ECOLE POLYTECHNIQUE DE BRUXELLES

Computational statistics - Homework 1

Model selection in GLM : logistic regression

Matteo Silvestri

Professor Maarten Jansen

2024

Table des matières

1 Problem

Let's consider the Data Generating Process

$$Y_i \sim \text{Bernoulli}(p_i) \text{ with } p_i = \frac{1}{2}[\cos(5x_i^2) + 1] \text{ for } n = 100 \text{ covariates } x_i \in [0,1]$$

where x_i is uniformly distributed on the interval $[0,1]$.

Select and estimate a GLM of the form $g(\boldsymbol{p}) = \boldsymbol{\theta} = \boldsymbol{X}\boldsymbol{\beta}$ where the design matrix $\boldsymbol{X}$ is given by

$$X_{ik} = x_i^k \text{ for } k = 0, 1, ..., m$$

and using the logit link function $g(u) = \log\left(\frac{u}{1-u}\right)$.

Use AIC and BIC on nested submodels $S = \{1, 2, ..., p\}$ with $p \leq m$ to select the appropriate polynomial degree. Let n grow towards infinty and see what happens. What can you say about the consistency of BIC?

2 Solution

Let's consider the case $n = 100$.

As can be seen from figures 2.1,2.2 , both criteria show how the linear part is fundamental to the model. Moreover, AIC gives as optimal dimension 6 while BIC gives 3. It's fair to note that there are some numerical issues, as some points in the graphs are missing. This could be related to the fact that we are working with powers of numbers in $[0,1]$ very close to 0, and therefore many columns of $\boldsymbol{X}$ are collinear. Since the task is to determine the true underlying model, BIC is more appropriate. Thus, the appropriate polynomial degree is $p = 2$. This is confirmed by the figure 2.3, which shows that the model can be described quite well by quadratic polynomials and also shows the tendency of AIC to overestimate the number of parameters in the model causing overfitting.

Now let's consider the case $n = 1000$.

As can be seen from figures 2.4,2.5 , there are no more numerical issues. In this case, the quadratic part is essential to the model. In particular, AIC gives 9 as the optimal dimension while BIC gives 5. However, it is clear that even AIC shows that polynomials of degree $p = 4$ are sufficient to estimate the model; also, the 9-model is not overfitting much (figure 2.6).

The selection consistency of BIC in this case is not effective, as the true model for the DGP is not among the proposed models.

