Advanced Quantitative Methods

Course for the master programme in peace and conflict studies,
Uppsala University March–May 2019

Håvard Hegre, Mihai Croicu and David Randahl

April 6, 2019

UPPSALA UNIVERSITET

Assignment 2: Out-of-sample evaluation I

Overfitting can be a severe problem when estimating models. This is particularly true when the dependent variable is dichotomous and data are scarce. In this assignment we want you to design a Monte Carlo experiment to assess the extent of the overfitting problem and explore whether out-of-sample evaluation of predictive performance can alleviate it. In the motivation section, briefly discuss over-fitting and out-of-sample evaluation.

1. Assume that the true data-generating process is

$$ln(\frac{p}{1-p}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

where $\beta_0 = -6, \beta_1 = -6, \beta_2 = 1$, and X_1 and X_2 are uncorrelated and drawn from a uniform distribution with range $[-2, 2]$
In `R`, this DGP is specified as: `Y <- rbinom(n, 1, inv.logit(b0 + b1*X1 + b2*X2))`

2. Draw a dataset based on this DGP with $N = 400$ observations
3. Also draw 8 additional X variables with the same distribution as X_1 and X_2.
4. Present descriptive statistics for all variables in the model

UPPSALA UNIVERSITET

Assignment 2: Out-of-sample evaluation II

5. Estimate a logit model with Y as the dependent variable and all 10 X variables as predictors. Comment on the results. Identify the three variables in the model with lowest p-values
6. Split the data into two partitions (e.g. 300 in-sample, 100 out-sample), or do a 4-fold cross-validation. Estimate the original model on the in-sample plus three models where you remove each of the three most significant variables one at the time (i.e., these three models will have 9 X variables). Generate predictions for the out-sample.
7. Compare the out-of-sample predictive performance of the four models using area under the ROC and PR curves. Discuss any differences in conclusions drawn from the p-values and from the out-of-sample evaluation
8. Generalize the experiment by repeating it 1000 times in a loop. Identify the repetitions where X1 is significant at the 5% level, report how often this happened, and calculate the average difference i AUROC when X1 is removed from the model. Then do the same for X3. Discuss.

UPPSALA UNIVERSITET

Bibliography I