MATH 423/533 - FINAL PROJECT

To be handed in not later than 10pm, 20th December 2016.
Please submit your solutions as pdf via myCourses.

This project focusses on the NHANES health and nutrition study

`https://www.cdc.gov/nchs/nhanes/`

which a large ongoing survey in the US. In `R`, a version of the study data is contained in the data frame `NHANES` stored in the library `NHANES`: this data frame contains 10000 observations made on 76 variables. It is available in `R` by installing the package

```
install.packages('NHANES')
library(NHANES)
```

The help file for the data frame is here

`http://127.0.0.1:23156/library/NHANES/html/NHANES.html`

The data frame contains observations on two survey years (2009-10 and 2011-12). It represents a representative (random) sample of the US population obtained from raw survey data by weighted sampling.

This project focusses on a subset of the data, from the 2011-12 year, for which complete observations are available on 21 of the variables (plus the a subject identifier); the data subset is stored in the comma separated file `nhanes-sub.csv`

`http://www.math.mcgill.ca/dstephens/Regression/Data/Project/nhanes-sub.csv`

The objective of the project is to find regression models that explain observed variation on the two blood pressure measurements, recorded in the data frame as `bpdia` and `bpsys`.

(a) Using the model building and selection techniques (eg F-testing, selection criteria), use the data to find models that best represent the variation in the two response variables `bpdia` and `bpsys`.

(b) Check your best models using numerical and graphical model adequacy assessments.

(c) Interpret your results in context: that is, explain in the context of the study how the statistical results should be interpreted.

20 Marks

Please limit your project report to no more than 12 pages, and also upload your computational code as an `R` script.