GENIOMHE

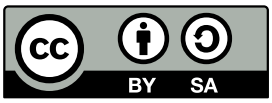
Multivariate Statistics

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.

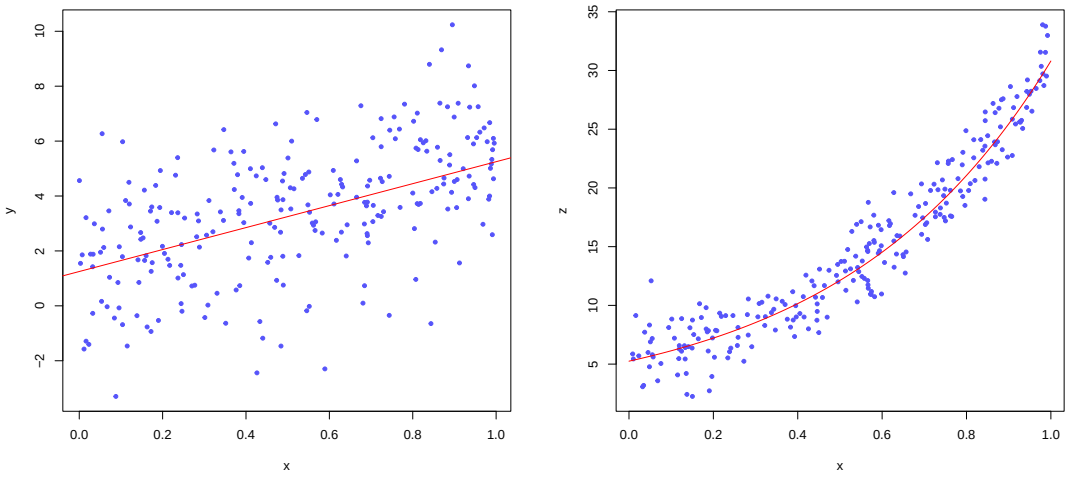


Figure 1.1. Illustration of two models fitting observed values

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.

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

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).

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

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 $\mathbf{X}$, $\mathbf{XY}$ is the projection on $\mathbf{Y}$ on this subspace, that corresponds to $\mathbf{X}\hat{\beta}$.

2.4. Sum of squares

$\mathbf{Y} - \mathbf{X}\hat{\beta} \perp \mathbf{X}\hat{\beta} - \mathbf{Y}\,\mathbf{1}$ if $\mathbf{1} \in V$, so

$$\underbrace{\|\mathbf{Y} - \bar{\mathbf{Y}}\,\mathbf{1}\|}_{\text{Total SS}} = \underbrace{\|\mathbf{Y} - \mathbf{X}\hat{\beta}\|^2}_{\text{Residual SS}} + \underbrace{\|\mathbf{X}\hat{\beta} - \bar{\mathbf{Y}}\,\mathbf{1}\|^2}_{\text{Explicated SS}}$$

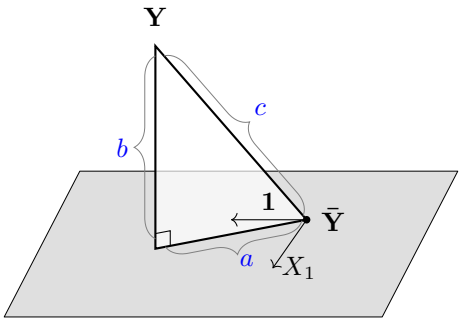


Figure 2.1. Orthogonal projection of $\mathbf{Y}$ on plan generated by the base described by $\mathbf{X}$. a corresponds to $\|\mathbf{X}\hat{\beta} - \bar{\mathbf{Y}}\|^2$ and b corresponds to $\hat{\varepsilon} = \|\mathbf{Y} - \hat{\beta}\mathbf{X}\|^2$ and c corresponds to $\|Y - \bar{Y}\|^2$.

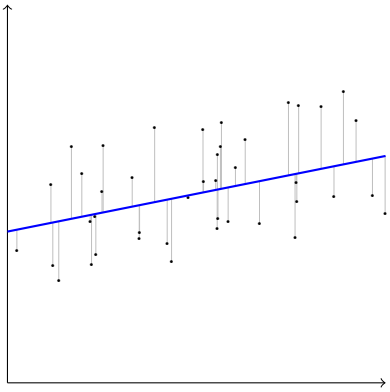


Figure 2.2. Ordinary least squares and regression line with simulated data.

2.5. 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.

