MSc DS & AI

2023 - 2024

statistical inference part 2

Logistic regression

①

I/ The model

Let consider a dataset with $\underline{y} = (y_1, \dots, y_n)'$ and $\underline{x} = (\underline{x}_1', \dots, \underline{x}_n')'$ when $y_i \in \{0, 1\}$ and $\underline{x}_i = (x_{i1}, \dots, x_{id})' \in \mathbb{R}^d$.

Let also denote $\mathcal{D} = \{(\underline{x}_1, y_1), \dots, (\underline{x}_n, y_n)\}$ the dataset composed with each instance with its label.

Ex

Y	X_1	.....	X_d
1	20	.....	5
1	25	.....	4
0	14	.....	0

The goal is to predict y_i based on $\underline{x}_i$

$\boxed{y_i \rightsquigarrow \mathcal{B}(\pi(\underline{x}_i))}$: y_i is assumed to follow a Bernoulli distribution (success or fail) where the probability of success depends on $\underline{x}_i$ (the covariates).

$$p(y_i \mid \underline{x}_i) = \begin{cases} \pi(\underline{x}_i) & \text{if } y_i = 1 \\ 1 - \pi(x_i) & \text{if } y_i = 0 \end{cases} \quad \text{or equivalently } p(y_i \mid \underline{x}_i) = \pi(\underline{x}_i)^{y_i} (1 - \pi(x_i))^{1-y_i}$$

Let also notice that $E[y_i \mid \underline{x}_i] = \pi(\underline{x}_i)$

Logistic regression assumes that

$$\boxed{\begin{aligned} \operatorname{logit}(\pi(\underline{x}_i)) &:= \log\left(\frac{\pi(\underline{x}_i)}{1 - \pi(\underline{x}_i)}\right) \\ &= \beta_0 + \beta_1 x_{i1} + \dots + \beta_d x_{id} \\ &= \beta' \underline{\tilde{x}}_i \end{aligned}}$$

$\Longleftrightarrow$ $\operatorname{logit}^{-1}$, $\pi(\underline{x}_i) = \operatorname{logit}^{-1}(\beta' \underline{\tilde{x}}_i)$

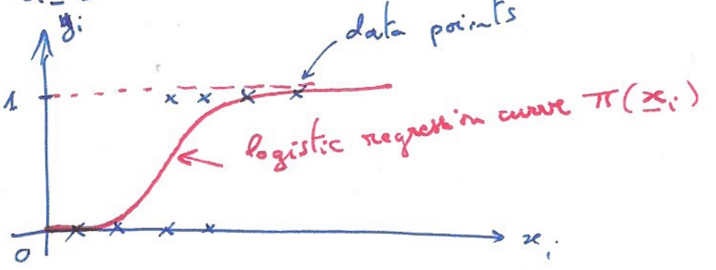
$$= \sigma(\beta' \underline{\tilde{x}}_i)$$

with $\sigma : z \longmapsto \dfrac{1}{1 + \exp(-z)}$

with $\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_d \end{pmatrix}$ the vector of parameters to estimate

and $\underline{\tilde{x}}_i = (1, \underline{x}_i')' = \begin{pmatrix} 1 \\ x_{i1} \\ \vdots \\ x_{id} \end{pmatrix}$

Case $d = 1$



The statistical inference issue consist in estimating β (the vector of parameters) based on the dataset $\mathcal{D}$.

Strategy : → Build an estimator $\hat{\beta}$ of β by maximum likelihood

⚠ $\hat{\beta}$ is a random variable since it depend on the data y_i which is random

→ Study the properties of $\hat{\beta}$: asymptotically normal/Gaussian thus allows to compute confidence interfals and to make tests on the parameters.

