Generalized nonparametric regression by local likelihood

Pedro Delicado
Departament d'Estadística i Investigació Operativa
Universitat Politècnica de Catalunya

1 Introduction

2 Nonparametric regression with binary response
   Logistic regression model
   Nonparametric binary regression

3 Generalized nonparametric regression model

4 Estimation by maximum local likelihood

5 Bandwidth choice in the generalized nonparametric regression model

Generalized nonparametric regression model

- Nonparametric version of the Generalized Linear Model (GLM).
- Different types of response variable Y are allowed:

  binary, count variable, non-negative, with values in $[0, 1]$, ...
- One explanatory variable (only one for the moment): X.
- The conditional distribution of $(Y|X = x)$ is in a parametric (exponential) family.
- One of the parameters (or a transformation of it) of the conditional distribution of $(Y|X = x)$ is a smooth function of x.

Regression with binary response. Logistic regression

- To fix ideas, let us remember the logistic regression model.
- Let (X, Y) be two random variable with Y a binary variable and X a continuous variable.
- We assume that the conditional distribution of Y, given $X = x$, is

$$(Y|X = x) \sim \text{Bernoulli}(p(x))$$

and that there exist parameters β_0 and β_1 such that

$$\log\left(\frac{p(x)}{1 - p(x)}\right) = \beta_0 + \beta_1 x$$

or equivalently

$$p(x) = \frac{e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+\beta_1 x}}.$$

- The link function $log(p/(1 - p))$ is the logistic function.

Logistic regression model estimation

- Let (x_i, y_i), $i = 1, \dots, n$, n independent observations of two random variables (X, Y) following the logistic regression model.
- The estimation of parameters β_0 and β_1 is done by the maximization of the log-likelihood function,

$$\ell(\beta_0, \beta_1) = \sum_{i=1}^{n} \left(y_i \log\left(\frac{p_i}{1-p_i}\right) + \log(1-p_i) \right),$$

  where $p_i = p(x_i)$.
- The maximization of $\ell(\beta_0, \beta_1)$ is done by numerical methods. The most used algorithm (equivalent to the Newton-Raphson algorithm) is known as IRWLS: iteratively re-weighted least squares.
- This algorithm is used also for fitting other GLMs.

Iteratively re-weighted least squares algorithm (IRWLS) for logistic regression.

- Choose starting values $\beta^0 = (\beta_0^0, \beta_1^0)$ (the choice $\beta_0^0 = \beta_1^0 = 0$ is usually appropriate).
- Set $s = 0$ and iterate the following steps until convergence.
  1. Set
  $$p_i^s = \frac{e^{\beta_0^s + \beta_1^s x_i}}{1 + e^{\beta_0^s + \beta_1^s x_i}},$$
  $$z_i^s = \beta_0^s + \beta_1^s x_i + \frac{y_i - p_i^s}{p_i^s(1 - p_i^s)},\ i = 1, \ldots, n.$$
  2. Let $(\nu_1^s, \ldots, \nu_n^s)$ be the weight vector with $\nu_i^s = p_i^s(1 - p_i^s)$.
  3. Fit the linear regression with responses z_i^s and explanatory variable values x_i, (plus the constant term) by weighted least squares using the weights ν_i^s, $i = 1, \ldots, n$.
  Let $\beta^{s+1} = (\beta_0^{s+1}, \beta_1^{s+1})$ be the estimated regression coefficients.
  4. Set $s = s + 1$ and go back to the step 1.

Nonparametric binary regression

- It is the nonparametric version of logistic regression.
- The bivariate random variable (X, Y) has joint distribution such that

$$(Y|X = x) \sim \text{Bernoulli}(p(x))$$

with

$$\log\left(\frac{p(x)}{1-p(x)}\right) = \theta(x)$$

where $\theta(x)$ is a smooth function of x.
- Equivalently,

$$p(x) = \frac{e^{\theta(x)}}{1+e^{\theta(x)}}.$$

- The logistic link, $\log(p/(1-p))$, is used also in this context.
- Function $\theta(x)$ is free of constrains (it is not the case for function $p(x)$, that must be in $[0, 1]$).
- Observe that $p(x) = E(Y|X = x)$ is the regression function.