Github repository

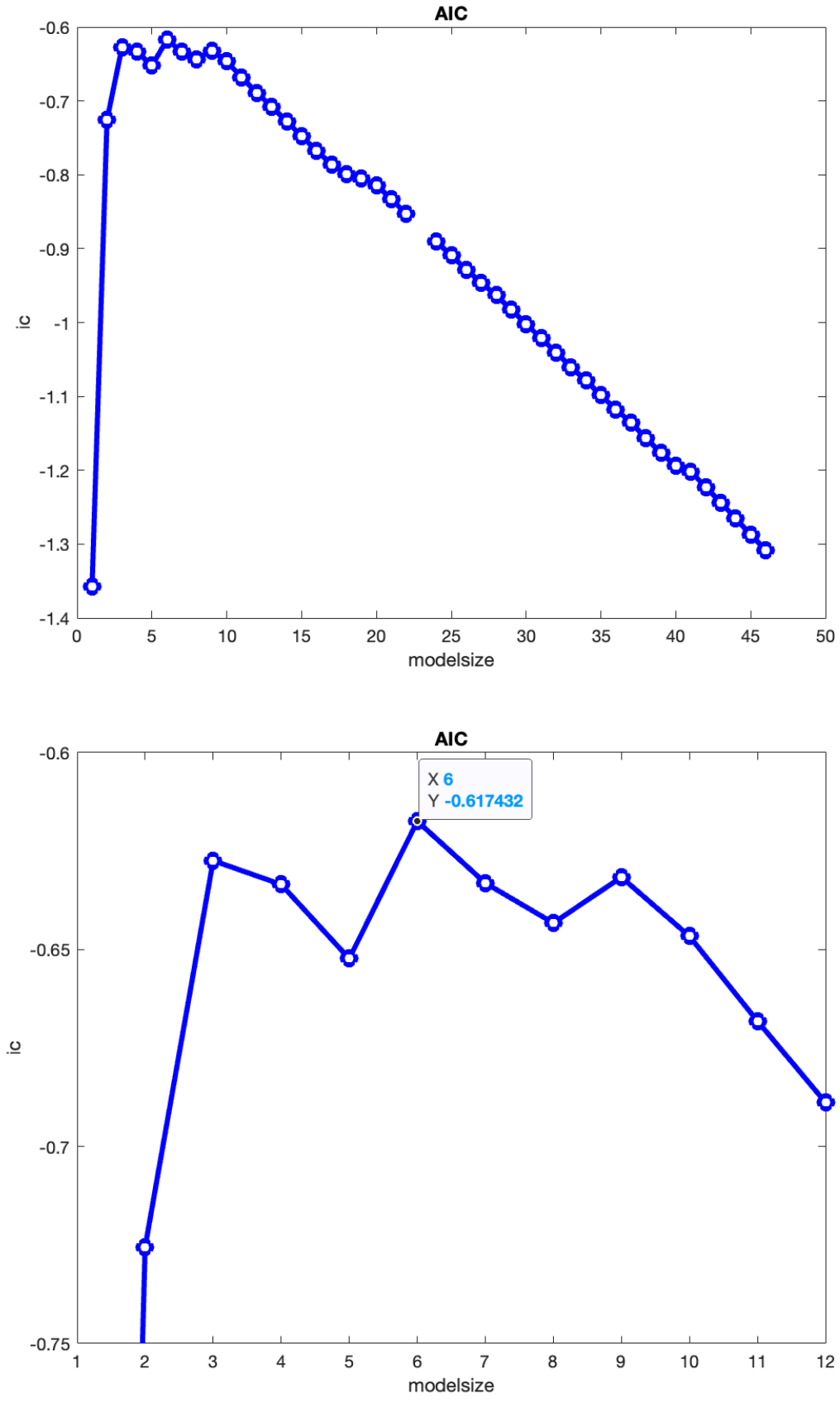


FIGURE 2.1 – Representation of AIC values, zooming on the optimal model size ($n = 100$)

