STAT 425 Final Study 1 (Fall 2024)

Wine Data

Instructor: A. Chronopoulou

Case Study Overview

In the final case study, we are going to investigate the relationship between alcohol content of a wine versus its physiochemical information.

The data set `wines.csv` (found in Canvas) contains the following variables: `fixed acidity`, `volatile acidity`, `citric acid`, `residual sugar`, `chlorides`, `free sulfur dioxide`, `total sulfur dioxide`, `density`, `pH`, `sulphates`, `alcohol`, `type`. All the variables are continuous, except for the variable `type` which is binary with values *'red'* or *'white'*.

Specifically, we have two *separate* goals/tasks:

Model A: Build a regression model to **predict** the alcohol percent of a wine.
Model B: Build a regression model to **describe** the alcohol percent of a wine.

Learning Objectives

By the end of this case study, you will

1. enhance your skills in using R for the purpose of performing a multiple linear regression.
2. independently apply the regression tools discussed in class in a real-world problem.
3. evaluate the applicability of the regression model.
4. draw conclusions, and make decisions about the initially stated research question(s).
5. interpret your statistical outcomes using plain English.
6. demonstrate your team collaboration skills.
7. improve your presentation skills.

More Specific Tasks for the Statistical Analysis

1. You should start with an exploratory data analysis: choose appropriate summary statistics and plots to describe the data.

2. **Model A Building:**
   Your target is the *best* model to predict the alcohol content of the wine. To find it, you should start by creating Training/Testing data sets. Then, you should use the following methods:

(a) Greedy algorithm with AIC and BIC criteria.

(b) Leap-and-bounds algorithm using at least 3 different criteria.

(c) Principal Components Regression.

(d) Ridge and LASSO penalized regression.

You are free to decide the size of training/testing data sets. The best model will be the one that **minimizes** the (Root) Mean Square Prediction Error, that is the (square root of the) mean square error in the testing data set. For simplicity, in this part, **briefly** test for diagnostics, report your findings **but** *do not proceed* with remedial measures.

3. **Model B Building:**
Your target now is different: you want to find the *best* model for estimating the alcohol percent of the wine. That is, you want to find which combination of predictors are better for describing the alcohol content of a wine. In this case, spitting the data into training/testing is not necessary, and "model selection" may be done based on p-values. You need to remember that if the goal is estimation, then you need to:

   (a) check for unusual observations.

   (b) check diagnostics for model assumptions, including collinearity.

   (c) if diagnostics fail, try appropriate remedial measures (measures that we discussed in class only). If you are not able to address all issues, this is ok - you just need to make sure that you explain which are the departures and which are the potential shortcomings of the final model.

4. Which are the similarities/differences between Models A and B. Comment on those.

Deliverables

The case study should be submitted on Gradescope as a group (only one case study per group) and should contain the following files:

(1) an **R Markdown**-generated HTML or PDF technical report containing all the steps in your analysis with discussion of the results along with the corresponding files. This should be professionally and clearly written addressed to someone *who knows statistics*. Do not forget to include an introduction and a conclusion. In your report, please make sure that you remove/hide any R-generated output that is not necessary, e.g. do not print the data or the residuals.

(2) a **PDF** file containing **up to 7 slides presentation** of your project. All members of the group will need to present **in person** and answer project-related questions. Presentations should be up to 10 minutes long and one of the slides should explain the division of work between the group members.
Presentation time slots are posted on Canvas (and here) for the groups to sign up. *For conflicts, please reach out to the instructor in advance.*

Grading

The grading of the case study consists of two-parts:

(1) 60% report (rubric attached)

(2) 40% presentation (rubric attached)

Deadline

Submit *one case study report and presentation per group* on Gradescope by **Monday, December 16** @ 11.59PM. There will be no presentations after **Monday, December 16** @ 6.00PM.

Additional Guidelines

1. The case study should be done in a group of **2–4 students**, which may be the same as before. If you want to change group, please use this form to let me know, and I will do my best to accommodate your request: https://forms.gle/fM2VQNfvKZTKP7Hc6 . The form will remain active until **Friday, December 6**.

2. Use of AI in any part of the project is strictly prohibited. The code and text (i.e. intro, interpretation, conclusion etc) should not be AI-generated or corrected using AI - that is you should not use AI to correct the grammar or syntax of your paragraphs. If such a violation is detected, it will result in a failing grade (i.e. a zero) for the case study, and will be reported to FAIR.

Good luck!