Generalized nonparametric regression model

- The bivariate random variable (X, Y) has joint distribution such that

$$(Y|X = x) \sim f(y; m(x), \psi)$$

where $m(x) = E(Y|X = x)$ is a smooth function of x, possibly subjected to certain constrains (non-negativity or boundedness, for instance), and ψ represents other parameters (variance, for instance) not depending on x.

- There exists an invertible link function $g(\cdot)$ such that

$$\theta(x) = g(m(x)), \;\; m(x) = g^{-1}(\theta(x))$$

where $\theta(x)$ is a smooth function of x free of constrains ($\theta(x)$ can take any real value).

- Alternatively, $(Y|X = x) \sim f_2(y; \theta(x), \psi) = f(y; g^{-1}(\theta(x)), \psi)$.

Estimation by maximum local likelihood (Loader 1999)

- We focus on the nonparametric binary response model:

$$(Y|X=x) \sim \text{Bernoulli}(p(x)),\ \log\left(\frac{p(x)}{1-p(x)}\right) = \theta(x)$$

- n data from this model have been observed, $(y_i; x_i)$, $i = 1, \ldots, n$, and the goal is to estimate $p(t) = E(Y|X=t)$ for a generic value $t \in \mathbb{R}$.
- Given that $\theta(x)$ is a smooth function, a first order Taylor expansion of $\theta(x)$ around t gives that for x close to t we have that

$$\theta(x) \approx \theta(t) + \theta'(t)(x-t).$$

- Then, in a neighborhood of t the standard logistic model is approximately valid:

$$\theta(x) \approx \beta_0^t + \beta_1^t(x-t),$$

- We fit this logistic model, $\theta(x) \approx \beta_0^t + \beta_1^t(x-t)$, by maximum *local* likelihood.
- The contribution to the log-likelihood function of each observation is

$$y_i \log\left(\frac{p_i^t}{1-p_i^t}\right) + \log(1-p_i^t),$$

where

$$p_i^t = \frac{e^{\beta_0^t+\beta_1^t x_i}}{1+e^{\beta_0^t+\beta_1^t x_i}} \approx p(x_i).$$

- The closer x_i is to t, the better is the approximation $p_i^t \approx p(x_i)$.
- Adding up all the data contributions, weighted by $w_i^t \propto K((t-x_i)/h)$, the local log-likelihood function is obtained:

$$\ell_t(\beta_0^t, \beta_1^t) = \sum_{i=1}^n w_i^t \left(y_i \log\left(\frac{p_i^t}{1-p_i^t}\right) + \log(1-p_i^t) \right).$$

Maximizing the local log-likelihood function

- The local logistic model is estimated by a weighted version of the IRWLS algorithm.
- The standard IRWLS algorithm is modified multiplying at each iteration the usual weights $p_i^s(1 - p_i^s)$ by the kernel weights w_i^t.

IRWLS algorithm for maximizing the local log-likelihood.

- Choose starting values $\beta^0 = (\beta_0^0, \beta_1^0)$ (the choice $\beta_0^0 = \beta_1^0 = 0$ is usually appropriate).
- Set $s = 0$ and iterate the following steps until convergence.
  1. Set
  $$p_i^s = \frac{e^{\beta_0^s + \beta_1^s x_i}}{1 + e^{\beta_0^s + \beta_1^s x_i}},$$
  $$z_i^s = \beta_0^s + \beta_1^s x_i + \frac{y_i - p_i^s}{p_i^s(1 - p_i^s)}, \; i = 1, \ldots, n.$$
  2. Let $(\nu_1^s, \ldots, \nu_n^s)$ be the weight vector with $\nu_i^s = p_i^s(1 - p_i^s)w_i^t$.
  3. Fit the linear regression with responses z_i^s and explanatory variable values x_i, (plus the constant term) by weighted least squares using the weights ν_i^s, $i = 1, \ldots, n$.
  Let $\beta^{s+1} = (\beta_0^{s+1}, \beta_1^{s+1})$ be the estimated regression coefficients.
  4. Set $s = s + 1$ and go back to the step 1.

