

GENIOMHE

# Multivariate Statistics

---

*by* Samuel Ortion

*Prof.:* Cyril Dalmasso

Fall 2023

# Contents

<b>1. Introduction</b>	<b>4</b>
<b>2. Linear Model</b>	<b>7</b>
2.1. Simple Linear Regression	7
2.2. Generalized Linear Model	7
2.2.1. Penalized Regression	7
2.3. Parameter Estimation	8
2.3.1. Simple Linear Regression	8
2.3.2. General Case	8
2.3.3. Ordinary Least Square Algorithm	8
2.4. Sum of squares	8
2.5. Coefficient of Determination: $R^2$	9
2.6. Gaussian vectors	10
2.6.1. Estimator's properties	12
2.6.2. Estimators properties	12
2.7. Statistical tests	12
2.7.1. Student $t$ -test	12
2.8. Student test of nullity of a parameter	13
2.8.1. Model comparison	14
2.8.2. Fisher $F$ -test of model comparison	14
2.9. Model validity	15
2.9.1. $\mathbf{X}$ is full rank	15
2.9.2. Residuals analysis	15
2.10. Model Selection	16
2.10.1. Information criteria	16
2.10.2. Stepwise	17
2.11. Predictions	17
<b>3. Generalized Linear Model</b>	<b>18</b>
3.1. Logistic Regression	19
3.2. Maximum Likelihood estimator	19
3.3. Test for each single coordinate	19
3.3.1. Comparison of nested models	19
3.4. Relative risk	20
3.5. Odds	20
3.6. Odds Ratio	20
3.7. Poisson model	21

<b>4. Tests Reminders</b>	<b>22</b>
4.1. $\chi^2$ test of independence . . . . .	22
4.2. $\chi^2$ test of goodness of fit . . . . .	22
<b>5. Regularized regressions</b>	<b>23</b>
5.1. Ridge regression . . . . .	23
5.2. Cross validation . . . . .	23
5.2.1. Leave-one-out <i>jackknife</i> . . . . .	23
5.2.2. K-fold cross-validation . . . . .	24
5.3. Lasso regression . . . . .	24
5.4. Elastic Net . . . . .	24
<b>II. Linear Algebra</b>	<b>26</b>
<b>6. Elements of Linear Algebra</b>	<b>27</b>


This work is licensed under a [Creative Commons “Attribution-ShareAlike 4.0 International”](https://creativecommons.org/licenses/by-sa/4.0/) license.



# 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.

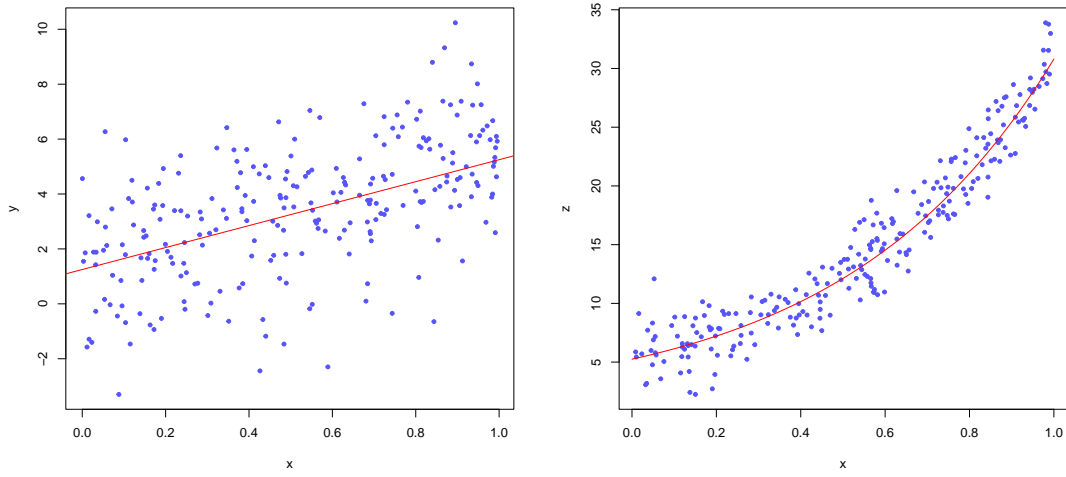


Figure 1.1. Illustration of two models fitting observed values



# 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<sub>1</sub>)  $\varepsilon_i$  are independent;
- (A<sub>2</sub>)  $\varepsilon_i$  are identically distributed;
- (A<sub>3</sub>)  $\varepsilon_i$  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  $p \gg n$  ( $p$ : 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:

