Homework 4: Moving Beyond Linearity

Overview

Due Sunday by 11:59 pm.

Fork the `problem-set-4` repository

Non-linear regression

The General Social Survey is a biannual survey of the American public.[1]

> The GSS gathers data on contemporary American society in order to monitor and explain trends and constants in attitudes, behaviors, and attributes. Hundreds of trends have been tracked since 1972. In addition, since the GSS adopted questions from earlier surveys, trends can be followed for up to 70 years. The GSS contains a standard core of demographic, behavioral, and attitudinal questions, plus topics of special interest. Among the topics covered are civil liberties, crime and violence, intergroup tolerance, morality, national spending priorities, psychological well-being, social mobility, and stress and traumatic events.

In this problem set, you are going to predict individual feelings towards egalitarianism. Specifically, `egalit_scale` is an additive index constructed from a series of questions designed to measure how egalitarian individuals are – that is, the extend to which they think economic opportunities should be distributed more equally in society. The variable ranges from 1 (low egalitarianism) to 35 (high egalitarianism).

`gss_*.csv` contain a selection of variables from the 2012 GSS. Documentation for the other predictors (if the variable is not clearly coded) can be viewed here. Some data pre-processing has been done in advance for you to ease your model fitting, such as imputing missing values. Also note that `income06` has been converted into a continuous feature from its original categorical encoding (https://gssdataexplorer.norc.org/variables/117/vshow).

Egalitarianism and income

1. (20 points) Perform polynomial regression to predict `egalit_scale` as a function of `income06`. Use and plot 10-fold cross-validation to select the optimal degree d for the polynomial based on the MSE. Plot the resulting polynomial fit to the data, and also graph the average marginal effect (AME) of `income06` across its potential values. Be sure to provide substantive interpretation of the results.
2. (20 points) Fit a step function to predict `egalit_scale` as a function of `income06`, and perform 10-fold cross-validation to choose the optimal number of cuts. Plot the fit and interpret the results.
3. (20 points) Fit a natural regression spline to predict `egalit_scale` as a function of `income06`. Use 10-fold cross-validation to select the optimal number of degrees of freedom, and present the results of the optimal model.

Egalitarianism and everything

4. (20 points total) Estimate the following models using all the available predictors (be sure to perform appropriate data pre-processing (e.g., feature standardization) and hyperparameter tuning (e.g. lambda for PCR/PLS, lambda and alpha for elastic net). Also use 10-fold cross-validation for each model to estimate the model's performance using MSE):
   a. (5 points) Linear regression
   b. (5 points) Elastic net regression
   c. (5 points) Principal component regression
   d. (5 points) Partial least squares regression

[1] Conducted by NORC at the University of Chicago.

5. (20 points) For each final tuned version of each model fit, evaluate feature importance by generating feature interaction plots. Upon visual presentation, be sure to discuss the substantive results for these models and in comparison to each other (e.g., talk about feature importance, conditional effects, how these are ranked differently across different models, etc.).