II/ Parameters estimation

Let consider $p(\underline{y} \mid \underline{x}) = p(y_1, \dots, y_m \mid \underline{x}_1, \dots, \underline{x}_m)$ it is the probability of the responses given the covariates (conditionnal likehood) ~~our model~~

Since the data $(\underline{x}_1, y_1), \dots, (\underline{x}_m, y_m)$ are assumed to be independent we can write :

$$\underbrace{p(\underline{y} \mid \underline{x})}_{:\mathcal{L}(\beta)} = \prod_{i=1}^{m} p(y_i \mid \underline{x}_i) = \prod_{i=1}^{m} \pi(\underline{x}_i)^{y_i} (1 - \pi(\underline{x}_i))^{1-y_i}$$

The likelihood which is a function of β (since $\pi(\underline{x}_i)$ is a function of β)

We often consider the log-likelihood $\ell(\beta) = \log \mathcal{L}(\beta)$.

$$\ell(\beta) = \sum_{i=1}^{m} \left[y_i \log \pi(\underline{x}_i) + (1 - y_i) \log (1 - \pi(\underline{x}_i)) \right]$$

$$\ell(\beta) = \sum_{i=1}^{m} \left[y_i \log \left(\frac{1}{1 + \exp(-\beta' \tilde{\underline{x}}_i)} \right) + (1 - y_i) \log \left(\frac{\exp(-\beta' \tilde{\underline{x}}_i)}{1 + \exp(-\beta' \tilde{\underline{x}}_i)} \right) \right]$$

$$\ell(\beta) = \sum_{i=1}^{m} y_i \beta' \tilde{\underline{x}}_i - \log(1 + \exp(\beta' \tilde{\underline{x}}_i))$$

~~Ef~~ $\ell : \mathbb{R}^{d+1} \to \mathbb{R}$. We want to maximise ℓ with respect to β thus we will compute the grandient.

$$\frac{\partial \ell(\beta)}{\partial \beta} = \sum_{i=1}^{m} y_i \tilde{\underline{x}}_i - \tilde{\underline{x}}_i \frac{\exp(\beta' \tilde{\underline{x}}_i)}{1+\exp(\beta' \tilde{\underline{x}}_i)} = \sum_{i=1}^{m} \tilde{\underline{x}}_i (y_i - \pi(\underline{x}_i))$$

$\frac{\partial \ell(\beta)}{\partial \beta} \in \mathbb{R}^{d+1}$. Solving $\frac{\partial \ell(\beta)}{\partial \beta} = 0$ has not closed-form

thus it is needed to use an iterative algorithm such as Newton-Raphson or gradient descent.

<u>Newton Raphson</u>

$$\beta^{(\kappa+1)} = \beta^{(\kappa)} - \underbrace{\left(\frac{\partial^2 \ell(\beta)}{\partial \beta \partial \beta'}\right)^{-1}_{\Big|\beta=\beta^{(\kappa)}}}_{\text{Hessian matrix}} \times \left(\frac{\partial \ell(\beta)}{\partial \beta}\right)_{\Big|\beta=\beta^{(\kappa)}}$$

We have $$\frac{\partial^2 \ell(\beta)}{\partial \beta \partial \beta'} = -\tilde{X}' V \tilde{X}$$

where $\tilde{X}$ is the matrix ~~with~~ $m\times(d+1)$ composed with $\tilde{\underline{x}}_i$ in rows

and V is the diagonal matrix of $\pi(\underline{x}_i)(1-\pi(\underline{x}_i))$

$$V = \begin{pmatrix} \pi(\tilde{\underline{x}}_1)(1-\pi(\tilde{\underline{x}}_1)) & & & & 0 \\ & \ddots & & & \\ & & \pi(\tilde{\underline{x}}_i)(1-\pi(\tilde{\underline{x}}_i)) & & \\ & & & \ddots & \\ 0 & & & & \pi(\tilde{\underline{x}}_m)(1-\pi(\tilde{\underline{x}}_m)) \end{pmatrix}$$

Let notice that $\frac{\partial^2 \ell(\beta)}{\partial \beta \partial \beta'}$ is definite negative if $\tilde{X}$ is of maximal rank (i.e. $d+1$ if $d+1 \leq m$).

III / Properties of $\hat{\beta}$

Since $\hat{\beta}$ is the maximum likelihood estimator it is asymptotically ~~normal~~ unbiaised and follows asymptotically a normal distribution with asymptotic covariance matrix $\hat{V}(\hat{\beta}) = \left[- \frac{\partial^2 \ell(\beta)}{\partial \beta \partial \beta'}\Big|_{\beta=\hat{\beta}} \right]^{-1}$

which is the inverse of the Fisher information matrix.