Definition 3: Model dimension

Let $\mathcal{M}$ be a model. The dimension of $\mathcal{M}$ is the dimension of the subspace generated by $\mathbf{X}$, that is the number of parameters in the β vector.
Nb. The dimension of the model is not the number of parameter, as σ^2 is one of the model parameters.

2.6. Gaussian vectors

Definition 4: Normal distribution

$X \sim \mathcal{N}(\mu, \sigma^2)$, with density function f

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$$

Definition 5: Gaussian vector

A random vector $\mathbf{Y} \in \mathbb{R}^n$ is a gaussian vector if every linear combination of its component is a gaussian random variable.

Property 1. *$m = \mathbb{E}(Y) = (m_1, \dots, m_n)^T$, where $m_i = \mathbb{E}(Y_i)$*

$$\mathbf{Y} \sim \mathcal{N}_n(m, \Sigma)$$

where Σ is the variance-covariance matrix!

$$\Sigma = \mathrm{E}\left[(\mathbf{Y} - m)(\mathbf{Y} - m)^T\right].$$

Remark 1

$$\mathrm{Cov}(Y_i, Y_i) = \mathrm{Var}(Y_i)$$

Definition 6: Covariance

$$\mathrm{Cov}(Y_i, Y_j) = \mathbb{E}\left((Y_i - \mathbb{E}(Y_j))(Y_j - \mathbb{E}(Y_j))\right)$$

When two variable are linked, the covariance is large.
If two variables X, Y are independent, $\mathrm{Cov}(X, Y) = 0$.

Definition 7: Correlation coefficient

$$\mathrm{Cor}(Y_i, Y_j) = \frac{\mathbb{E}\left((Y_i - \mathbb{E}(Y_j))(Y_j - \mathbb{E}(Y_j))\right)}{\sqrt{\mathbb{E}(Y_i - \mathbb{E}(Y_i)) \cdot \mathbb{E}(Y_j - \mathbb{E}(Y_j))}}$$

Covariance is really sensitive to scale of variables. For instance, if we measure distance in millimeters, the covariance would be larger than in the case of a measure expressed in metters. Thus the correlation coefficient, which is a sort of normalized covariance is useful, to be able to compare the values.

Remark 2

$$\begin{aligned}\mathrm{Cov}(Y_i, Y_i) &= \mathbb{E}((Y_i - \mathbb{E}(Y_i))(Y_i - \mathbb{E}(Y_i))) \\ &= \mathbb{E}((Y_i - \mathbb{E}(Y_i))^2) \\ &= \mathrm{Var}(Y_i)\end{aligned}$$

$$\Sigma = \begin{pmatrix} \mathbb{V}(Y_1) & \cdots & \cdots \\ \cdots & \mathrm{Cov}(Y_i, Y_j) & \mathbb{V}(Y_i) \cdots \\ \cdots & \cdots & \mathbb{V}(Y_n) \end{pmatrix} \tag{2.2}$$

Definition 8: Identity matrix

$$\mathcal{I}_n = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

Theorem 1: Cochran Theorem (Consequence)

Let $\mathbf{Z}$ be a gaussian vector: $\mathbf{Z} \sim \mathcal{N}_n(0_n, I_n)$.

- If V_1, V_n are orthogonal subspaces of $\mathbb{R}^n$ with dimensions n_1, n_2 such that

  $$\mathbb{R}^n = V_1 \overset{\perp}{\oplus} V_2.$$

- If Z_1, Z_2 are orthogonal of $\mathbf{Z}$ on V_1 and V_2 i.e. $Z_1 = \Pi_{V_1}(\mathbf{Z}) = \Pi_1 \mathbf{Y}$ and $Z_2 = \Pi_{V_2}(\mathbf{Z}) = \Pi_2 \mathbf{Y}$ (Π_1 and Π_2 being projection matrices) then:
- z_1, Z_2 are independent gaussian vectors, $Z_1 \sim \mathcal{N}_{n_1}(0_n, \Pi_1)$ and $Z_2 \sim \mathcal{N}(0_{n_2}, \Pi_2)$.
  In particular $\|Z_1\| \sim \chi^2(n_1)$ and $\|Z_2\| \sim \chi^2(n_2)$.

$Z_2 = \Pi_{V_1}(\mathbf{Z})$ is the projection of $\mathbf{Z}$ on subspace V_1.
...

