SF2930 VT25: Regression Analysis
Project 1 instructions
For questions or comments, please contact Isaac Ren (isaacren@kth.se) and/or Hanqing Xiang (hanqingx@kth.se).

This project should be done in groups of **at most two**.

A self-contained PDF report (LaTeX is recommended) of the subjects presented below should be handed in on Canvas no later than **2025-03-11**. Please **name the document** as follows:

```
SF2930Project1-FullName1-FullName2.pdf
```

Note that the deadline is the day of the exam. Please consider submitting early: it makes life easier for you and for the graders! You may be asked to correct your report and resubmit.

Introduction

This project aims at illustrating two typical scenarios in modern regression modeling. Choose **only one** of the two following alternative scenarios. Your choice must be clearly specified in the project report.

- **Scenario I:** Large-Sample Regression, $p < n$.
  You are expected to work with classical methods of statistical inference which are applicable in this case.
- **Scenario II:** High-Dimensional Regression, $p > n$ or $p \gg n$.
  In this situation, the scientist is often looking for a few informative predictor variables hidden in an ocean of uninformative ones. There is a variety of approaches to tackle this "curse of dimensionality" problem in regression analysis; you are expected to focus on a subset of such.

All the R-functions needed to perform this project will be presented during the exercise sessions and are available from the course Canvas page.

Scenario I: Body mass fat data ($p < n$)

The World Health organization (WHO) has reported that obesity is a major risk factor for a number of chronic diseases, including diabetes, cardiovascular diseases, and cancer. Obesity is defined as "the disease in which excess of body fat has accumulated to such extents that health may be adversely affected." Once considered a problem only for high-income countries, obesity is now rising in low- and middle-income countries. An important issue for medical purposes thus is to reliably identify people with excess fat.

Data Description

The well-known body mass index (BMI=weight/height^2), while widely used in practice and simple to calculate, is only an indirect measure of fatness and is empirically shown to be a poor predictor of actual fatness. Instead, a person's *body fat mass* (BFM)[1] is considered. Highly accurate methods for measuring the BFM, such as X-ray densitometry (DXA) or hydrodensitometry, while being precise, have little practical applicability because of the high costs and methodological efforts. Thus, cheaper and more portable methods such as regression models attract a lot of interest in body composition research.

A number of anthropologic measurements such as waist circumference and waist-to-hip-ratio, combined with skin-fold thickness are known to be related to the BFM. These variables can be used as predictor variables in multiple linear regression models, which can then be used for predicting BFM instead of measuring it exactly.

For deeper understanding of the topic, see Garcia et al. (2005).

Goals

The goal of this project is to develop and validate your own regression model for prediction of BFM (density). We recommend that you follow the strategy for model building and variable selection presented in Montgomery et al., Section 10.3; see also the flow chart in Montgomery et al., Figure 10.11. Use **one** of the datasets described on the next page. The following aspects of the model development are expected to be discussed in the project.

- Thorough residual analysis for model adequacy checking, including various types of residual scaling and plotting.
- Diagnostics and handling of outliers, leverage and influential observations using e.g. Cook's distance and CovRatio.
- Possible transformations of the variables to correct model inadequacies.
- Multicollinearity diagnostics and treatments.
- Different types of variable selection (e.g. all possible regressions, forward/backward elimination) using model evaluation criteria such as e.g. MSE, AIC, BIC, Mallows' C_p and adjusted R^2. Evaluate MSE and adjusted R^2 using cross validation (CV) (see p. 250 in ISL), i.e. calculate the evaluation criteria on the test set.

[1] BFM is in fact a person's body density (mass/volume)

- Computer-intensive procedures for the model assessment, such as bootstrap residuals in (Montgomery et al., 2012, §15.4.1) or bootstrap based confidence intervals for regression coefficients using the percentile method (Montgomery et al., 2012, §15.4.2).

Specify clearly your final model.

Datasets

Two different datasets are available, one for men and one for women. Choose one of them to use in your project for scenario I. The datasets come from different studies and do not contain exactly the same columns. Short descriptions of the datasets along with references are given below. The datasets are also available on the course Canvas page.

BFM men. This dataset contains measurements of the body density (`density`) of 252 men assessed by hydrodensitometry (underwater wighting) along with their age and a number of anthropometric variables. The dataset is available on the course Canvas page.

Observe that this is a modified version of the orginal dataset available in the package `mfp`. In this version of the dataset, the columns `case`, `brozek` and `siri` are removed. Note also that certain data points that may be erroneous; see the data description found at `https://cran.r-project.org/web/packages/mfp/mfp.pdf`. For more detailed presentation of the dataset, see `http://lib.stat.cmu.edu/datasets/bodyfat` and (Izenman, 2009, Example 5.5.2).