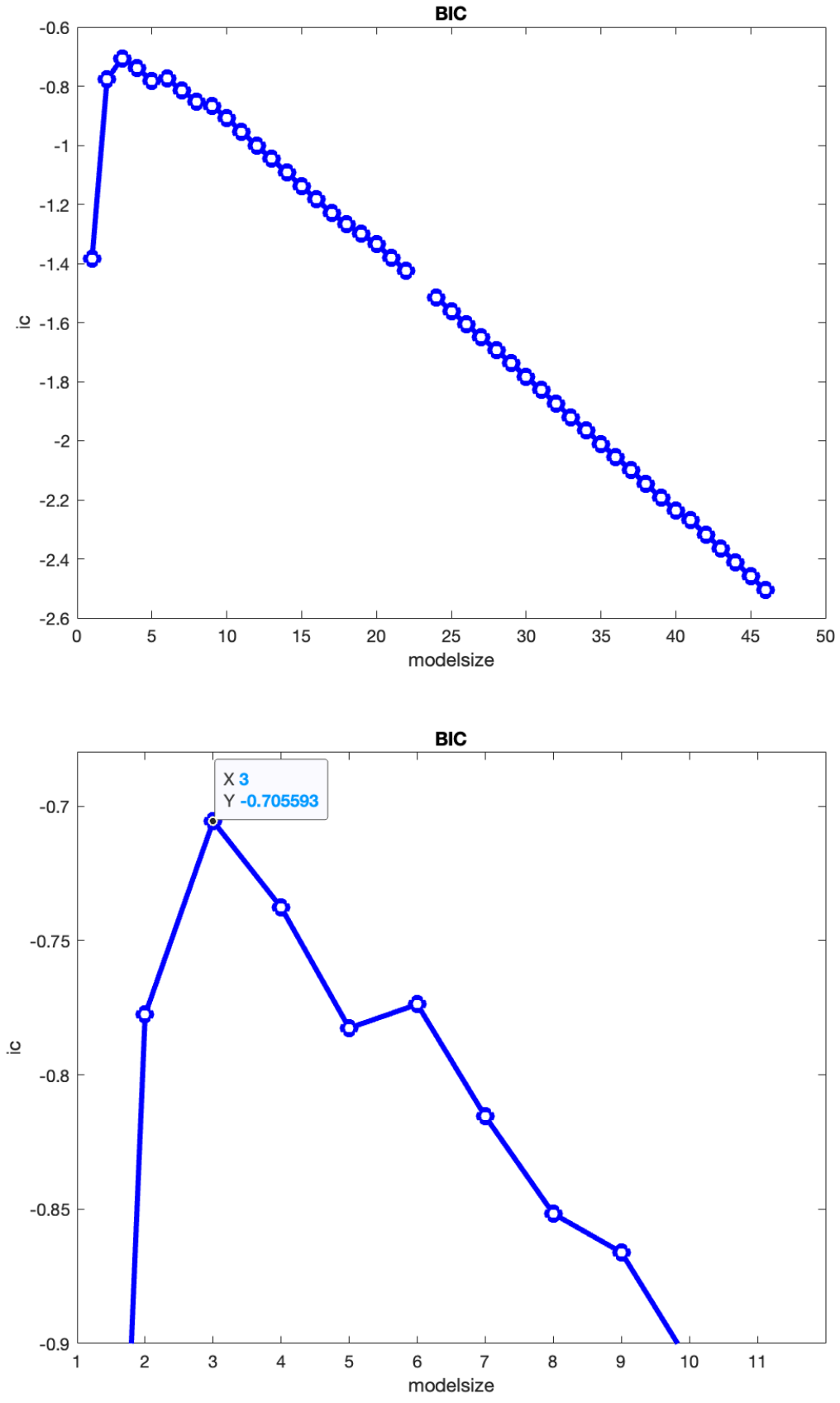
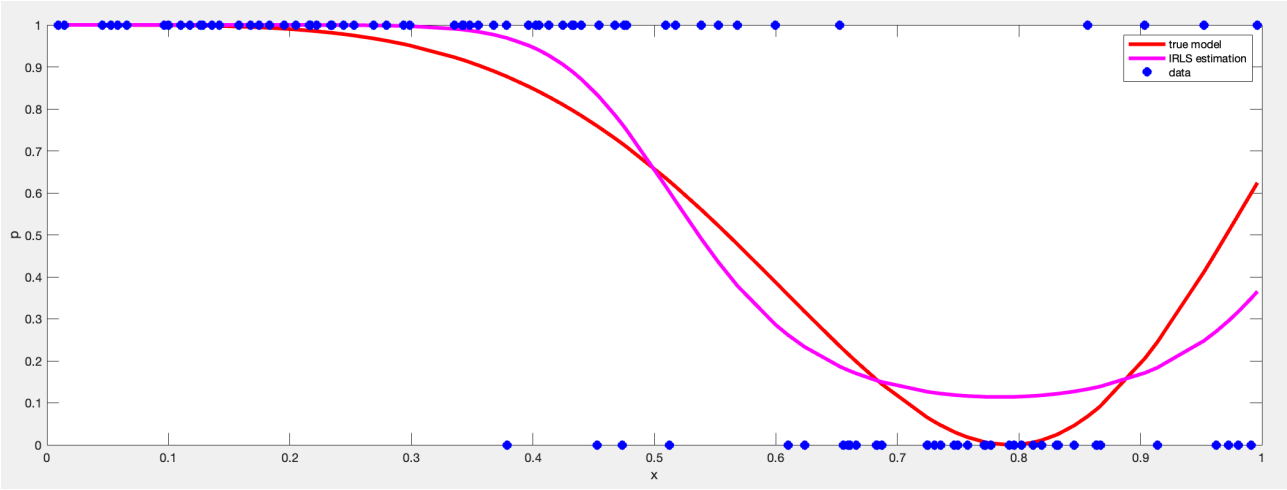
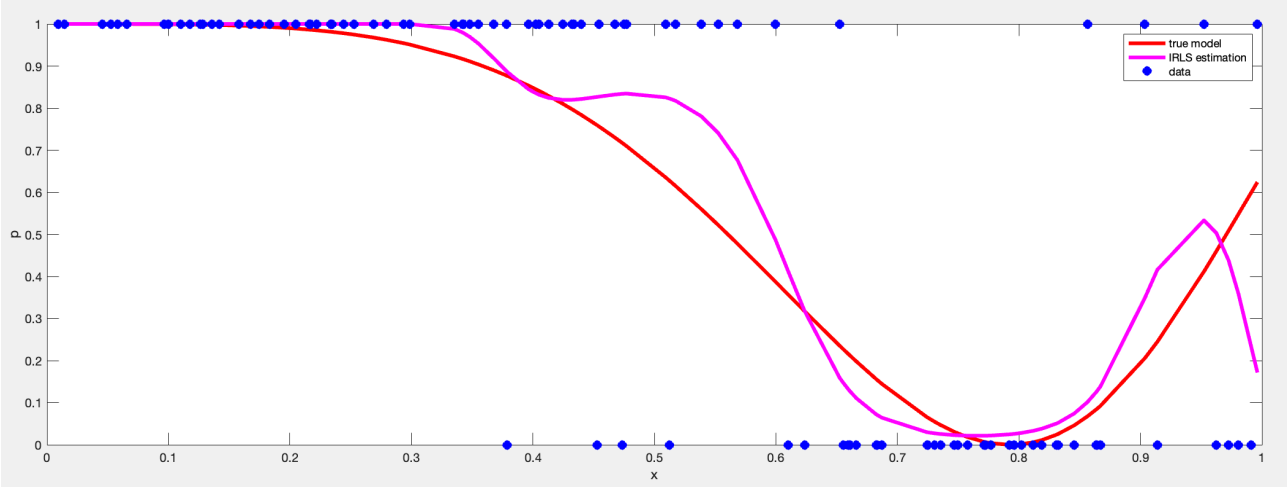


