Elements of Statistics and Econometrics

Assignment 4

Problem 6: Regression techniques

In this problem we use the same data set on service usage of 1000 clients of a telecommunication company as in Assignment 3. The variable `tenure` is taken as the dependent variable and the remaining variables as explanatory.

1. The lasso regression is an alternative approach to variable selection.

   (a) Explain in your own words the idea of the lasso regression. Sketch a situation when a simple linear regression fails, but the lasso regression still can be estimated.

   (b) For the usual regression model the variables are rarely normalized/standardized. However, in the case of the lasso regression the scaling becomes crucial. Why? Scale your data by $(x_i - \bar{x})/\hat{\sigma}_x$. Can/should the binary variables be scaled in the same fashion? How would you handle the variable `ed`?

   (c) Run a lasso regression for data with $\alpha \in (0, 1)$. Plot the estimated parameters as functions of α. Which value of α would you recommend? If it is easy to implement, then determine the optimal λ by cross-validation.

2. A nonlinear regression offers a flexible technique for modelling complex relationships. We wish to explain the `tenure` by the long distance calls per month `longmon`.

   (a) Make a bivariate scatter plot and estimate an appropriate linear (!) model. Add the regression line to the plot.

   (b) Estimate now an appropriate nonlinear regression which might fit the data better. Add the regression curve to the plot and compare (quantitatively) the fit with the fit of the linear model.

   (c) Explain in your own words, why all the classical tests and inferences are not directly applicable to the NLS estimators.

   (d) What kind of problems might arise if we decide to fit a non-linear regression using all explanatory variables?

3. Next we model the relationship between `tenure` and `address` using the nonparametric Nadaraya-Watson regression.

   (a) An important calibration parameter of a nonparametric regression is the bandwidth. Explain what happens with the regression/the weights in the Nadaraya-Watson regression if the bandwidth is too high or too small.

(b) Fit a Nadaraya-Watson regression with Gaussian kernel and "optimal" bandwidth to the longmon/address data. Check and explain how the "optimal bandwidth" is determined in your software. Compare the (in-sample) fit of the nonparametric regression and the nonlinear regression in the previous subproblem.

4. Next we consider classification of the clients using the `churn` variable as the dependent variable and the logistic regression.

   (a) Fit a logistic regression to explain `churn` by the remaining explanatory variables.

   (b) Consider the explanatory variable `tenure`. Obviously its parameter cannot be interpreted in the same way as for a linear regression. Provide the correct interpretation using the parameter and using odds.

   (c) Run a stepwise model selection using AIC as criterion. Further consider only the optimal model chosen here. From the final model, which of the variables do increase the probability of churn and which variables decrease this probability? Is this consistent with economic intuition?

   (d) Randomly pick up five clients. Determine their probabilities of leaving the company. Provide for the first of them the formula which may be used to compute this probability with inserted values of parameters and variables. If you want to predict the membership in one of the two groups for a particular client, what is the simplest way to proceed using these probabilities?

   (e) Compute the classification table and calculate the specificity and sensitivity. Provide verbal interpretation for the elements of the classification table and the performance measures.

   (f) To improve the performance it makes sense to change the threshold used for classification. This can be done using the ROC curve. Plot this curve and determine the optimal threshold.

   (g) Recompute the classification table, sensitivity and specificity for the new threshold. Provide interpretation of the obtained values. Compare the results with the original values. Is the procedure now more strict/conservative?

5. In the next step we model `tenure` using regression trees.

   (a) Assume the first variable to be used for splitting is `longmon`. Write down the corresponding optimization problem and explain how the optimization works.

   (b) Obviously you can get very long trees. Tree prunning helps to get trees of a reasonable size. Fit a CART to the data and prune it to have at most 10 splits. What is the value of the corresponding complexity parameter? Check you software for the implementation of the prunning, particularly the form of the loss function.

   (c) Check the value of the improvement in the first split. Explain the idea of improvement and provide numerical expression how this improvement is computed for the first split.

   (d) Compare the in-sample fit of the tree to the in-sample fit of the lasso regression above. Compare the importance of the variables for the two models. In the case of the lasso regression the importance is mirrored by the parameters if the variables are standardized.