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

### 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 6](#)).

$$\Rightarrow \mathbf{X}\beta = \text{proj}^{(1,\mathbf{X})}\mathbf{Y}$$

$$\Rightarrow \forall v \in w, v\mathbf{y} = v\text{proj}^w(y)$$

$$\Rightarrow \forall i :$$

$$\mathbf{X}_i\mathbf{Y} = \mathbf{X}_i\mathbf{X}\hat{\beta} \quad \text{where } \hat{\beta} \text{ is the estimator of } \beta$$

$$\Rightarrow \mathbf{X}^T\mathbf{Y} = \mathbf{X}^T\mathbf{X}\hat{\beta}$$

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

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

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  $\mathbf{X}$ .

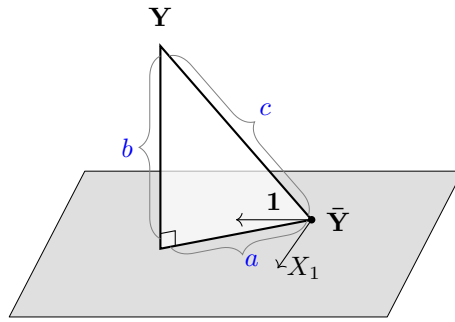
If  $H$  is the projection matrix of the subspace generated by  $\mathbf{X}$ ,  $\mathbf{X}\hat{\beta}$  is the projection on  $\mathbf{Y}$  on this subspace, that corresponds to  $\mathbf{X}\hat{\beta}$ .

## 2.4. Sum of squares

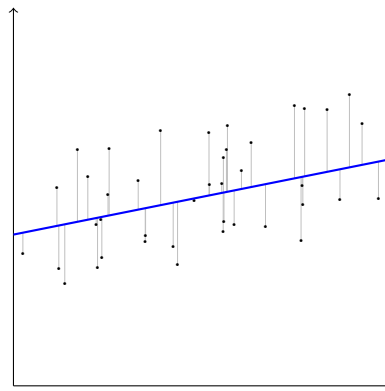
$\mathbf{Y} - \mathbf{X}\hat{\beta} \perp \mathbf{X}\hat{\beta} - \bar{\mathbf{Y}}\mathbf{1}$  if  $\mathbf{1} \in V$ , so

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





**Figure 2.1.** Orthogonal projection of  $\mathbf{Y}$  on plan generated by the base described by  $\mathbf{X}$ .  $a$  corresponds to  $\|\mathbf{X}\hat{\beta} - \bar{\mathbf{Y}}\mathbf{1}\|^2$  and  $b$  corresponds to  $\hat{\varepsilon} = \|\mathbf{Y} - \hat{\beta}\mathbf{X}\|^2$  and  $c$  corresponds to  $\|\mathbf{Y} - \bar{\mathbf{Y}}\mathbf{1}\|^2$ .



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

## 2.5. Coefficient of Determination: $R^2$

$\pi$  **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.

$\pi$  **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  $\beta$  vector.

*Nb.* The dimension of the model is not the number of parameter, as  $\sigma^2$  is one of the model parameters.

## 2.6. Gaussian vectors

### $\pi$ Definition 4: Normal distribution

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

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

### $\pi$ 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  $\Sigma$  is the variance-covariance matrix!

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

### $i$ Remark 1

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

### $\pi$ Definition 6: Covariance

$$\text{Cov}(Y_i, Y_j) = \mathbb{E}((Y_i - \mathbb{E}(Y_i))(Y_j - \mathbb{E}(Y_j)))$$

When two variable are linked, the covariance is large.

If two variables  $X, Y$  are independent,  $\text{Cov}(X, Y) = 0$ .

### $\pi$ Definition 7: Correlation coefficient

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

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 meters. Thus the correlation coefficient, which is a sort of normalized covariance is useful, to be able to compare the values.