FIGURE 2.2 – Representation of BIC values, zooming on the optimal model size ($n = 100$)



(a) BIC model size



(b) AIC model size

FIGURE 2.3 – Estimation using IRLS with suggested model size and true model ($n = 100$)

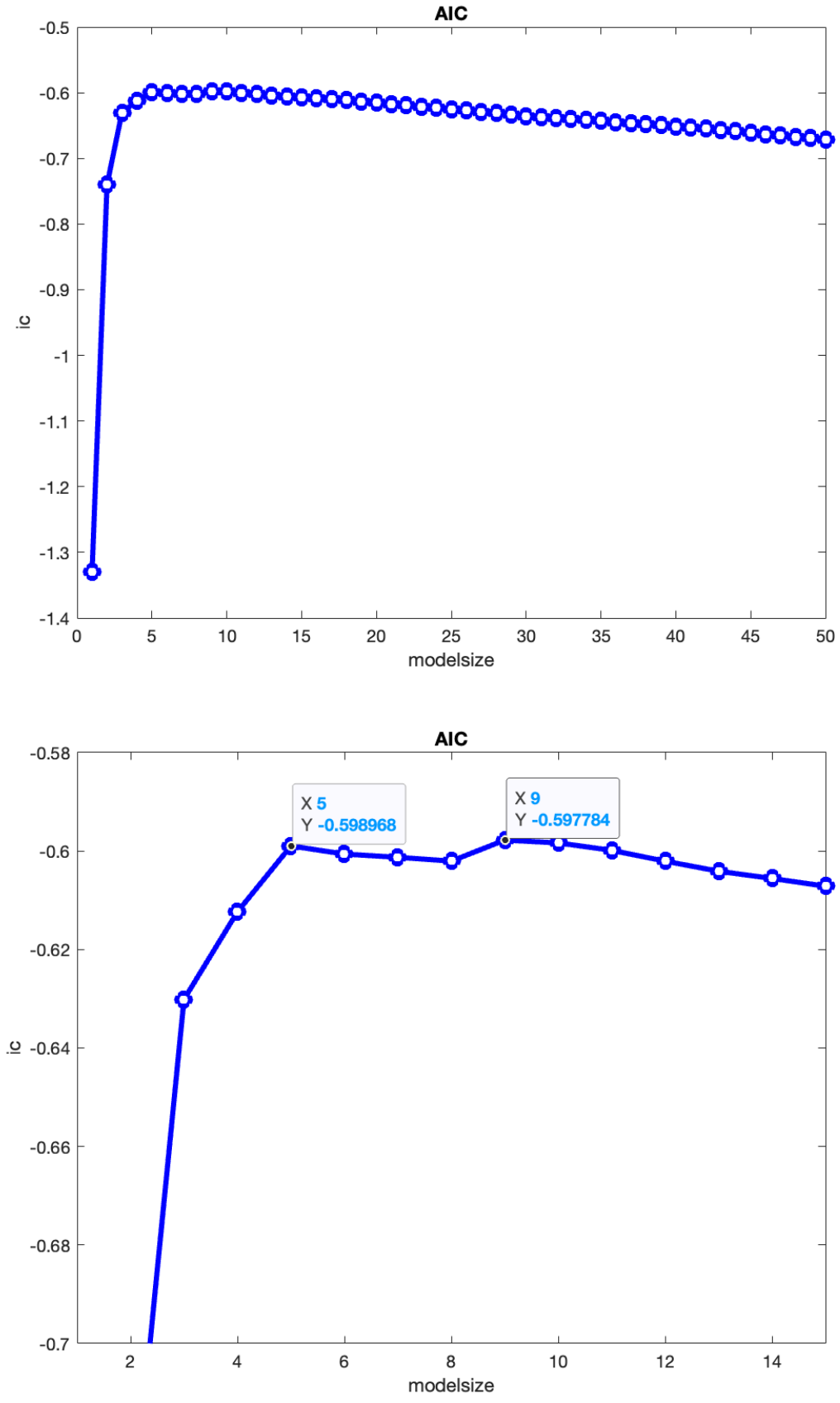


FIGURE 2.4 – Representation of AIC values, zooming on the optimal model size ($n = 1000$)

