MAP 569 Regression and Classification

Data project: Abalone dataset

Instructions

- You will use Rmarkdown to compile your report in PDF format describing your reasoning, the R outputs (and not the code) for the data experiments you will carry out as well as your conclusions. The following questions are provided as a guideline for your investigation. You should not answer all of them linearly.
- You are required to submit the pdf compiled file and the Rmd filed used to generate it on moodle. No late work will be accepted.

Abalone Data

Abalones are onetype of reef-dwelling marine snails. It is difficult to tell the ages of abalones because their shellsizes not only depend on how old they are, but also depend on the availability of food. The study of age is usually by obtaining a stained sample of the shell and looking at the number of rings through a microscope. We are interested in using some of abalones physical measurements, especially the height measurement to predict their ages. Biologists believe that a simple linear regression model with normal error assumption is appropriate to describe the relationship between the height of abalones and their ages. In particular, that a larger height is associated with an older age.

The dataset and its description is available at https://archive.ics.uci.edu/ml/datasets/Abalone.

We upload the dataset and correct the types of the variables when needed.

```
library(readr)
abalone <- read_csv("Abalone_data.csv")
abalone$Sex <- as.factor(abalone$Sex)
```

We split the original dataset into a training set (70%) and a test set (30%). The test set is set aside for the prediction related questions of Parts II and III.

```
set.seed(42)
#Splitting dataset in train and test using 70/30 method
indexes <- sample(1:nrow(abalone), size = 0.3 * nrow(abalone))
abalone_train <- abalone[-indexes,]
abalone_test <- abalone[indexes,]
```

Part I: EDA and Model validation

Question 1.

Write a mathematical formula modelling the several assumptions inn the above description. Describe what kind of statistical techniques you are going to use to study these hypothesis (confidence intervals, test,...)

Question 2.

Find summary measures of each variables (mean, variance, range, etc). Examine the variables individually (univariate). Graphically display each. Describe what you see.

Question 3.

Generate a labeled scatterplot of the data. Describe interesting features trends. Does it agree with the biologists' hypothesis?

Question 4.

Fit a simple linear regression to the data predicting number of rings using height of the abalones.

Question 5.

Generate a labeled scatterplot that displays the data and the estimated regression function line. Describe the line's fit.

Question 6.

Do diagnostics to assess whether the model assumptions are met; if not, appropriately transform height and/or number of rings and refit your model. Justify your decisions (and recheck your diagnostics).

Question 7.

Interpret your final parameter estimates in context of the problem. Is there a statistically significant relationship between the height and the number of rings (and hence, the age) ofabalones?

Question 8.

Consider now all variables. Look at the scatterplot of the data (`GGally:ggpairs`). Look for correlations betwen predictors. Select some additional variables to add to the simple linear model in order to better predict number of rings. Justify your choices (keep in mind that we want a practical method to predict number of rings). Perform a multiple linear regression. Check the validity of the model. If validity conditions are not met, transform some variables, add/delete some variables, check for outliers and recheck until you find an acceptable model.

Part II ANOVA, ANCOVA

Question 9.

Perform an ANOVA test to decide between the simple linear model and the multiple linear model.

Question 10.

We want to determine whether the covariate *sex* has an impact on the number of rings? To this end, add the variable *sex* to the previous multiple linear model (if you did not already) and perform and ANCOVA analysis of the dataset.

Part III Model Selection and Prediction

Question 11.

Build an estimator of the generalization error of the simple linear model and the multiple linear model using the test set.

Question 12.

Build a model including all variables and maybe including some polynomials. Perform model selection using the BIC criterion. Compare the prediction error on the test set of the selected model to that of the ANOVA selected model of Part II.

Question 13.

Use the package *caret* to implement some ML predictors on the train set including all covariates. You can pick 3 methods among LASSO, Elastic Net, Random Forest, XGBoost, extra trees regression, etc,... Compare their prediction performance to the manually built model.