E401/M518: Problem Set 1b

Linear Regression

Fall 2023

Due: September 12 2023

Please work on the following questions and hand in your solutions in groups of at most 3 students.

Part 1: R Questions

Suppose you are hired as a consultant to provide advice on how to improve sales of a particular product. The available data set consists of the sales (in thousands of units) of that product in 200 different markets, along with advertising budgets (in thousands of dollars) for the product in each of those markets for three different media: TV, radio, and newspaper. Our goal is to develop an accurate model that can be used to predict sales on the basis of the three media budgets and obtain an optimal allocation of budgets based on the model. Your objective is to analyze this data set with R to answer your client's questions. For answering the following questions, use the data available in `Advertising.csv` on Canvas.

Question 1: Load and visualize the data

Load the data into R, summarize the data, and plot the data.

Question 2: Simple linear regression

Is there a relationship between sales and the advertising budget for the different media? As a preliminary analysis, run simple regressions of sales on each of the three regressors. Interpret the coefficients. Do all advertising media contribute to sales?

Question 3: Multiple linear regression

To provide a more detailed answer run a multiple linear regression of `Sales` on `TV`, `Radio`, and `Newspaper`. Interpret the slope estimates. Revisit the question of whether all media contribute to sales. How do you reconcile the results for the multiple and simple regressions for newspaper? How strong is the relationship between advertising and sales? Compute the R-squared and discuss your results. Provide a 3-dimensional plot of the model using just TV

and Radio as regressors. Use the library `car` and the command `scatter3d(Sales~TV+Radio)`. You may have to install the package `rgl` explicitly in order to get the `car`-package to work. On Mac OS X, you may also need to install the X11-environment. Experiment with the interactive 3D-graph that the command generates. There is no need to print the 3d-graph. Just the provide the code for it. `rgl` has some nice rotating and zoom features that can be handy sometimes when looking at it on the screen, but it rarely looks good on paper.

Question 4: Models with interaction terms

Are there synergies among the different advertising media? To see this run a multiple regression with an interaction between `TV` and `Radio`. Does this model fit the data better? Experiment with several additional specifications that contain interaction terms.

Question 5: Optimize sales

Your client has an advertising budget of USD 300K. Based on your previous results, how should the budget be divided between TV and radio? (Ignore newspaper advertising since it was not significant), i.e., estimate a model that contains only `TV`,`Radio`, and the interaction term, but not `Newspaper` as a regressor. What is the optimal allocation based on this fitted model with interactions? Maximize the predicted sales by substituting `Radio=300-TV` in the predicted sales and solve the first order condition. What are the optimal sales for the optimal combination? Compare your predictions with the observed sales. Compute a confidence interval for the prediction. Hint: With the fitted model parameters solve the optimization problem by hand to obtain the optimal values of TV and Radio. Then use the R-command `predict`.

Part 2: Reading

Read the paper by Varian (2014), which is posted on Canvas. This is a nice paper that contains many ideas that we will cover in class. Do not worry if you do not understand these ideas yet. The goal is that you will be able to understand them by the end of this course. After reading the paper answer the following questions concisely. What is the goal of *Machine Learning*? What does Varian mean by *good out of sample predictions*? What is *overfitting*? What is *model complexity*? What is the *training data*?