# SF2930 Regression analysis VT2017

Project 1

The project should be done in groups of **two**.

A computer written[1], self-containing, report of the subjects presented below should be handed in no later than **2017-03-06** by email to `flrios@math.kth.se`. The **subject** of the email should be

```
SF2930 Project 1: Full Name 1, Full Name 2
```

and **name the document**

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

## Introduction

The World Health organization (WHO) reported that the obesity is a major risk factors 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 extend that health may be adversely affected". Once being considered as a problem only for high income countries, obesity is now rise in low- and middle-income countries. An important issue for medical purposes is thus is to reliably identify people with the fat excess.

The well known body mass index (BMI=weight/height$^2$), while being widely used in practice and simple to calculate is only an indirect measure of the fatness and is empirically shown to be a poor predictor of actual fatness. Instead, a persons *body fat mass* (BFM)[2] 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 portable methods such as regression models attract a lot of interest in the body composition research.

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

For deeper understanding of the topic see and in particular Garcia et al. [2005].

[1] Preferably using LaTeX

[2] BFM is in fact a persons body density (mass/volume)

## The project

The goal of this project is to develop and validate your own regression model for prediction of BFM (density). Following the strategy for model building and variable selection presented in Montgomery et al., Section 10.3, and flow chart in Montgomery et al., Figure 10.11, perform all the steps of regression analysis. The following aspects of the model development are expected to be discussed in the project.

1. Thorough residual analysis for model adequacy checking, including various types of residual scaling and plotting.
2. Diagnostics and handling of outliers, leverage and influential observations using e.g. Cook's distance and CovRatio.
3. Possible transformations of the variables to correct model inadequacies.
4. Multicollinearity diagnostics and treatments. Ridge and Lasso regression with penalty traces and principal component regression.
5. Different types of variable selection (e.g. all possible regressions, forward/backward elimination) using model evaluation criteria such as e.g. MSE, AIC, BIC, Mallow's $C_p$ and adjusted $R^2$.
6. Computer-intensive procedures for model evaluation, e.g. Bootstrap based confidence intervals for the coefficient and cross validation for estimating the prediction error.

You should put more emphasis on at least two of the subjects and analyze them more thoroughly. For example in 3., you can have a look at the scatter plot (using e.g. `pairs` in R) matrix as well as the histograms (using e.g. `hist` in R) of each variable and answer the following questions. Which variables should be transformed? Do the transformations and use the transformed values in further analysis. In 4. you can e.g. do the following; fit a ridge (Lasso) model on the training set, with the penalty parameter $k$ chosen by cross-validation. Report the test error obtained, along with the number of non-zero coefficient estimates and comment on the results obtained. See [James et al., 2013, Ch. 6] for guidelines of R-implementation of ridge, Lasso and cross validation.

All graphs and tables in the report should be properly presented and discussed. Specify clearly the final model you arrive to and summarize it in an ANOVA table. Such important model characteristics as $p$-values, confidence intervals for the coefficients of the included variables and all other important measures of model adequacy should be included.

**All the R-function needed to perform this project will be presented during the exercise sessions.**

## Datasets

Two different datasets are available, one for men and one for women. The datasets come from different studies and do not contain exactly the same columns. Short descriptions of the datasets along with links are given below.

**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. Download the dataset **directly** from the course web page by typing

```
bodyfat <- read.csv('http://www.math.kth.se/
                      matstat/gru/sf2930/bodyfat_men.csv')
```

Observe that his is a modified version of the orginal dataset available in the package `TH.data`. In this version of the dataset, the columns `case, brozek` and `siri` are removed and the rows corresponding to case 48, 76, and 96 are removed following 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`.

# References

Ada L. Garcia, Karen Wagner, Torsten Hothorn, Corinna Koebnick, Hans-Joachim F. Zunft, and Ulrike Trippo. Improved prediction of body fat by measuring skinfold thickness, circumferences, and bone breadths. *Obesity Research*, 13(3):626–634, 2005. ISSN 1550-8528.

A.J. Izenman. *Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning*. Springer Texts in Statistics. Springer New York, 2009. ISBN 9780387781891. URL `https://books.google.se/books?id=1CuznRORa3EC`.

G. James, D. Witten, T. Hastie, and R. Tibshirani. *An Introduction to Statistical Learning: with Applications in R*. Springer Texts in Statistics. Springer New York, 2013. ISBN 9781461471387.

D.C. Montgomery, E.A. Peck, and G.G. Vining. *Introduction to Linear Regression Analysis*. Wiley Series in Probability and Statistics. Wiley, 2012. ISBN 9780470542811.