ML 2020-2021 – Fall 2020
Homework 1

2020-10-22

A problem in linear regression, using the **`homes.Rdata`** dataset (that you can find in the Virtual Campus), extracted from the Homes USA repository of data analyzed in the paper Barberan et al. (2015) *The ecology of microscopic life in household dust*. Proceedings of the Royal Society B, Biological Sciences Volume 282, Issue 1814.

The data are dust samples from the ledges above doorways from n=1,059 homes (after removing samples with missing data) in the continental US. Bioinformatics processing detects the presence or absence of 763 species (technically operational taxonomic units) of fungi.

The response is the log of the number of fungi species present in the sample, which is a measure of species richness.

The objective is to determine which factors influence a home's species richness. For each home, eight covariates are included in this example:

1.	long:	Longitude,
2.	lat:	Latitude,
3.	temp:	Temperature,
4.	precip:	Precipitation,
5.	NPP:	Net Primary Productivity (NPP) (See the Appendix) ,
6.	elev:	Elevation,
7.	house:	Single-family home (binary indicator),
8.	bedrooms:	Number of bedrooms.

These covariates are all centered and scaled to have mean zero and variance one.

1. Start with an OLS regression, including main effects, quadratic terms, and all second order interactions. Pre-select relevant predictors using the `step()` function.

2. Apply ridge regression to the resulting model. Compare prediction errors. Which one is better?
3. Apply the Lasso to the (same) resulting model. Any predictor(s) can be discarded due to the feature selection property?
4. PCR
5. PLS

Guidelines

Perform a statistical description of the data. In the first place individually, summarizing each variable, both graphically, e.g. with boxplot and histogram, and numerically. Do these variables have a normal appearance? Or, rather, do these variables show an asymmetric shape? Check correlations between pairs of predictors and between individual predictors and response. It will be useful to truncate to 2 or 1 decimal places, to avoid clutter: In this way we can see at a glance which correlations are large or small. Is there multicollinearity?

Note

It is possible that you happen to find a study in some online sources. It is OK to learn from and even reproduce parts of these sources. This is acceptable provided that you:

1. Give credit where it is due, in particular, including full reference (URL) of any cited work
2. Do not copy/paste "*in extenso*" large chunks of code.
3. Understand everything you write, explaining with sufficient details the steps you take and the results you obtain.