Property 2 (Estimators properties in the linear model)**.** *According to Theorem 2.6,*

$\hat{m}$ *is independent from* $\hat{\sigma}^2$

$\|\mathbf{Y} - \Pi_V(\mathbf{Y})\|^2 = \|\varepsilon - \Pi_V(\varepsilon)\|^2 = \|\Pi_V^{\perp}(\varepsilon)\|^2$

$\hat{m} = \mathbf{X}\hat{\beta}$
$\hat{m}$ *is the estimation of the mean.*

Definition 9: Chi 2 distribution

If $X_1, \ldots, X_n$ i.i.d. $\sim \mathcal{N}(0, 1)$, then;,

$$X_1^2 + \ldots X_n^2 \sim \chi_n^2$$

2.6.1. Estimator's properties

$$\Pi_V = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$$

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

so

$$= \Pi_V\mathbf{Y}$$

According to Cochran theorem, we can deduce that the estimator of the predicted value $\hat{m}$ is independent $\hat{\sigma}^2$

All the sum of squares follows a χ^2 distribution.

2.6.2. Estimators properties

- $\hat{m}$ is unbiased and estimator of m;
- $\mathbb{E}(\hat{\sigma}^2) = \sigma^2(n-q)/n$ $\hat{\sigma}^2$ is a biased estimator of σ^2.

  $$S^2 = \frac{1}{n-q}\|\mathbf{Y} - \Pi_V\|^2$$

  is an unbiased estimator of σ^2.

We can derive statistical test from these properties.

2.7. Statistical tests

2.7.1. Student t-test

$$\frac{\hat{\theta} - \theta}{\sqrt{\frac{\hat{\mathbb{V}}(\hat{\theta})}{n}}} \underset{H_0}{\sim} t_{n-q}$$

where

Estimation of σ^2 A biased estimator of σ^2 is:

$$\hat{\sigma^2} = ?$$

S^2 is the unbiased estimator of σ^2

$$\begin{aligned} S^2 &= \frac{1}{n-q}\|\mathbf{Y} - \Pi_V(\mathbf{Y})\|^2 \\ &= \frac{1}{n-q}\sum_{i=1}^{n}(Y_i - (\mathbf{X}\hat{\beta})_i)^2 \end{aligned}$$

Remark 3: On $\widehat{m}$

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon \Leftrightarrow \qquad \mathbb{E}(\mathbf{Y}) = \mathbf{X}\beta$$

2.8. Student test of nullity of a parameter

Let β_j be a parameter, the tested hypotheses are as follows:

$$\begin{cases} (H_0) : \beta_j = 0 \\ (H_1) : \beta_j \neq 0 \end{cases}$$

Under the null hypothesis:

$$\frac{\hat{\beta}_j - \beta_j}{S\sqrt{(\mathbf{X}^T\mathbf{X})^1_{j,j}}} \sim \mathcal{S}\mathrm{t}(n-q).$$

The test statistic is:

$$W_n = \frac{\hat{\beta}_j}{S\sqrt{(\mathbf{X}^T\mathbf{X})^{-1}_{j,j}}} \underset{H_0}{\sim} \mathcal{S}\mathrm{t}(n-q).$$

$\hat{\beta}$ is a multinormal vector.
Let's consider a vector of 4 values:

$$\begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \end{pmatrix} \sim \mathcal{N}_4 \left(\begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} ; \sigma^2 \left(\mathbf{X}^T\mathbf{X}\right)^{-1} \right)$$

Let $\mathcal{M}$ be the following model

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \varepsilon_i$$

Why can't we use the following model to test each of the parameters values (here for X_2)?

$$Y_i = \theta_0 + \theta_1 X_{2i} + \varepsilon_i$$

We can't use such a model, we would probably meet a confounding factor: even if we are only interested in relationship X_2 with Y, we have to fit the whole model.

Example 2: Confounding parameter

Let Y be a variable related to the lung cancer. Let X_1 be the smoking status, and X_2 the variable 'alcohol' (for instance the quantity of alcohol drunk per week).
If we only fit the model $\mathcal{M} : Y_i = \theta_0 + \theta_1 X_{2i} + \varepsilon_i$, we could conclude for a relationship between alcohol and lung cancer, because alcohol consumption and smoking is strongly related. If we had fit the model $\mathcal{M} = Y_i = \theta_0 + \theta_1 X_{1i} + \theta_2 X_{2i} + \varepsilon_i$, we could indeed have found no significant relationship between X_2 and Y.

Definition 10: Student law

Let X and Y be two random variables such as $X \perp\!\!\!\perp Y$, and such that $X \sim \mathcal{N}(0,1)$ and $Y \sim \chi^2_n$, then

$$\frac{X}{\sqrt{Y}} \sim \mathcal{S}\mathrm{t}(n)$$

2.8.1. Model comparison

Definition 11: Nested models

Let $\mathcal{M}_2$ and $\mathcal{M}_4$ be two models:
$\mathcal{M}_2 : Y_i = \beta_0 + \beta_3 X_{3_i} + \varepsilon_i$
$\mathcal{M}_4 : Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \varepsilon_i$
$\mathcal{M}_2$ is nested in $\mathcal{M}_4$.