Thus $\hat{\beta} \underset{\text{as } n \to +\infty}{\approx} \mathcal{N}_{d+1}\left(\beta, (\tilde{X}'\hat{V}\tilde{X})^{-1}\right)$

$\Rightarrow$ Thus possible to make similar computations as in the linear regression. Since n assumed to be large we only rely on ~~Student (t)~~ ~~distri~~ normal distribution and not anymore on Student (t)

IV / Tests, confidence intervals and model choice

A/ Test on β_j

Let consider testing if variable j has an effect or not:

$$H_0 : \beta_j = 0 \quad \text{against} \quad H_1 : \beta_j \neq 0$$

There are three possibilities:

- Likelihood ratio test:

$$LRT = 2 \log \frac{\max_{\beta} \mathcal{L}(\beta)}{\max_{\beta \text{ s.c } \beta_j = 0} \mathcal{L}(\beta)} = 2 \log \frac{\max_{\beta} \mathcal{L}_{H_1}(\beta)}{\max_{\beta} \mathcal{L}_{H_0}(\beta)}$$

$$= 2\Big[\underbrace{\ell(\hat{\beta})}_{\substack{\text{maximum} \\ \text{log-likelihood}}} - \underbrace{\ell(\hat{\beta}_{H_0})}_{\substack{\text{maximum log-likelihood} \\ \text{under } H_0}} \Big] = D_0 - D_1 \quad \text{with } D_0 = -2\,\ell(\hat{\beta}_{H_0}) \text{ and } D_1 = -2\,\ell(\hat{\beta})$$

Under H_0 : $LRT \rightsquigarrow \chi^2_1$

- Wald test: we know that $\dfrac{\hat{\beta}_j - \beta_j}{\hat{\sigma}(\hat{\beta}_j)} \approx \mathcal{N}(0,1)$

Thus under H_0 : $\dfrac{\hat{\beta}_j}{\hat{\sigma}(\hat{\beta}_j)} \approx \mathcal{N}(0,1) \Rightarrow \dfrac{\hat{\beta}_j^2}{\hat{\sigma}(\hat{\beta}_j)^2} \rightsquigarrow \chi^2_1$

- Score test

$$U(\hat{\beta}_{H_0})' \, \hat{V}(\hat{\beta}_{H_0}) \, U(\hat{\beta}_{H_0}) \longrightarrow \chi^2_1$$

where $\hat{V}(\hat{\beta}_{H_0})$ is the inverse of the Fisher information matrix and $U(\hat{\beta}_{H_0})$ the vector of partial derivative of the log-likelihood both estimated under H_0.

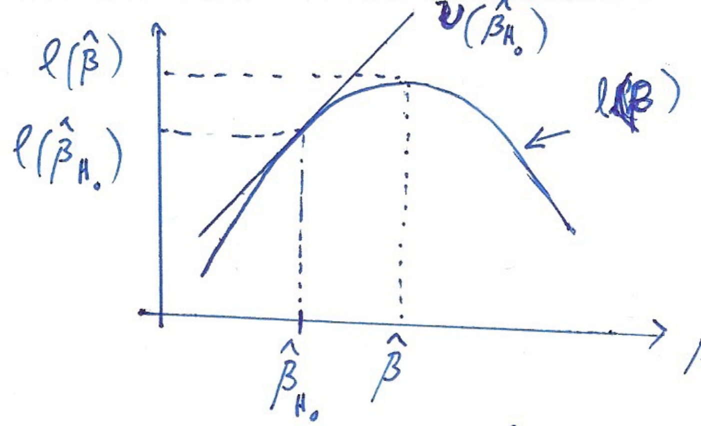


Illustration for only one parameter

B/ Test on ~~linear~~ ~~comb~~ a linear combination of the parameters $\overset{q\times(k-1)}{C}\beta = 0$

We can still use the test defined in previous section extended to the multivariate framework.

Let first notice that $C\hat\beta \approx \mathcal{N}\left(C\beta, C(\tilde X'\hat V\tilde X)^{-1}C'\right)$

Thus under H_0 : $C\hat\beta \approx \mathcal{N}\left(0, C(\tilde X'\hat V\tilde X)^{-1}C'\right)$

And consequently $(C\hat\beta)'\left(C(\tilde X'\hat V\tilde X)^{-1}C'\right)^{-1}(C\hat\beta) \approx \chi^2_q$

Thus possible to use ~~test~~ ~~similar~~ ~~to~~ Wald test.

