Lecture 9 : Methods for Regression

Generalized linar model

K. Meziani

Section 1

1. Introduction

Introduction

The best prediction of Y conditionnaly to x is the regression function $h(x) = \mathbb{E}[Y|x]$.
In previous chapter, we assumed $h(x)$ is linear with respect to x
$h(x) = \mathbb{E}[Y|x] = x^T\beta$, $s.t.$

$$\boxed{Y = x^T\beta + \varepsilon, \text{ with } \varepsilon \sim \mathcal{N}(0, \sigma^2).}$$

Problem : One can not deal with categorial responses, classification

Introduction

Introduce new models but keep the linear link $\eta(x) = x^T\beta$ $s.t.$

$$g(E_\beta[Y|x]) = x^T\beta,$$

where $g(\cdot) = h^{-1}$ is called the link function. Therefore,

$$\mathbb{E}[Y|x] = g^{-1}(\eta(x)) = g^{-1}(x^T\beta). \tag{1}$$

Method

1. Choose the prob. distribution of $Y|x$ among the natural exponential family.
2. Set $\eta(x) := x^T\beta$ and choose a "good" link function. Usually, one choose the canonical link function.
3. Estimate the unknown parameter β by $\widehat{\beta}_n$ from a n-sample $(Y_i, x_i)_{i=1,\cdots,n}$. Therefore,

$$g^{-1}(X\widehat{\beta}_n) \quad \text{where} \quad X = (x_1, \cdots, x_n)^T.$$

Natural exponential family

Definition

We say that a random variable Y has a probability density, with respect to a dominant measure ν, denoted by $f_{\theta,\phi}$ belonging to the natural exponential family $\mathcal{F}_\theta^{Nat}$ if $f_{\theta,\phi}$ is written

$$f_{\theta,\phi}(y) = \exp\left(\frac{y\theta - b(\theta)}{\phi} + c(y,\phi)\right), \qquad (2)$$

where $b(\cdot)$ and $c(\cdot)$ are known and differentiable functions such as

- $b(\cdot)$ is 3 times differentiable,
- $b'(\cdot)$ is invertible, *i.e.* $(b')^{-1}(\cdot)$ exists.
- $\theta \in \Theta \subseteq \mathbb{R}$, $\phi \in \mathcal{B} \subseteq \mathbb{R}_*^+$ is the natural parameter and ϕ the dispersion parameter.

Natural exponential family

Proposition

If Y admits a density belonging to the natural exponential family $\mathcal{F}_\theta^{Nat}$ then

1. $\mathbb{E}_\theta[Y] = b'(\theta)$.
2. $\mathbb{V}\mathrm{ar}_\theta[Y] = b''(\theta)\phi$.

Natural exponential family

Definition

Let Y be a random variable which admits a density belonging to the natural exponential family$\mathcal{F}_{\theta}^{Nat}$, $s.t.$

$$\mathbb{E}_{\theta}[Y] = b'(\theta) = \mu,$$

alors la fonction

$$g(\mu) = (b')^{-1}(\mu) \tag{3}$$

is called **the canonical link**.

Canonical link

Choice of the law of $Y\|x$	Ber(p)/Bin(N, p)	Poisson	Gamma	Gausian
Link function canonique	$g(\mu) = \text{logit}(\mu)$ $= \log\left(\frac{\mu}{N-\mu}\right)$	$g(\mu) = \log(\mu)$	$g(\mu) = -\frac{1}{\mu}$	$g(\mu) = \mu$
Name link	logit	log	reciprocal	identity

with $\mu(x) = \mathbb{E}[Y|x] = g^{-1}(\eta(x)) = g^{-1}(x^T\beta)$.

Remarks

☛ In the setting of the “logit link”, we speak of logistic regression, and in the setting of a “logarithmic link”, we speak of poisson regression.

☛ Other non-canonical link functions are used in practice. The probit link: : $g(\mu) = \Phi^{-1}(\mu)$ where $\Phi(\cdot)$ is the distribution function of a reduced centered Gaussian. The log-log : $g(\mu) = \log(-\log(1-\mu))$ with $\mu \in]0, 1[$.

Logistic regression

For sake of simplicity, consider a binary variable Y, *i.e.* Y takes its values in $\{0, 1\}$.