Principle We compare the residual variances of the two models, that is, the variance that is not explained by the model.

The better the model is, the smallest the variance would be.

If everything is explained by the model, the residual variance would be null.

Here $\mathcal{M}_4$ holds all the information found in $\mathcal{M}_2$ plus other informations. In the worst case It would be at least as good as $\mathcal{M}_2$.

2.8.2. Fisher F-test of model comparison

Let $\mathcal{M}_q$ and $\mathcal{M}_{q'}$ be two models such as $\dim(\mathcal{M}_q) = q$, $\dim(\mathcal{M}_{q'}) = q'$, $q > q'$ and $\mathcal{M}_{q'}$ is nested in $\mathcal{M}_q$.

Tested hypotheses

$$\begin{cases} (H_0) : \mathcal{M}_{q'} \text{ is the proper model} \\ (H_1) : \mathcal{M}_q \text{ is a better model} \end{cases}$$

ESS Estimated Sum of Squares

RSS Residual Sum of Squares

EMS Estimates Mean Square

RMS Residual Mean Square

$$ESS = RSS(\mathcal{M}_{q'}) - RSS(\mathcal{M}_q)$$

$$RSS(\mathcal{M}) = \|\mathbf{Y} - \mathbf{X}\hat{\beta}\| = \sum_{i=1}^{n} \hat{\varepsilon}_i^2$$

$$EMS = \frac{ESS}{q - q'}$$

$$RMS = \frac{RSS(\mathcal{M}_q)}{n - q}$$

Under the null hypotheses:

$$F = \frac{EMS}{RMS} \underset{H_0}{\sim} \mathcal{F}(q - q'; n - q)$$

2.9. Model validity

Assumptions:

- $\mathbf{X}$ is a full rank matrix;
- Residuals are i.i.d. $\varepsilon \sim \mathcal{N}(0_n, \sigma^2 \mathcal{I}_n)$;

We have also to look for influential variables.

2.9.1. X is full rank

To check that the rank of the matrix is $p+1$, we can calculate the eigen value of the correlation value of the matrix. If there is a perfect relationship between two variables (two columns of $\mathbf{X}$), one of the eigen value would be null. In practice, we never get a null eigen value. We consider the condition index as the ratio between the largest and the smallest eigenvalues, if the condition index $\kappa = \frac{\lambda_1}{\lambda_p}$, with $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p$ the eigenvalues.

If all eigenvalues is different from 0, $\mathbf{X}^T\mathbf{X}$ can be inverted, but the estimated parameter variance would be large, thus the estimation of the parameters would be not relevant (not good enough).

Variance Inflation Factor Perform a regression of each of the predictors against the other predictors.

If there is a strong linear relationship between a parameter and the others, it would reflect that the coefficient of determination R^2 (the amount of variance explained by the model) for this model, which would mean that there is a strong relationship between the parameters.

We do this for all parameters, and for parameter $j = 1, ..., p$, the variance inflation factor would be:

$$VIF_j = \frac{1}{1 - R_j^2}.$$

Rule If $VIF > 10$ or $VIF > 100$...

In case of multicollinearity, we have to remove the variable one by one until there is no longer multicollinearity. Variables have to be removed based on statistical results and through discussion with experimenters.

2.9.2. Residuals analysis

Assumption

$$\varepsilon \sim \mathcal{N}_n(0_n, \sigma^2 I_n)$$

Normality of the residuals If ε_i $(i = 1, ..., n)$ could be observed we could build a QQ-plot of ε_i/σ against quantiles of $\mathcal{N}(0, 1)$.

Only the residual errors $\hat{e}_i$ can be observed:

Let e_i^* be the studentized residual, considered as estimators of ε_i

$$e_i^* = \frac{\hat{e}_i}{\sqrt{\sigma^2_{(i)(1-H_{ii})}}}$$

$$\begin{aligned}\hat{Y} &= X\hat{\beta} \\ &= X\left((X^TX)^{-1}X^TY\right) \\ &= \underbrace{X(X^TX)^{-1}X^T}_{H}Y\end{aligned}$$

Centered residuals If $(1, \dots, 1)^T$ belongs to $\mathbf{X}$ $\mathbb{E}(\varepsilon) = 0$, by construction.

Independence We do not have a statistical test for independence in R, we would plot the residuals e against $\mathbf{X}\hat{\beta}$.

