EC3389 - Homework 5

Due Monday, April 17th

1 Programming

> This time we have decided to give an slightly open question for you to think about. This will prepare you for the final project, which will (probably) feature a similar problem.

Data In the zip file accompanying this homework, you will find a csv file containing the regressors and regressand to be used. The number of observations is 5000, and the number of regressors is 3000. The regressand is binary.

You are supposed to estimate the conditional expectaion of Y, given by a generalized nonlinear model:

$$E[Y|X] = f(X^T\beta)$$

However, there is a catch: most of the regressors are just noise, and they are not related to the regressand in any way. The number of relevant regressors is not certain[1], but it is known to be "small". Moreover, it is known that there are no higher-order terms (e.g.., there are no squared terms X_m^2) or interaction terms (e.g., no cross-terms X_mX_n) in the true model.

Problem Your task is to:

1. Discover the relevant regressors.
2. Estimate their associated coefficients as accurately as possible.
3. Provide a sensible measure of your estimator's performance.

Hopefully, this you have you use all techniques you have learned in the course so far, and have you explore some new tools as well. There is no need to document every step of your work. Extra points will be given for succint answers.

[1] By which we mean it is not certain to you. Since we have simulated the data ourselves, we know the answer and will be able to evaluate your work.

Python tips Since the regressand is discrete, you will have to use slightly different tools as we have used until now. Make sure to check out the the scikit-learn documentation[2] for useful tools.

2 Theory

- Textbook exercise: 7.9.2.
- Textbook exercise: 4.7.1.
- Textbook exercise: 4.7.6.

[2] `http://scikit-learn.org/stable/modules/linear_model.html`