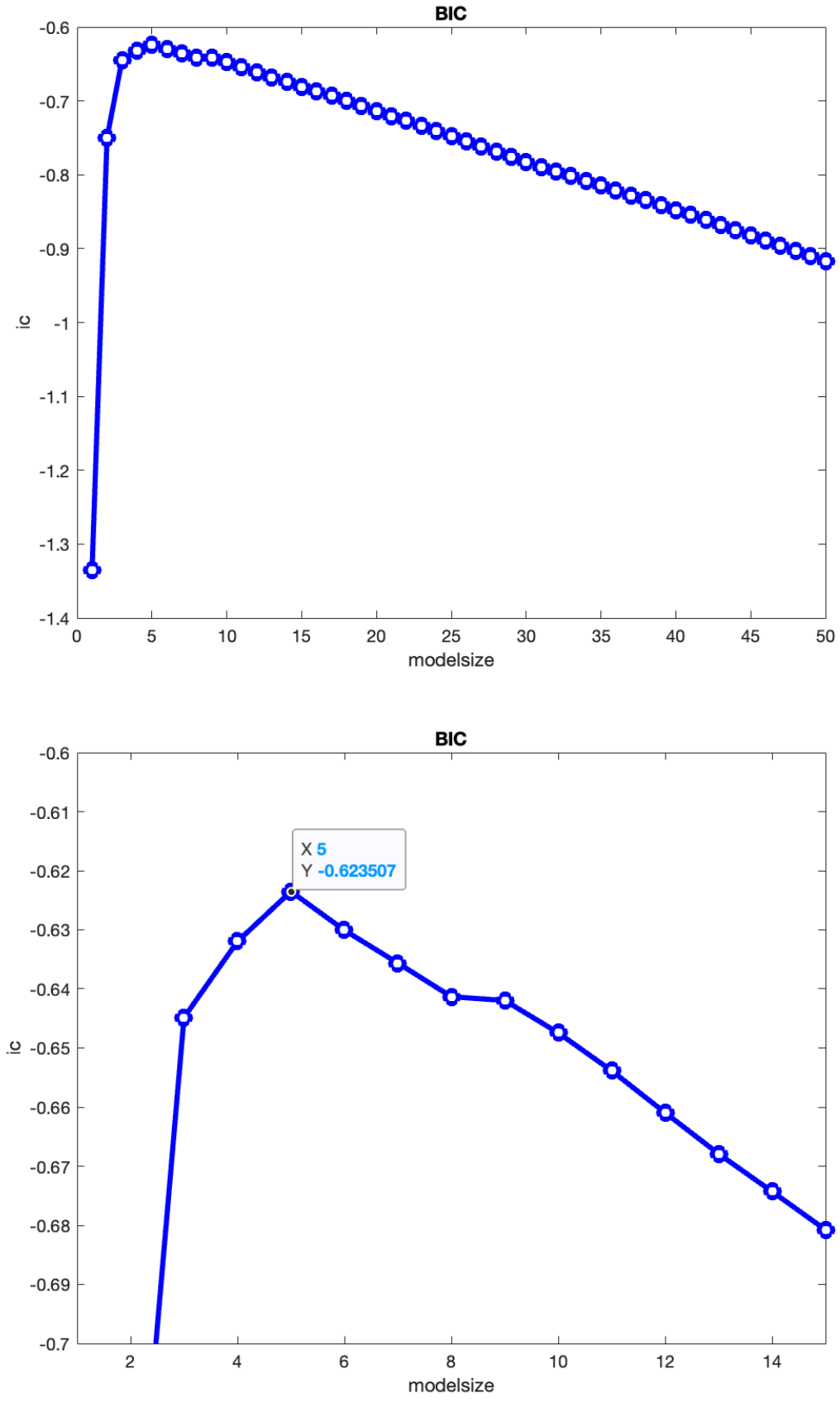
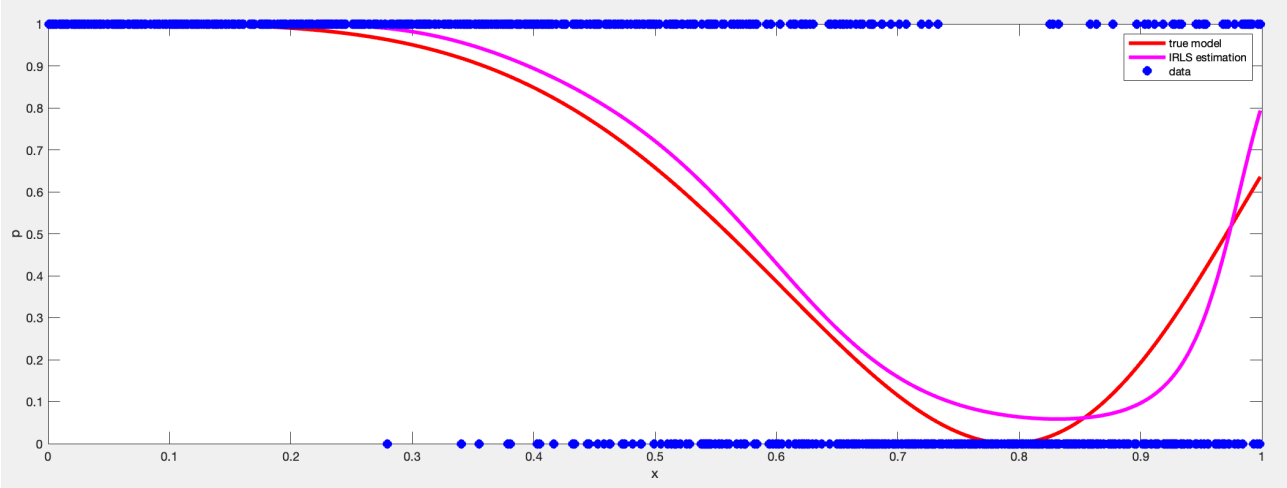
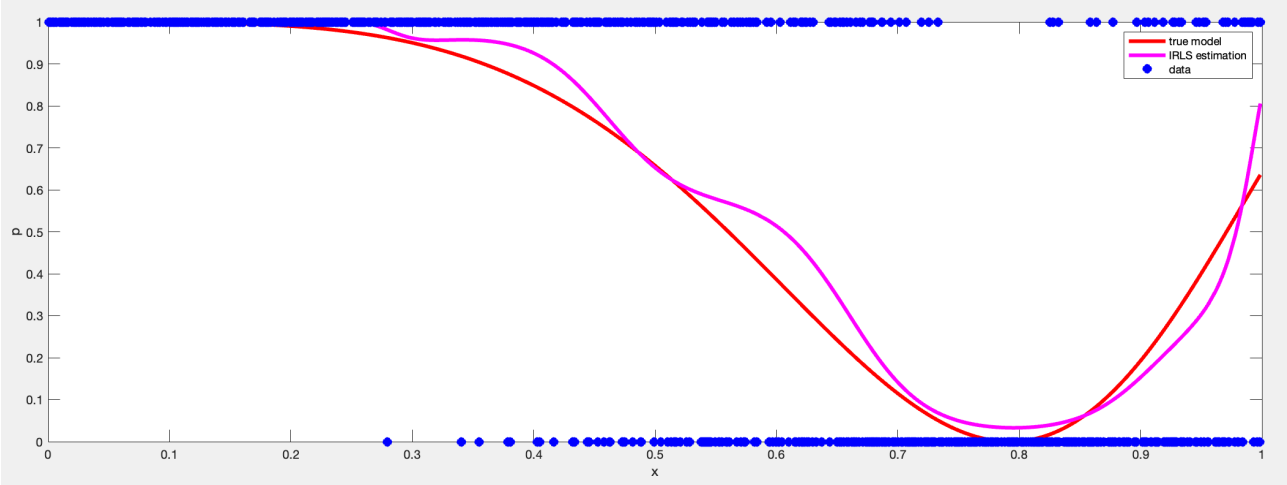


FIGURE 2.5 – Representation of BIC values, zooming on the optimal model size ($n = 1000$)



(a) BIC model size



(b) AIC model size

FIGURE 2.6 – Estimation using IRLS with suggested model size and true model ($n = 1000$)