Multivariate Statistics: Assignment 1

Submit your answers in p.d.f. or html format by Monday 1st December 2025, 13:00 to me at `noble@mimuw.edu.pl`

Exercise 1

The file `bodyfat2.xlsx` contains measurements of the percentage of bodyfat for 252 men. The Y-variable is the bodyfat percentage; there are 13 explanatory variables (X-variables). *Do not include 'density' as an explanatory variable; the* 13 *variables to the right of 'bodyfat' are the explanatory variables.* One way to load an `.xlsx` file is by:

```
> www =
"https://www.mimuw.edu.pl/~noble/courses/MultivariateStatistics/data/bodyfat2.xlsx"
> library(rio)
> bodyfat = import(www)
```

1. Consider the correlations between the 13 explanatory variables. Are there grounds to suspect ill-conditioning?

2. Perform regression analysis on this data, using the principal component approach, partial least squares and the other methods considered in the course. Use leave-one-out cross-validation for estimating the mean prediction error as a criterion for model selection. Which subset of variables gives the best model, based on this criterion? And which method gives the best results.

3. Now give special attention to the LASSO method applied to the bodyfat data set. Indicate the LASSO path and decide on a suitable model. Justify your choice.

Exercise 2: Boston Housing Data

Least Angle Regression and LASSO can be carried out using routines in the **lars** package and also the **glmnet** package. The Boston housing data can be found in the file `boston_corrected.txt` in the course data directory. Information may be found here:

`https://www.cs.toronto.edu/~delve/data/boston/bostonDetail.html`

There are 506 observations on census tracts in the Boston Standard Metropolitan Statistical Area (SMSA) in 1970. For the response variable, use the *logarithm* of `MEDV`, the median value of owner-occupied houses in thousands of dollars. There are 13 input variables (plus information on location of each observation). The 13 explanatory variables are: `CRIM, ZN, INDUS, CHAS, NOX, RM, AGE, DIS, RAD, TAX, PTRATIO, B, LSTAT`. The url above gives information on these variables. Compute the OLS estimates.

- **Note 1** The first few lines are text and should be removed from the file.
- **Note 2** Not all the instantiations are complete. Remove the incomplete instantiations before performing a regression analysis.

Is OLS regression effective? Or do the penalised regression techiques give a better answer? Decide on the best regression technique (from those dealt with in the course) and analyse the data according to this method.

Exercise 3: Doctor Visits Data

Consider the `DoctorVisits` data in the **AER** package. Use a Poisson regression for the number of visits. Is the Poisson model satisfactory? If not, where are the problems and what can be done about them? (**Note** Please note the limitations of diagnostics for count data. For example, if we have $X_1, \ldots, X_n$ i.i.d. Bernoulli($\frac{1}{2}$), then each observation will be either 0 or 1, so even if we have the 'correct' model and the 'correct' estimate $\widehat{p} = \frac{1}{2}$, the 'error sum of squares' will still be $\sum_{j=1}^{n}(X_j - \frac{1}{2})^2 = \frac{n}{4}$ which is substantial; a large residual sum of squares does not necessarily imply that the model is bad.)

Exercise 4: Diabetes

The package **pacman** is useful. Having installed it, use the following to get a data set of observations of 9 variables on 392 women, where the instantiations are complete. The last variable gives positive or negative for diabetes.

```
knitr::opts_chunk$set(echo = TRUE,
                      message = F,
                      warning = F,
                      fig.align = 'center')
# Loading any packages
pacman::p_load(tidyverse, mlbench, broom, pander)

# Getting the PimaIndiansDiabetes2 data set from mlbench package
data("PimaIndiansDiabetes2",
     package = "mlbench")



pid2 <-
  PimaIndiansDiabetes2 %>%
  filter(complete.cases(.)) |>
  mutate(diabetes = relevel(diabetes, ref = "neg"))
```

```
rm(PimaIndiansDiabetes2)
```

Build a model for predicting the probability of diabetes. Which of the models: Logistic, Probit, Extreme value works best? Try logistic LASSO using **glmnet**.