1. The choice of the law of $Y|x$ will naturally be carried on a Bernoulli law of parameter

$$p(x) = P(Y = 1|x) \text{ and } \mu(x) = \mathbb{E}[Y|x] = p(x).$$

2. We choose the canonical link logit

$$g(\mu(x)) = g(p(x)) = \text{logit}(p(x)) = \log\left(\frac{p(x)}{1 - p(x)}\right).$$

3. For $\eta(x) = x^T\beta$ and for $\widehat{\beta}_n$ a "good" estimator of β built from n observations, we estimate $\mathbb{E}[Y|x] = p(x)$ by

$$\widehat{p}(x) = g^{-1}(\widehat{\eta}(x)) = g^{-1}(x^T\widehat{\beta}_n) = \frac{e^{x^T\widehat{\beta}_n}}{1 + e^{x^T\widehat{\beta}_n}}.$$

4. We assign the value 1 to $\widehat{Y}_i$ if $\widehat{p}_i = \widehat{p}(x_i) > s$ where $s = 0.5$ for example.

Section 2

2. Maximum likehood estimator (MLE)

Maximum likehood estimator (MLE)

Denote by $Y = (Y_1, \cdots, Y_n)^\top$ and the design matrix

$$X = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix} = (X_1, \cdots, X_p) = \begin{pmatrix} x_1^T \\ \vdots \\ x_n^T \end{pmatrix},$$

where the X_j, $j = 1 \cdots, p$ are the *explanatory variables*.

Maximum likehood estimator (MLE)

Let us denote by $\mathcal{L}(\beta)$ the log of the likelihood function. The Y_i being independent, it comes

$$\mathcal{L}(\beta) = \sum_{i=1}^{n} \log f_{\theta_i,\phi}(Y_i) = \sum_{i=1}^{n} \mathcal{L}_i(\beta),$$

where $\mathcal{L}_i(\beta)$ is the contribution of the $i^{\text{ième}}$ observation (Y_i, x_i), to the log of the likelihood

$$\mathcal{L}_i(\beta) = \ell(Y_i, \theta_i, \phi, \beta) = \log f_{\theta_i,\phi}(Y_i) = \frac{Y_i\theta_i - b(\theta_i)}{\phi} + c(Y_i, \phi).$$

The likelihood equations

Proposition

The likelihood equations are

$$\frac{\partial \mathcal{L}(\beta)}{\partial \beta_j} = \sum_{i=1}^{n} \frac{Y_i - \mu_i}{\mathbb{V}\mathrm{ar}[Y_i]} h'(\eta_i) x_{i,j} = 0, \quad j = 1, \cdots, p$$

In matrix form, the gradient is written:

$$\nabla \mathcal{L}(\beta) = \left[\frac{\partial \mathcal{L}(\beta)}{\partial \beta_1}, \cdots, \frac{\partial \mathcal{L}(\beta)}{\partial \beta_p} \right]^T = 0_p.$$

For the canonical link, the likelihood equations are simplified:

$$\sum_{i=1}^{n} \frac{(Y_i - \mu_i) x_{i,j}}{\phi} = 0, \quad j = 1, \cdots, p. \tag{4}$$

Example

$\pi(x_i)$

Let $Y_i|x_i \sim \mathcal{B}(\pi_i)$, then $\mu_i = \pi_i = \frac{e^{x_i^T\beta}}{1+e^{x_i^T\beta}}$ et $\phi = 1$. therefore, the likelihood equations are

$$\sum_{i=1}^{n}\left(Y_i - \frac{e^{x_i^T\beta}}{1+e^{x_i^T\beta}}\right)x_{i,j} = 0, \quad \forall j = 1, \cdots, p.$$

Remarks

- ☛ No closed form solution in general
- ☛ Efficient approximation alogorithm are used : **Newton Raphson algorithm**