**i Remark 2**

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

$$\Sigma = \begin{pmatrix} \mathbb{V}(Y_1) & & & \\ & \ddots & & \\ & & \text{Cov}(Y_i, Y_j) & \\ & & & \mathbb{V}(Y_i) & \ddots & \\ & & & & & \mathbb{V}(Y_n) \end{pmatrix} \quad (2.2)$$

**$\pi$  Definition 8: Identity matrix**

$$J_n = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

**$\pi$  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  $Z_2 = \Pi_{V_2}(\mathbf{Z}) = \Pi_2 \mathbf{Y}$  ( $\Pi_1$  and  $\Pi_2$  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.

**$\pi$  Definition 9: Chi 2 distribution**

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

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

## 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  $\chi^2$  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  $\sigma^2$ .

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

is an unbiased estimator of  $\sigma^2$ .

We can derive statistical test from these properties.

## 2.7. Statistical tests

### 2.7.1. Student $t$ -test

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

where

**Estimation of  $\sigma^2$**  A biased estimator of  $\sigma^2$  is:

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

$S^2$  is the unbiased estimator of  $\sigma^2$

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

**i Remark 3:** On  $\hat{m}$

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

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

## 2.8. Student test of nullity of a parameter

Let  $\beta_j$  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:

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

The test statistic is:

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

$\hat{\beta}$  is a multinormal vector.

Let's consider a vector of 4 values:

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

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  $Y$ .

**π 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  $Y \sim \chi_n^2$ , then

$$\frac{X}{\sqrt{Y}} \sim \mathcal{St}(n)$$

## 2.8.1. Model comparison

**π Definition 11:** Nested models

Let  $\mathcal{M}_2$  and  $\mathcal{M}_4$  be two models:

$$\mathcal{M}_2 : Y_i = \beta_0 + \beta_3 X_{3i} + \varepsilon_i$$

$$\mathcal{M}_4 : Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \varepsilon_i$$

$\mathcal{M}_2$  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. $\mathbf{X}$ 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  $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p$  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  $j = 1, \dots, p$ , the variance inflation factor would be:

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

**Rule** If  $VIF > 10$  or  $VIF > 100\dots$

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**

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

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

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

Let  $e_i^*$  be the studentized residual, considered as estimators of  $\varepsilon_i$

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

$$\begin{aligned} \hat{Y} &= X\hat{\beta} \\ &= X((X^T X)^{-1} X^T Y) \\ &= \underbrace{X(X^T X)^{-1} X^T Y}_H \end{aligned}$$

**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  $\mathbf{X}\hat{\beta}$ .

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

**Influential observations** We make the distinction between observations:

- With too large residual  $\rightarrow$  Influence on the estimation of  $\sigma^2$
- Which are too isolated  $\rightarrow$  Influence on the estimation of  $\beta$

$$e_i^* \sim \mathcal{St}(n - p - 1)$$

**Rule** We consider an observation to be aberrant if:

$$e_i^* > F_{\mathcal{St}(n-p-1)}^{-1}(1 - \alpha)$$

quantile of order  $1 - \alpha$ ,  $\alpha$  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.** •  $0 \leq H_{ii} \leq 1$

- $\sum_i H_{ii} = p$

**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

**$\pi$  Definition 12:** Likelihood

Probability to observe what we observed for a particular model.

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



**π** **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.

**π** **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:

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

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

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

$$\text{Var}(\hat{Y}_i) = \dots$$

$$S^2 = \|\dots\|$$

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

...

**Prediction error of  $Y$**

# 3 Generalized Linear Model

## Example 3

**Ex. 1 - Credit Card Default** Let  $Y_i$  be a boolean random variable following a Bernoulli distribution.

**Ex. 2 - Horseshoe Crabs** Let  $Y_i$ , be the number of satellites males.

$Y_i$  can be described as following a Poisson distribution.

## Remark 4

A Poisson distribution can be viewed as an approximation of binomial distribution when  $n$  is high and  $p$  low.

We will consider the following relation:

$$\mathbb{E}(Y_i) = g^{-1}X_i\beta,$$

equivalently:

$$g(\mathbb{E}(Y_i)) = X_i\beta.$$

- $\beta$  is estimated by the maximum likelihood;
- $g$  is called the link function.

## Remark 5

In standard linear model, the OLS estimator is the estimator of maximum of likelihood.

