Elements of Statistics, Econometrics and Time Series Analysis (2017)

Assignment 3

Problem 5: Regression techniques

In this problem we will apply several alternative modeling techniques to the CEO salary data which was used in the lectures. The data set consists of the following variables:

```
salary  = 1999 salary + bonuses
totcomp = 1999 CEO total compensation
tenure  = # of years as CEO (=0 if less than 6 months)
age     = age of CEO
sales   = total 1998 sales revenue of firm i
profits = 1998 profits for firm i
assets  = total assets of firm i in 1998
```

The `salary` 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?

   (c) Run a lasso regression for scaled $((x_i - \bar{x})/\hat{\sigma}_x)$ data with $\alpha \in (0, 1)$. Plot the estimated parameters as functions of α. Which value of α would you recommend?

2. A nonlinear regression offers a flexible technique for modelling complex relationships. We wish to explain the profits of the companies using their sales. Take logarithms of the both variables and omit negative values if any.

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

3. Next we model the relationship between log-profits and log-sales using a nonparametric 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 profits/sales data. Check and explain how the "optimal bandwidth" is determined in your software. Plot the data and the regression curve.

   (c) Compare the fit of the nonparametric regression and the nonlinear regression in the previous subproblem.

4. We manually classify all CEO into those who earn more than 2000 and those who earn less than 2000. This implies that we can use the `salary` variable to group the students into two categories:

$$Y_i = \begin{cases} 1, & \text{if } \texttt{salary}_i > 2000 \\ 0, & \text{if } \texttt{salary}_i \leq 2000 \end{cases}.$$

   Next we consider the logistic regression and use Y as the dependent variable.

   (a) Fit a logistic regression to explain Y by the remaining explanatory variables. Run a stepwise model selection using AIC as criterion. Further consider only the optimal model chosen here.

   (b) Consider the explanatory variable `sales`. Obviously its parameter cannot be intepreted in the same way as for a linear regression. Provide the correct intepretation using odds.

   (c) Randomly pick up five CEOs. Determine their probabilities of having the salary of more or less than 2000. 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 CEO, what is the simplest way to proceed using these probabilities?

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

   (e) 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.

   (f) Recompute the classification table, sensitivity and specifity 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 the salary of CEOs using regression trees.

   (a) Assume the first variable to be used for splitting is `assets`. 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) Which properties of the trees guarantees that prunning using a single complexity parameter works? Give short verbal summary of these properties.