BFM women. This dataset, introduced in Garcia et al. (2005) contains measurements of the BFM (`DEXfat`) of 71 women assessed by DXA. The data can be found in the R-package `TH.data` which is installed by typing `install.packages("TH.data")`. After installation, you can use the dataset by typing

```
library("TH.data")
data("bodyfat")
```

The data is now available in the variable `bodyfat`. Type `??bodyfat` to get the explanation of the columns. This dataset is not exactly the same as in Garcia et al. (2005). Firstly, it contains contains measurement of women only. Secondly, some of the variable are transformed, e.g.

```
anthro3a = log(chin) + log(triceps) +  log(subcapular)
```

and some of the variable are presented as log transformed product of anthropological measures, e.g. `anthro3b`.

Scenario II: Riboflavin production by Bacillus subtilis ($p \gg n$)

In this scenario, we consider a high-dimensional regression model dealing with the gene expression microarray data on riboflavin (vitamin B_2) production with *Bacillus subtilis*, as presented in

Bühlmann et al. (2014). Riboflavin belongs to the vitamin B group and is responsible for cellular respiration in the body. It is included in the *WHO Model list of essential medicines* representing most efficient and safe medicines in a health care system.

The dataset consists of $n = 71$ samples that were hybridized repeatedly during a fed-batch fermentation process where different engineered strains and strains grown under different fermentation conditions were analyzed. The samples were normalized using the default in the R-package `affy` (Gautier et al., 2004). For deeper understanding of the topic see Bühlmann et al. (2014) and references therein.

Data description

The $n = 71$ samples each has a single real-valued response variable which is the logarithm of the riboflavin production rate; furthermore, there are $p = 4088$ (co-)variables measuring the logarithm of the expression level of 4088 genes.

Goals

The overall goal is to explore the linear model selection and regularization strategies for the prediction of the riboflavin production rate using the expression levels of $p = 4088$ genes as a set of potential predictor variables.

To be approved for the project, you should work with **all three sub-goals** specified below.

The **first** goal is to develop and validate the PCA- and PLS-regression models for prediction of the riboflavin production. Start by splitting the data into a training set and a test set, use the ratio 3:1. The following aspects of the model development are expected to be discussed.

(a) For both approaches use cross-validation (CV) to fit the models on the training set and motivate your choice of number of folds.

(b) Report and comment on the number of principal components for PCA and partial least squares directions for PLS regression which minimizes the training set MSE.

(c) Using your results from step 2, choose the models and evaluate its performance accuracy with the test set MSE.

The **second** goal is to apply the regularization (shrinkage or penalized LS) techniques. Specifically, fit the ridge- and Lasso-regression models to the riboflavin data. Start by splitting the data into a training set and a test set, using the ratio 3:1. The following aspects of the model development are expected to be discussed.

(a) Fit both of the models using the training set, plot and interpret both the Ridge trace and Lasso path of the regression coefficients as a function of the log penalty parameter.

(b) For both approaches, report (plot) the cross-validated MSE as a function of the log penalty parameter on the training set and motivate your choice of number of folds in the cross-validation.

(c) For both approaches, specify the optimal (according to the training set MSE) value of the penalty parameter, build the corresponding linear regression model and evaluate its prediction accuracy on the test data (using the test set MSE). Report and comment your results.

See (James et al., 2013, Ch. 6) for guidelines of R-implementation of ridge, Lasso and cross validation.

The **third** goal is to explore post-selection inference for the Lasso.

To work with this part of the project, read carefully the introduction and sub section 2.1.1 in Dezeure et al. (2015). Discuss the principles of *single* and *multi sample-splitting* for construction of hypothesis test or confidence intervals. Motivate the need of the multi-sample splitting approach in the post-selection inference.

The multi sample-splitting functionality is implemented in the `hdi` package as a function also named `hdi`. Apply this approach to the `riboflavin` dataset using 100 splits by the following commands.

```
library("hdi")
data("riboflavin")
fit.multi <- hdi(riboflavin[,-1], riboflavin[,1], B=100)
```

Which regression coefficient(s) do you find as significant on the 10% level? Report them along with their corresponding p-values and confidence intervals and comment on your findings. The p-values are then obtained from `fit.multi$pval.corr`.

Dataset

The dataset is accessible from the data frame `riboflavin` and can e.g. be viewed by typing

```
install.packages("hdi")
library("hdi")
data("riboflavin")
View(riboflavin)
```

References

Bühlmann, P., Kalisch, M., and Meier, L. (2014). High-dimensional statistics with a view toward applications in biology. *Annual Review of Statistics and Its Application*, 1(1):255–278.

Dezeure, R., Bühlmann, P., Meier, L., Meinshausen, N., et al. (2015). High-dimensional inference: Confidence intervals, p-values and r-software hdi. *Statistical science*, 30(4):533–558.

Garcia, A. L., Wagner, K., Hothorn, T., Koebnick, C., Zunft, H.-J. F., and Trippo, U. (2005). Improved prediction of body fat by measuring skinfold thickness, circumferences, and bone breadths. *Obesity Research*, 13(3):626–634.

Izenman, A. (2009). *Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning*. Springer Texts in Statistics. Springer New York.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An Introduction to Statistical Learning: with Applications in R*. Springer Texts in Statistics. Springer New York.

Montgomery, D., Peck, E., and Vining, G. (2012). *Introduction to Linear Regression Analysis*. Wiley Series in Probability and Statistics. Wiley.