## 3.1. Logistic Regression

$$\begin{aligned}
 \log\left(\frac{\Pi}{1-\Pi}\right) &= \mathbf{X}\beta \\
 \Leftrightarrow e^{\ln\frac{\Pi}{1-\Pi}} &= e^{\mathbf{X}\beta} \\
 \Leftrightarrow \frac{\Pi}{1-\Pi} &= e^{\mathbf{X}\beta} \\
 \Leftrightarrow \Pi &= (1-\Pi)e^{\mathbf{X}\beta} \\
 \Leftrightarrow \Pi &= e^{\mathbf{X}\beta} - \Pi e^{\mathbf{X}\beta} \\
 \Leftrightarrow \Pi + \Pi e^{\mathbf{X}\beta} &= e^{\mathbf{X}\beta} \\
 \Leftrightarrow \Pi(1 + e^{\mathbf{X}\beta}) &= e^{\mathbf{X}\beta} \\
 \Leftrightarrow \Pi &= \frac{e^{\mathbf{X}\beta}}{1 + e^{\mathbf{X}\beta}}
 \end{aligned}$$

## 3.2. Maximum Likelihood estimator

log-likelihood: the probability to observe what we observe.

Estimate  $\beta$  by  $\hat{\beta}$  such that  $\forall \beta \in \mathbb{R}^{p+1}$ :

$$L_n(\hat{\beta}) \geq L_n(\beta)$$

These estimators are consistent, but not necessarily unbiased.

## 3.3. Test for each single coordinate

### Example 4: Payment Default

Let  $Y_i$  be the default value for individual  $i$ .

$$\log\left(\frac{\Pi(X)}{1-\Pi(X)}\right) = \beta_0 + \beta_1 \text{student} + \beta_2 \text{balance} + \beta_3 \text{income}$$

In this example, only  $\beta_0$  and  $\beta_2$  are significantly different from 0.

### Remark 6

We do not add  $\varepsilon_i$ , because  $\log\left(\frac{\Pi(X)}{1-\Pi(X)}\right)$  corresponds to the expectation.

### 3.3.1. Comparison of nested models

To test  $H_0 : \beta_0 = \dots = \beta_p = 0$ , we use the likelihood ratio test:

$$T_n = -2 \log(\mathcal{L}^{\text{null}}) + 2 \log(\mathcal{L}(\hat{\beta})) \xrightarrow[n \rightarrow \infty]{H_0} \chi^2(p).$$

**i Remark 7:** Family of Tests

- Comparison of estimated values and values under the null hypothesis;
- Likelihood ratio test;
- Based on the slope on the derivative.

## 3.4. Relative risk

$RR_i$  is the probably to have the disease, conditional to the predictor  $X_{i1}$  over the probability of having the disease, conditional to the predictor  $X_{i2}$ .

$$RR(j) = \frac{\mathbb{P}(Y_{i_1} = 1 | X_{i_1})}{\mathbb{P}(Y_{i_2} = 1 | X_{i_2})} = \frac{\mathbb{E}(Y_{i_1})}{\mathbb{E}(Y_{i_2})}.$$

$\pi(X_i)$  is the probability of having the disease, according to  $X_i$ .  
The relative risk can be written as...

## 3.5. Odds

Quantity providing a measure of the likelihood of a particular outcome:

$$odd = \frac{\pi(X_i)}{1 - \pi(X_i)}$$

$$odds = \exp(X_i\beta)$$

odds is the ratio of people having the disease, if Y represent the disease, over the people not having the disease.

## 3.6. Odds Ratio

$$OR(j) = \frac{odds(X_{i_1})}{odds(X_{i_2})} = \frac{\frac{\pi X_{i_1}}{1 - \pi(X_{i_1})}}{\frac{\pi X_{i_2}}{1 - \pi(X_{i_2})}}$$

The OR can be written as:

$$OR(j) = \exp(\beta_j)$$

**Exercise 1:**

Show that  $OR(j) = \exp(\beta_j)$ .

$$\begin{aligned} OR(j) &= \frac{odds(X_{i_1})}{odds(X_{i_2})} \\ &= \frac{\exp(X_{i_1}\beta)}{\exp(X_{i_2}\beta)} \end{aligned}$$

$$\log \left( \frac{\mathbb{P}(Y = 1 | X_{i_1})}{1 - \mathbb{P}(Y = 1 | X_{i_1})} \right) = \beta_0 + \beta_1 X_1^{(1)} + \beta_2 X_2^{(1)} + \dots + \beta_p X_p^{(1)}$$

