DSA 8020 R Lab 6: Non-parametric Regression and Shrinkage Methods

your name here

Contents

Non-parametric Regression

The dataset `teengamb` concerns a study of teenage gambling in Britain. Type ?teengamb to get more details about the the dataset. In this lab, we will take the variables `gamble` as the response and `income` as the predictor.

Data Source: Ide-Smith & Lea, 1988, *Journal of Gambling Behavior*, 4, 110-118

1. Make a scatterplot to examine the relationship between the predictor `income` and the response `gamble`.

Code:

```
library(faraway)
data(teengamb)
```

2. Fit a curve to the data using a regression spline with `df = 8`. Produce a plot for the fit and a 95% confidence band (using *RegSplinePred <- predict(RegSplineFit, data.frame(income = xg), interval = "confidence")*) for that fit. Is a linear fit plausible?

Code:

Answer:

3. Fit regression curves using *generalized additive models* and *smoothing splines*, respectively, and compare them with the regression spline fit in problem 2.

Code:

Answer:

Ridge Regression andd LASSO: Meat spectrometry to determine fat content

A Tecator Infratec Food and Feed Analyzer, operating in the wavelength range 850 - 1050 nm based on the Near Infrared Transmission (NIT) principle, was employed to collect data on samples of finely chopped pure meat. A total of 215 samples were measured, with both the fat content and a 100-channel spectrum of absorbances recorded for each sample. Due to the time-consuming nature of determining fat content through analytical chemistry, our objective is to construct a model for predicting the fat content of new samples using the 100 absorbances, which can be measured more easily.

Data Source: H. H. Thodberg (1993) "Ace of Bayes: Application of Neural Networks With Pruning", report no. 1132E, Maglegaardvej 2, DK-4000 Roskilde, Danmark

Load the data and partition it into the *training set* (the first 150 observations) and the *testing set* (the remaining 65 observations).

Code:

```
data(meatspec, package = "faraway")
train <- 1:150; test <- 151:215
trainmeat <- meatspec[train,]
testmeat <- meatspec[test,]
```

4. Fit a linear regression with all the 100 predictors to the training set. Compute the root mean square error (RMSE) for the testing set. The code below shows how to compute the RMSE for the training set (also known as in-sample prediction), and you will need to modify the code to compute the RMSE for the testing set.

Code:

```
lmFit <- lm(fat ~ ., data = trainmeat)
# Define a function to calculate RMSE
rmse <- function(pred, obs) sqrt(mean((pred - obs)^2))
# Computing RMSE for the training set
rmse(fitted(lmFit), trainmeat$fat)
```

```
## [1] 0.5919489
```

Answer:

5. Fit a ridge regression (using cross-validation to select the 'best' λ) to the training set and compute the RMSE for the testing set.

Code:

Answer:

6. Fit a LASSO (again using Cross-Validation to select the 'best' λ) to the and training set and compute RMSE for the test set.

Code:

Answer:

7. Fit a LASSO with all the data points (using the best λ from question 6) and report the number of non-zero regression coefficients.

Code:

Answer: