Building Non-Linear Models, Part 1

Anuar Konkashbaev

School of Technology & Engineering, National University

Course code: 7057984436

Professor: Dr. Yavarimanesh

December 4th, 2024

Building Non-Linear Models, Part 1

For this assignment, I will complete three tasks: answer question 1 (for the conceptual part), answer question 8 (for the applied part) from the ISLR Python book (James G., et al., 2023), and compete in the Kaggle competition.

Answers to conceptual part questions

1. We perform best subset, forward stepwise, and backward stepwise selection on a single data set. For each approach, we obtain p + 1 models, containing 0, 1, 2,...,p predictors. Explain your answers:

(a) Which of the three models with k predictors has the smallest training RSS?

The “Best subset” would have smallest training RSS, because it finds the best set of predictors by going over all possible combinations of the set of *p+1* models with *k* predictors. The “Best subset” should have the smallest residual.

(b) Which of the three models with k predictors has the smallest test RSS?

Assuming that the training and testing datasets have similar properties, e.g., distributions, means, standard deviations etc., the model produced by utilizing the “Best subset” method in the training set should also have the smallest residual in test set, for the same reason as in question (a).

(c) True or False:

i. The predictors in the *k*-variable model identified by forward stepwise are a subset of the predictors in the (*k*+1)-variable model identified by forward stepwise selection.

The assertion is true. The forward stepwise method adds a predictor to *k*-variable model with the greatest additional gain in model fit to generate the (*k*+1)-variable model.

ii. The predictors in the k-variable model identified by backward stepwise are a subset of the predictors in the (k + 1)-variable model identified by backward stepwise selection.

The assertion is true. To find the *k*-variable model, the backward stepwise method removes one variable from the existing set (ultimately beginning with the full set of predictors), thus the *k* predictors from the next step are already in the set of *k+*1 predictors.

iii. The predictors in the k-variable model identified by backward stepwise are a subset of the predictors in the (k + 1)-variable model identified by forward stepwise selection.

The assertion is *not* true. In contrast to the forward stepwise method, the backward stepwise method starts from the full set of the predictors, and at each step, it recalculates the model parameters and removes the least useful, so the order in which it removes parameters, can change at any step. Therefore, the set of predictors from the forward stepwise method, that is adding parameters in the predefined order, is not necessarily the same set of predictors from the backward stepwise.

iv. The predictors in the k-variable model identified by forward stepwise are a subset of the predictors in the (k+1)-variable model identified by backward stepwise selection.

This assertion is again not true for the same reasoning as in the answer to question iv.

v. The predictors in the k-variable model identified by best subset are a subset of the predictors in the (k + 1)-variable model identified by best subset selection.

This assertion is not true. The best subset is the result of testing all combinations of the predictors, and those predictors that are best for *k*-variable set, are not necessarily the best contributors in the best subset for a *(k+*1*)*-variable model.

Applied part question

8. In this exercise, we will generate simulated data and will then use this data to perform forward and backward stepwise selection.

(a) Create a random number generator and use its normal() method to generate a predictor X of length n = 100, as well as a noise vector " of length n = 100.

(b) Generate a response vector Y of length n = 100 according to the model Y = β0 + β1X + β2X2 + β3X3 + ", where β0, β1, β2, and β3 are constants of your choice.

(c) Use forward stepwise selection in order to select a model containing the predictors X, X2,...,X10. What is the model obtained according to Cp? Report the coefficients of the model obtained.

(d) Repeat (c), using backwards stepwise selection. How does your answer compare to the results in (c)?

(e) Now ft a lasso model to the simulated data, again using X, X2, ...,X10 as predictors. Use cross-validation to select the optimal value of λ. Create plots of the cross-validation error as a function of λ. Report the resulting coefficient estimates, and discuss the results obtained.

(f) Now generate a response vector Y according to the model Y = β0 + β7X7 + and perform forward stepwise selection and the lasso. Discuss the results obtained.

The answer for this question can be found in the Jupyter notebook “week3.assignment3.ipynb”.

Kaggle competition “Regression with an Abalone Dataset”

The competition description and code are in “week3.assignment3.kaggle.ipynb”

Refences:

Hanke, M., Dijkstra, L., Foraita, R., & Didelez, V. (2024). Variable selection in linear regression models: Choosing the best subset is not always the best choice. *Biometrical Journal, 66(1), 2200209*. https://doi.org/10.1002/bimj.202200209

James G., Witten D., Hastie T., Tibshirani R., T. J. 2023. *An Introduction to Statistical Learning: with Applications in Python (Springer Texts in Statistics) (pp. 283,286). Springer International Publishing*. ISBN: 978-3031387463

Misman M et al., "Prediction of Abalone Age Using Regression-Based Neural Network," 2019 1st International *Conference on Artificial Intelligence and Data Sciences (AiDAS), Ipoh*, Malaysia, 2019, pp. 23-28, doi: 10.1109/AiDAS47888.2019.8970983.