Similarly

$$\log \left( \frac{\mathbb{P}(Y = 1 | X_{i_2})}{1 - \mathbb{P}(Y = 1 | X_{i_2})} \right) = \beta_0 + \beta_1 X_1^{(2)} + \beta_2 X_2^{(2)} + \dots + \beta_p X_p^{(2)}$$

We subtract both equations:

$$\begin{aligned} & \log \left( \frac{\mathbb{P}(Y = 1 | X_{i_1})}{1 - \mathbb{P}(Y = 1 | X_{i_1})} \right) - \log \left( \frac{\mathbb{P}(Y = 1 | X_{i_2})}{1 - \mathbb{P}(Y = 1 | X_{i_2})} \right) \\ &= \beta_0 + \beta_1 X_1^{(1)} + \beta_2 X_2^{(1)} + \dots + \beta_p X_p^{(1)} - \beta_0 - \beta_1 X_1^{(2)} - \beta_2 X_2^{(2)} - \dots - \beta_p X_p^{(2)} \\ &= \log OR(j) \\ &= (\cancel{\beta_0 - \beta_0}) + \beta_1 (\cancel{X_1^{(1)} - X_1^{(2)}}) + \beta_2 (\cancel{X_2^{(1)} - X_2^{(2)}}) + \dots + \beta_j (\cancel{X_j^{(1)} - X_j^{(2)}}) + \dots + \beta_p (\cancel{X_p^{(1)} - X_p^{(2)}}) \\ &\Leftrightarrow \log(OR_j) = \beta_j \\ &\Leftrightarrow OR(j) = \exp(\beta_j) \end{aligned}$$

OR is not equal to RR, except in the particular case of probability (?)

If OR is significantly different from 1, the  $\exp(\beta_j)$  is significantly different from 1, thus  $\beta_j$  is significantly different from 0.

If we have more than two classes, we do not know what means  $X_{i_1} - X_{i_2} = 0$ . We will have to take a reference class, and compare successively each class with the reference class.

$\hat{\pi}(X_+) = \hat{\mathbb{P}}(X = 1 | X_{i_1})$  for a new individual.

## 3.7. Poisson model

Let  $Y_i \sim \mathcal{P}(\lambda_i)$ , corresponding to a counting.

$$\begin{aligned} \mathbb{E}(Y_i) &= g^{-1}(X_i \beta) \\ \Leftrightarrow g(\mathbb{E}(Y_i)) &= X_i \beta \end{aligned}$$

where  $g(x) = \ln(x)$ , and  $g^{-1}(x) = e^x$ .

$$\lambda_i = \mathbb{E}(Y_i) = \text{Var}(Y_i)$$

# 4 Tests Reminders

## 4.1. $\chi^2$ test of independence

[...]

## 4.2. $\chi^2$ test of goodness of fit

Check if the observations is in adequation with a particular distribution.



### Example 5: Mendel experiments

Let  $AB$ ,  $Ab$ ,  $aB$ ,  $ab$  be the four possible genotypes of peas: colors and grain shape.

$AB$	$Ab$	$aB$	$ab$
315	108	101	32

The test statistics is:

$$D_{k,n} = \sum_{i=1}^k \frac{(N_i - np_i)^2}{np_i} \xrightarrow[n \rightarrow \infty]{\mathcal{L}_{H_0}} \chi_{(n-1)(q-1)}^2$$

# 5 Regularized regressions

Let  $\mathbf{Y}$  be a vector of observations and  $\mathbf{X}$  a matrix of dimension  $n \times (p + 1)$ . Suppose the real model is:

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

if  $p$  is large compared to  $n$ :

- $\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}$  is not defined as  $\mathbf{X}^T \mathbf{X}$  is not invertible.
- $m^*$  is the number of true predictors, that is, the number of predictor with non-zero values.
- 
- 

## 5.1. Ridge regression

Instead of minimizing the mean square error, we want to minimize the following regularize expression:

$$\hat{\beta}_\lambda^{\text{ridge}} = \arg \min_{\beta \in \mathbb{R}^p} \|\mathbf{Y} - \mathbf{X}\beta\|^2 + \lambda \sum_{j=1}^p \beta_j^2$$

it is a way to favor the solution with small values for parameters. where  $\lambda$  is used to calibrate the regularization.

$$\sum_{j=1}^p \beta_j^2 = \|\beta_j\|^2$$

is the classical square norm of the vector.