Theorem

Theorem

Under some assumptions, the maximum likelihood estimator

$$\widehat{\beta}_n^{MV} := \arg\max_{\beta} \sum_{i=1}^{n} \frac{Y_i x_i^T \beta - b(x_i^T \beta)}{\phi}$$

is *s.t.*

- $\widehat{\beta}_n^{MV} \xrightarrow{P_{\beta_0}} \beta_0,$
- $\sqrt{n}(\widehat{\beta}_n^{MV} - \beta_0) \xrightarrow{\mathcal{D} \text{ under } P_{\beta_0}} \mathcal{N}(0_p, I^{-1}(\beta_0)).$

estimateur non consistant
→ Bolec

Moreover,

$$I^{1/2}(\widehat{\beta}_n^{MV}) \sqrt{n}(\widehat{\beta}_n^{MV} - \beta_0) \xrightarrow{\mathcal{D}} \mathcal{N}(0_p, I_p).$$

Coefficients nullity test

Wald Test Consider the test

$$H_0 \ : \ \beta_j = 0, \ vs \ H_1 \ : \ \beta_j \neq 0.$$

Under some assumptions and under H_0

$$S := n \left[I(\widehat{\beta}^{MV}) \right]_{jj} \left(\widehat{\beta}_j^{MV} \right)^2 \xrightarrow{\mathcal{D}} \mathcal{X}_1^2.$$

For a fixed $\alpha \in]0, 1[$ fixé, the rejected zone is

$$\left\{ S \geq q_{1-\alpha}^{\mathcal{X}_1^2} \right\},$$

where $q_{1-\alpha}^{\mathcal{X}_1^2}$ is the quantile of order $1 - \alpha$ of a Khi2 distribution with 1 degrees of freedom.

Coefficients nullity test

Note that for categorial variable and under the constraint $\alpha_1 = 0$, the Wald test is different.

$\boxed{\textbf{Wald Test}}$ Considern the test

$$H_0 \ : \ \alpha_{(-1)} = (\alpha_2, \cdots, \alpha_J)^T = \mathbf{0}_{J-1}, \ vs \ H_1 \ : \ \alpha_{(-1)} \neq \mathbf{0}_{J-1}.$$

Under some assumptions and under H_0

$$S := \left\| \sqrt{n}\, \mathrm{I}\left(\widehat{\beta}^{MV}_{(-1)}\right) \widehat{\alpha}^{MV}_{(-1)} \right\|^2 \xrightarrow{\mathcal{D}} \mathcal{X}^2_{J-1}.$$

For a fixed $\alpha \in]0, 1[$ fixé, the rejected zone is

$$\left\{ S \geq q_{1-\alpha}^{\mathcal{X}^2_1} \right\},$$

where $q_{1-\alpha}^{\mathcal{X}^2_{J-1}}$ is the quantile of order $1 - \alpha$ of a Khi2 distribution with $J - 1$ degrees of freedom.

Section 3

3. About the quality of the model

Discussion

- Denote $[m_{sat}]$ the saturated model, *i.e.* when $p \geq n \Rightarrow \widehat{\mathbb{E}[Y_i|x_i]} = Y_i$ (Overfitting).
- $[m_{sat}]$ is the most complex model and all others models are such $[m] \subseteq [m_{sat}]$.

☛ Compare $\mathcal{L}$ the log of the likelihood of our model with $\mathcal{L}_{[m]}$ the log of the likelihood of the saturated model $[m_{sat}]$

tel que $p > n$
toutes les variables
+ leurs transformations

modèle saturé
→ estimation parfaite.
modèle parfait (sur train)

on s'intéresse donc à un modèle plus simple
mais dont $\mathcal{L}(m)$ se rapproche de $\mathcal{L}(m_{sat})$.

Discussion

☛ If $Y_i|x_i \sim \mathcal{B}(p(x_i))$, then for the saturated model $[m_{sat}]$

$$\widehat{\mathbb{E}[Y_i|x_i]} = \widehat{p}(x_i) = Y_i.$$

and the log-likelihood is zero

$$\mathcal{L}_{[m_{sat}]} = \sum_{i=1}^{n} \log \underbrace{\left(\widehat{p}(x_i)^{Y_i}(1-\widehat{p}(x_i))^{1-Y_i}\right)}_{= 1 \text{ si } Y_i = 1 \text{ ou } 0} = 0$$

☛ If $Y_i|x_i \sim \mathcal{B}(n, p(x_i))$, then for the saturated model $[m_{sat}]$

$$\widehat{\mathbb{E}[Y_i|x_i]} = n\widehat{p}(x_i) = Y_i.$$

and the log-likelihood is not zero

$$\mathcal{L}_{[m_{sat}]} = \sum_{i=1}^{n} \log \left(\binom{n}{Y_i} (\widehat{p}(x_i))^{Y_i}(1-\widehat{p}(x_i))^{1-Y_i}\right) \neq 0.$$

$\mathcal{L}_{[m_{sat}]}$ est la plus petite likelihood.