- Once the estimates $\hat{\beta}_0^t$ and $\hat{\beta}_1^t$ have been obtained, the function $\theta(t)$ is estimated as

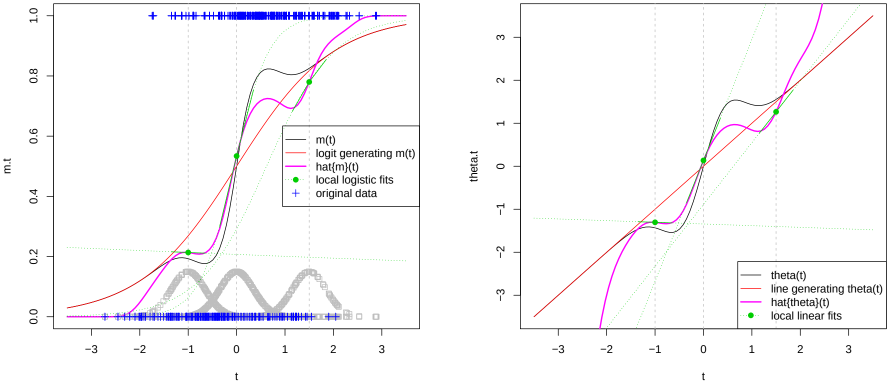
$$\hat{\theta}(t) = \hat{\beta}_0^t + \hat{\beta}_1^t(t - t) = \hat{\beta}_0^t$$

and $p(t)$ as

$$\hat{p}(t) = \frac{e^{\hat{\theta}(t)}}{1 + e^{\hat{\theta}(t)}}.$$

- The logistic model has been used as the parametric model for the local approximation, but other parametric regression models for binary response can be used instead.

Example. Local logistic regression, artificial data



Practice:

- Write your own local logistic regression function.
  `03_my_own_local_glm.Rmd`
- Illustrating the local logistic regression.
  `03_local.logistic.R`
  `03_Gener_NPRM_country_data.Rmd`, points 1, 2.
- Generalized nonparametric regression model.
  `03_Gener_NPRM_country_data.Rmd`, point 3.

Bandwidth choice when fitting local logistic regression

- We present two possibilities (there exist other alternatives):
  - Minimizing in h the probability of misclassification of a new observation.
  - Maximizing in h the expected log-likelihood of a new observation.
- Both quantities must be estimated, for instance, by cross-validation.

Minimizing in h the probability of misclassification of a new observation

This probability is estimated by leave-one-out cross-validation:

$$p_{CV}(h) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{y_i \neq \hat{y}_i^{(-i)}\}},$$

where $\hat{y}_i^{(-i)} = \mathbb{1}_{\{p_i^{(-i)} \geq 0.5\}}$, and $p_i^{(-i)} = \hat{p}_h^{(-i)}(x_i)$ is the estimation of $p(x_i) = \Pr(Y = 1 | X = x_i)$, when using h as bandwidth and all the observations, except the i-th.

Maximizing in h the expected log-likelihood of a new observation

- This value is estimated by leave-one-out cross-validation.
- The cross-validation estimation of the expected log-likelihood of a new observation, when using h as bandwidth, is

$$\ell_{CV}(h) = \frac{1}{n}\sum_{i=1}^{n} \log\left(\widehat{\Pr}_h^{(-i)}(Y = y_i|X = x_i)\right)$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left(y_i \log\left(\frac{p_i^{(-i)}}{1 - p_i^{(-i)}}\right) + \log(1 - p_i^{(-i)})\right),$$

where $p_i^{(-i)} = \hat{p}_h^{(-i)}(x_i)$ is the estimation of $p(x_i) = \Pr(Y = 1|X = x_i)$, when using h as bandwidth and all the observations, except the i-th.

Practice:

- Bandwidth choice when fitting local logistic regression.

  `03_Band_Choice_Local_Logistic.Rmd`

Bowman, A. W. and A. Azzalini (1997).
Applied Smoothing Techniques for Data Analysis.
Oxford: Oxford University Press.

Hastie, T., R. Tibshirani, and J. Friedman (2009).
The elements of statistical learning (2nd ed.).
Springer.

Loader, C. (1999).
Local regression and likelihood.
New York: Springer.

Wasserman, L. (2006).
All of Nonparametric Statistics.
New York: Springer.