Homoscedastiscity Plot the $\sqrt{e^*}$ against $\mathbf{X}\hat{\beta}$.

Influential observations We make the distinction between observations:

- With too large residual → Influence on the estimation of σ^2
- Which are too isolated → Influence on the estimation of β

$$e_i^* \sim \mathcal{S}\mathrm{t}(n-p-1)$$

Rule We consider an observation to be aberrant if:

$$e_i^* > \mathrm{F}^{-1}_{\mathcal{S}\mathrm{t}(n-p-1)}(1-\alpha)$$

quantile of order $1-\alpha$, α being often set as $1/n$, or we set the threshold to 2.

Leverage Leverage is the diagonal term of the orthogonal projection matrix(?) H_{ii}.

Property 3.
- $0 \leq H_{ii} \leq 1$
- $\sum_i H_i i = p$

Rule We consider that the observation is aberrant if the leverage is ??.

Non-linearity

2.10. Model Selection

We want to select the best model with the smallest number of predictors.

When models have too many explicative variables, the power of statistical tests decreases.

Different methods:

- Comparison of nested models;
- Information criteria;
- Method based on the prediction error.

2.10.1. Information criteria

Likelihood

Definition 12: Likelihood

Probability to observe what we observed for a particular model.

$$L_n(\mathcal{M}(k))$$

Definition 13: Akaike Information Criterion

$$AIC(\mathcal{M}(k)) = -2\log L_n(\mathcal{M}(k)) + 2k.$$

$2k$ is a penalty, leading to privilege the smallest model.

Definition 14: Bayesian Information Criterion

$$BIC(\mathcal{M}(k)) = -2\log L_n(\mathcal{M}(k)) + \log(n)k.$$

$\log(n)k$ is a penalty.

Usually AIC have smaller penalty than BIC, thus AIC criterion tends to select models with more variables than BIC criterion.

2.10.2. Stepwise

forward Add new predictor iteratively, beginning with the most contributing predictors.

backward Remove predictors iteratively.

stepwise Combination of forward and backward selection. We start by no predictors. We add predictor. Before adding the predictor, we check whether all previously predictors remain meaningful.

The problem with this iterative regression, is that at each step we make a test. We have to reduce the confidence level for multiple test.

In practice, the multiple testing problem is not taken into account in these approaches.

We can use information criteria or model comparison in these methods.

2.11. Predictions

Let X_i the i-th row of the matrix $\mathbf{X}$. The observed value Y_i can be estimated by:

$$\hat{Y}_i = (\mathbf{X}\hat{\beta})_i = X_i\hat{\beta}$$

$$\mathbb{E}(\hat{Y}_i) = (\mathbf{X}\beta)_i = X_i\beta$$

$$\sigma^{-1}(\mathbf{X}\hat{\beta} - \mathbf{X}\beta) \sim \mathcal{N}(0_{p+1}, (\mathbf{X}^T\mathbf{X})^{-1}), \quad \text{and}$$

$$\mathrm{Var}(\hat{Y}_i) = ...$$

$$S^2 = \|...\|$$

Prediction Confidence Interval We can build confidence interval for predicted values $(\mathbf{X}\hat{\beta})_i$

...

Prediction error of Y

Linear Algebra

3 Elements of Linear Algebra

Remark 4: 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 15: Scalar Product (Dot Product)

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

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

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

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

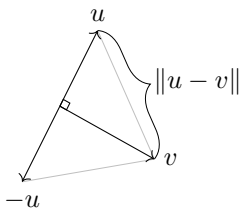


Figure 3.1. Scalar product of two orthogonal vectors.

Definition 16: Norm

Length of the vector.

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

$\|u, v\| > 0$

Definition 17: Distance

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

Definition 18: Orthogonality

Remark 5

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

and

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

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

$$\|u - v\|^2 = \|u\|^2 + \|v\|^2 - 2\langle u, v\rangle$$
$$\|u + v\|^2 = \|u\|^2 + \|v\|^2 + 2\langle u, v\rangle$$

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 2: Pythagorean theorem

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

Definition 19: 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 4. *$proj^w(y)$ is the closest vector to y that belongs to w.*

Definition 20: Matrix

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

Example 3: Matrix application

Let A be a matrix:

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

and

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

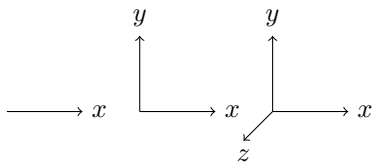


Figure 3.2. Coordinate systems

Example 3 continued

Then,

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

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 21: Tranpose of a Matrix

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