Discussion

- The saturated model is the most complex;
- all others model are such $[m] \subseteq [m_{sat}]$.
- Thus, if a simpler (more parsimonious) model $[m]$ has a $\mathcal{L}_{[m]}$ close to $\mathcal{L}_{[m_{sat}]}$, we will prefer it.

Deviance

Definition

The deviance of a model $[m]$ defined with respect to the saturated model $[m_{sat}]$ is noted $\mathcal{D}_{[m]}$ and is equal to

$$\mathcal{D}_{[m]} = 2\left(\mathcal{L}_{[m_{sat}]} - \mathcal{L}_{[m]}\right) \geq 0,$$

where$\mathcal{L}_{[m_{sat}]}$ and $\mathcal{L}_{[m]}$ are respectively the log likelihoods in the saturated model and in the model $[m]$.

Remark It seems clear that the greater the deviance $\mathcal{D}_{[m]}$, the less the model $[m]$ is good.

Deviance test of two nested models

Proposition

Consider $[m_0]$ and $[m_1]$, 2 nested models ($[m_0] \subset [m_1]$).

$$\begin{cases} H_0 : & [m_0] \text{ is adequat,} \\ H_1 : & [m_1] \text{ is adequat.} \end{cases}$$

Under H_0

$$\Delta\mathcal{D} := (\mathcal{D}_{[m_0]} - \mathcal{D}_{[m_1]}) = 2(\mathcal{L}_{[m_1]} - \mathcal{L}_{[m_0]}) \xrightarrow{\mathcal{D}} \mathcal{X}^2_{m_1 - m_0}.$$

And for $\alpha \in]0, 1[$, a asymptotic test of level α is

$$\left\{ \Delta\mathcal{D} \geq q_{1-\alpha}^{\mathcal{X}^2_{m_1 - m_0}} \right\}.$$

Asymptotic Goodness-of-fit tests

These tests allow to test if a model $[m]$ (with m parameters) is sufficient or not to explain our data:

$$\begin{cases} H_0 : & [m] \text{ is adequate,} \\ H_1 : & [m] \text{ is NOT adequate.} \end{cases}$$

Asymptotic Goodness-of-fit test by deviance

By Deviance Under some assumptions and under H_0

$$\mathcal{D}_{[m]} \xrightarrow{\mathcal{D}} \chi^2_{n-m}.$$

For a fixed $\alpha \in]0, 1[$ fixé, the rejected zone is

$$\left\{ \mathcal{D}_{[m]} \geq q_{1-\alpha}^{\chi^2_{n-m}} \right\},$$

where $q_{1-\alpha}^{\chi^2_{n-m}}$ is the quantile of order $1 - \alpha$ of a Khi2 distribution with $n - m$ degrees of freedom.

Asymptotic Goodness-of-fit test by Pearson

Pearson's generalized $\mathcal{X}^2$ Define the folloxing test statistic

$$\mathcal{X}^2_{\mathcal{P}} = \sum_{i=1}^{n} \frac{(Y_i - \widehat{\mu}_i)^2}{\mathbb{V}\mathrm{ar}(\widehat{\mu}_i)}.$$

Under some assumptions and under H_O

$$\mathcal{X}^2_{\mathcal{P}} \xrightarrow{\mathcal{D}} \mathcal{X}^2_{n-\mathrm{Rank}(X)}.$$

For a fixed $\alpha \in]0,1[$ fixé, the rejected zone is

$$\left\{ \mathcal{X}^2_{\mathcal{P}} > q_{1-\alpha}^{\mathcal{X}^2_{n-\mathrm{Rank}(X)}} \right\}.$$

where $q_{1-\alpha}^{\mathcal{X}^2_{n-\mathrm{Rank}(X)}}$ is the quantile of order $1-\alpha$ of a Khi2 distribution with $n - \mathrm{Rank}(X)$ degrees of freedom.