## 5.2. Cross validation

### 5.2.1. Leave-one-out *jackknife*

### Example 6

Let  $\mathcal{M}_0$  be the model  $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i}$   
 The model will be:

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \end{pmatrix} = \beta_0 + \beta_1 \begin{pmatrix} x_{11} \\ x_{12} \\ x_{13} \\ x_{14} \\ x_{15} \end{pmatrix} + \beta_2 \begin{pmatrix} x_{21} \\ x_{22} \\ x_{23} \\ x_{24} \\ x_{25} \end{pmatrix} + \beta_3 \begin{pmatrix} x_{31} \\ x_{32} \\ x_{33} \\ x_{34} \\ x_{35} \end{pmatrix}$$

1	2	3	4	5
.	×	×	×	×
×	.	×	×	×
×	×	.	×	×
×	×	×	.	×
×	×	×	×	.

We perform computation of  $\lambda$  for each dataset without one observation.

## 5.2.2. K-fold cross-validation

We will have as many tables as subsets.

We chose lambda such that the generalization error is the smallest.

## 5.3. Lasso regression

The difference with the Ridge regression lies in the penalty:

$$\hat{\beta}_\lambda^{\text{lasso}} = \arg \min \|Y - X\beta\|^2 + \sum_{j=1}^p |\beta_j|$$

$$\sum_{j=1}^p |\beta_j| = \|\beta\|_1$$

Instead of having a smooth increasing for each parameters, each parameters will enter iteratively in the model. Some parameters can be set to 0.

Lasso regression can be used to perform variable selection.

We can use the same methods (K-fold and Leave-one-out) to select the  $\lambda$  value.

## 5.4. Elastic Net

Combination of the Ridge and Lasso regression:

$$\hat{\beta}_\lambda^{\text{en}} = \arg \min \|Y - X\beta\|^2 + \lambda_1 \|\beta\|_1 + \lambda_2 \|\beta\|_2^2$$



 **Remark 8**

In the case of Lasso, Elastic net or Ridge regression, we can no longer perform statistical test on the parameters.

# Linear Algebra

# 6 Elements of Linear Algebra

## **i** Remark 9: vector

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

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

## **$\pi$** Definition 15: Scalar Product (Dot Product)

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

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

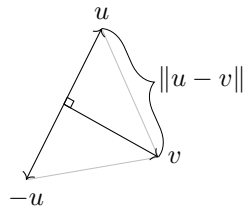
### Dot product properties

**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 6.1.** Scalar product of two orthogonal vectors.

**$\pi$  Definition 16:** Norm

Length of the vector.

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

$$\|u, v\| > 0$$

**$\pi$  Definition 17:** Distance

$$\text{dist}(u, v) = \|u - v\|$$

**$\pi$  Definition 18:** Orthogonality

**$i$  Remark 10**

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

and

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

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

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

**$\pi$  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 6.1](#).

$$\Leftrightarrow -2\langle u, v \rangle = 2\langle u, v \rangle$$

$$\Leftrightarrow 4\langle u, v \rangle = 0$$

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

□

**$\pi$  Theorem 2:** Pythagorean theorem

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

**$\pi$  Definition 19:** Orthogonal Projection

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

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

where

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

There is only one vector  $y$  that ?

The scalar product between  $z$  and (?) is zero.

**Property 4.**  $\text{proj}^w(y)$  is the closest vector to  $y$  that belongs to  $w$ .

**$\pi$  Definition 20:** Matrix

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

**$\text{📄}$  Example 7:** Matrix application

Let  $A$  be a matrix:

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

and

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

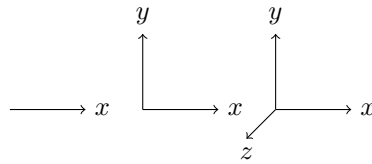


Figure 6.2. Coordinate systems

 Example 7 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  $A^T = \begin{pmatrix} a & c \\ b & d \end{pmatrix}$