle u,v\rangle\\&\Leftrightarrow 4\langle u,v\rangle = 0\\&\Leftrightarrow \langle u,v\rangle = 0\end{aligned}$$

□

Theorem 1: Pythagorean theorem

If $u \perp v$, then $\|u+v\|^2 = \|u\|^2 + \|v\|^2$.

Definition 7: Orthogonal Projection

Let $y = \begin{pmatrix} y_1 \\ . \\ y_n \end{pmatrix} \in \mathbb{R}^n$ and w a subspace of $\mathbb{R}^n$. $\mathcal{Y}$ can be written as the orthogonal projection of y on w:

$$\mathcal{Y} = proj^w(y) + z,$$

where

$$\begin{cases} z \in w^{\perp} \\ proj^w(y) \in w \end{cases}$$

There is only one vector $\mathcal{Y}$ that ?
The scalar product between z and (?) is zero.

Property 1. *$proj^w(y)$ is the closest vector to y that belongs to w.*

Definition 8: Matrix

A matrix is an application, that is, a function that transform a thing into another, it is a linear function.

Example 2: Matrix application

Let A be a matrix:

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

and

$$x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$$

Then,

$$\begin{aligned} Ax &= \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \\ &= \begin{pmatrix} ax_1 + bx_2 \\ cx_1 + dx_2 \end{pmatrix} \end{aligned}$$

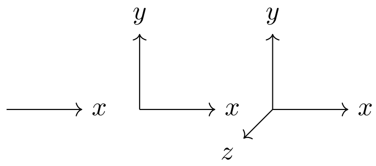

Figure 3.2 Coordinate systems

Example 2 continued

Similarly,

$$\begin{pmatrix} a & b & c & d \\ e & f & g & h \\ i & j & k & l \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} ax_1 + bx_2 + cx_3 + dx_4 \\ ex_1 + fx_2 + gx_3 + hx_4 \\ ix_1 + jx_2 + kx_3 + lx_4 \end{pmatrix}$$

The number of columns has to be the same as the dimension of the vector to which the matrix is applied.

Definition 9: Tranpose of a Matrix

Let $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$, then $A^T = \begin{pmatrix} a & c \\ b & d \end{pmatrix}$