Pseudo-R^2

- Unlike classical linear regression, the coefficient of determination R^2 does not make sense.
- However, a number of pseudo-R^2 metrics exist.
- Most notable is McFadden's pseudo-R^2.

Pseudo-R^2

$\boxed{\textbf{McFadden's pseudo-}R^2\textbf{.}}$ Let $[m_0]$ be the model resume to the intercept, and $[m]$ the complet model with p parameters. Define:

$$\text{pseudo } R^2_{McF} = \frac{\mathcal{L}_{[m]}}{\mathcal{L}_{[m_0]}} \in [0, 1)$$

- The interpretation remains almost identical to that of the classic one.
- The measure ranges from 0 to just under 1, with values close to zero indicating that the model has no predictive power.

Accuracy and variable selection

- Models are not necessarily nested $\Rightarrow$ deviance test has its limits.
- Other criteria make it possible to compare models which are not necessarily nested within each other (AIC, BIC, ...) coupled to the models selection methods seen previously (bakward, forward, ...).

Residuals analysis

- Due to the nature of the response variable Y, the classical analysis of residuals as a function of predicted values or the notion of heteroskedasticity must be redefined.
- In the linear setting, the residuals are as for the linear case defined as the difference between the observed values Y_i and the predicted values $\widehat{Y_i}$.
- Here, the residuals are defined as the difference between the observed values Y_i and the predicted values $\widehat{\mu_i} = g^{-1}(x_i^T \widehat{\beta})$:

$$\widehat{\epsilon_i} = y_i - \widehat{\mu_i}.$$

Standardized Pearson residuals

The standardized Pearson residuals r_{s_i} are obtained by renormalizing the residuals $\widehat{\epsilon_i}$ by the estimated variance of Y_i, $\widehat{\mathbb{V}\mathrm{ar}(y_i)}$

Example Logistic setting :

$$\widehat{\mathbb{V}\mathrm{ar}(y_i)} = \widehat{p}(x_i)(1 - \widehat{p}(x_i)).$$

In addition, it is also necessary to renormalize by the leverage effect

$$r_{s_i} = \frac{\widehat{\epsilon_i}}{\sqrt{(1 - \widehat{h_{ii})\mathbb{V}}\mathrm{ar}(y_i)}},$$

where h_{ii} is the i^{eme} diagonal element of the projection matrix $H = X(X^T X)^{-1} X^T$ in the **full rank** setting of the matrix X.

Standardized deviance residuals

Standardized deviance residuals

Let us introduce residuals adapted to generalized models. Let $\mathcal{L}_{[m]}(\beta, Y)$ and $\mathcal{L}_{[m_{sat}]}(\beta, Y)$ respectively be the log of the likelihood in the model $[m]$ and the saturated model $[m_{sat}]$.

Let $\widehat{\beta}$ and $\widehat{\beta}_{sat}$ be the maximum likelihood estimators calculated respectively in the models $[m]$ and $[m_{sat}]$.

The standardized deviance residuals measure how far $\mathcal{L}_{[m]}(\widehat{\beta}, y)$ for the i observation is from $\mathcal{L}_{[m_{sat}]}(\widehat{\beta}_S, y)$ for this same observation, all renormalized through the leverage effect. Thereby

$$r_{d_i} = \mathrm{sign}(y_i - \widehat{\mu}_i)\sqrt{\frac{2\left(\mathcal{L}_{[m_{sat}]}(\widehat{\beta}_S, y) - \mathcal{L}_{[m]}(\widehat{\beta}, y)\right)}{(1 - h_{ii})}}.$$

Remarks

☛ The standardized deviance residuals measure the deviance.

☛ The deviance of a model $[m]$ defined with respect to the saturated model $[m_{sat}]$ is

$$\mathcal{D}_{[m]} = 2\left(\mathcal{L}_{[m_{sat}]} - \mathcal{L}_{[m]}\right) \geq 0,$$

where $\mathcal{L}_{[m_{sat}]}$ and $\mathcal{L}_{[m]}$ are respectively the log likelihoods in the saturated model and in the model $[m]$.

Interpretation

- As in the linear setting, we can show that the residuals are asymptotically Gaussian (to be verified by a Q-Q-plot).
- It will be necessary to check that there is no structure or trend, in this case, it will be necessary to identify the cause (bad model, particular / quadratic structure of a variable, . . .).