The likelihood ratio test is defined by

$$D_0 - D_1 = 2\left(\ell(\hat\beta) - \ell(\hat\beta_{H_0})\right) \underset{H_0}{\approx} \chi^2_q$$

Wher $\ell(\hat\beta_{H_0}) = \max\limits_{\substack{\beta \\ \text{ss } C\beta=0}} \ell(\beta)$

And the score test

$$U(\hat\beta_{H_0})'\hat V(\hat\beta_{H_0})U(\hat\beta_{H_0}) \longrightarrow \chi^2_q$$

C/ Model choice

$$BIC = -2\ell(\hat\beta) + (d+1)\log n$$

$$AIC = -2\ell(\hat\beta) + (d+1)$$

where $d+1$ is the number of estimated parameters

AIC and BIC can be used to put models into competitions: the goal is to find the model (subset of variables) minimizing the criterium. This can be done for instance by using a stepwise approach. (forward, backward or forward-backward)

BIC tends to select model of lower dimension than AIC

By default the step function of R uses AIC.

D/ Confidence interval

Since $\hat{\beta}_j \underset{approx}{\approx} \mathcal{N}(\beta_j, \hat{\sigma}(\hat{\beta}_j))$

One can deduce a $(1-\alpha)$ confidence interval by the formula $\hat{\beta}_j \pm z_{\alpha/2}\,\hat{\sigma}(\hat{\beta}_j)$ with $z_{\alpha/2}$ the $\alpha/2$ upper quantile of the normal distribution.

V Case of categorical features

Let assume that a variable x_j is categorical, with $x_j \in \{1, \dots, J\}$

Thus a binary coding can be used

x_j	$x_{j,blue}$	$x_{j,red}$	$x_{j,green}$
blue	1	0	0
red	0	1	0
green	0	0	1

↑ Reference level not used in design matrix

By default in R the first level of the variable is used as reference level (the column) is not used to fit the model

Testing if variable j has an effect consist in testing

$$H_0 : \beta_{red} = \beta_{green} = 0 \quad vs \quad H_1 : \beta_{red} \neq 0 \text{ or } \beta_{green} \neq 0.$$

(possible : see IV B)

$$\text{Odds}(\underline{x}_i) = \frac{\pi(\underline{x}_i)}{1-\pi(\underline{x}_i)} = \exp(\beta' \tilde{\underline{x}}_i)$$

$$\text{odds-ratio}(\underline{x}_i, \underline{x}_{i'}) = \frac{\text{odds}(\underline{x}_i)}{\text{odds}(\underline{x}_{i'})} = \exp\left(\beta'(\tilde{\underline{x}}_i - \tilde{\underline{x}}_{i'})\right)$$

If i and i' differ for only one variable $x_{ij} \neq x_{i'j}$ then

$$\text{Odds-ratio}(\underline{x}_i, \underline{x}_{i'}) = \exp(\beta_j (x_{ij} - x_{i'j}))$$ ~~and this variable is categorical~~

$$\widehat{\text{Odds-ratio}}(\underline{x}_i, \underline{x}_{i'}) = \exp(\hat{\beta}_j (x_{ij} - x_{i'j}))$$

Then it is possible to derive a confidence interval on Odd-ratio $(\underline{x}_i, \underline{x}_{i'})$:

$$\left[\exp\left(\left(\hat{\beta}_j \pm z_{\alpha/2}\, \hat{\sigma}(\hat{\beta}_j)\right)(x_{ij} - x_{i'j})\right)\right]$$

If the variable is categorical $\exp(\beta_j)$ gives the odds ratio between the considered level and the reference level

Remark can add interaction in the model by creating new variables for instance $x_{new} = x_{i1} \times x_{i2}$

VI Multi-class logistic regression $y_i \in \{1, \dots, K\}$

$$p(y_i = k \mid x_i) = \frac{\exp(\beta_k' \tilde{x}_i)}{1 + \sum_{h=1}^{K-1} \exp(\beta_h' \tilde{x}_i)} \quad \text{for } k \in \{1, \dots, K-1\}$$

where $\beta_k = \begin{pmatrix} \beta_{k0} \\ \beta_{k1} \\ \vdots \\ \beta_{kd} \end{pmatrix}$

$$p(y_i = K \mid x_i) = \frac{1}{1 + \sum_{h=1}^{K-1} \exp(\beta_h' \tilde{x}_i)}$$

$\Rightarrow$ Maximum likelihood ...

Can need regularization if too many parameters to estimate