Assignment 2. ST464/ST684
Catherine Hurley
Due Friday March 12, 6pm.

- Do all questions.
- Upload solution to questions 1 and 3 only to Moodle by Friday March 12, 6pm.
- You should complete the assignment in Rmarkdown, knit the file to html and upload the html file to Moodle.
- Please use the skeleton Rmarkdown file h2student.Rmd available on Moodle.
- Fill in your name and student number on the solution file in the space provided.
- The upload must be completed by time and date given above or the assignment will not be accepted.
- The tutorial on week 4 will provide assistance with this material.

1. The data morpho gives a morphological description of 153 athletes split in five different sports. The dataset has the following variables bia (biacromial diameter (cm)), height (height (cm)), bhd (distance from the buttocks to the top of the head (cm)), arm (length of the upper limbs (cm)), weight ((kg)) and sport.

```
morpho <- read.csv("morpho.csv")
morpho$sport<- as.factor(morpho$sport)
```

   (a) Use pairs to construct a scatterplot matrix of the numeric variables, colour by sport. Are there any outliers? If so, which cases are they?What sport?

   [3 marks]

   (b) Carry out a principal components analysis of the numeric variables. Decide for yourself re scaling of the variables. What percentage of the variability in the dataset is accounted for by the first component? What percentage of the variability in the dataset is accounted for by the first two components? Examine the scree diagram and comment. (You will find the code for the screeplot in screeplot.R).

   [3 marks]

```
source("screeplot.R")
```

   (c) What does the first component measure? the second component? Make a biplot to assist your interpretations. Are there any outliers?

   [3 marks]

   (d) Make a plot of the first two PCs again, showing the points only. This time colour the points by their sport. Do athletes from the same sport cluster together in the PC plot?

   [3 marks]

2. A 1902 study obtained measurements on seven physical characteristics for each of 3000 criminals. The seven variables measured were (1) head length (2) head breadth (3) face breadth (4) left finger length (5) left forearm length (6) left foot length (7) height. Using the correlation matrix given below, find the principal components of the data and interpret the results. What percentage of the variability in the dataset is accounted for by the first component? What percentage of the variability in the dataset is accounted for by the first

two components? Examine the scree diagram and comment.

Hint: use the `eigen` function in R to calculate the eigen values and vectors.

$$\begin{bmatrix} 1.000 & & & & & & \\ 0.402 & 1.000 & & & & & \\ 0.396 & 0.618 & 1.000 & & & & \\ 0.301 & 0.150 & 0.321 & 1.000 & & & \\ 0.305 & 0.135 & 0.289 & 0.846 & 1.000 & & \\ 0.339 & 0.206 & 0.363 & 0.759 & 0.797 & 1.000 & \\ 0.340 & 0.183 & 0.345 & 0.661 & 0.800 & 0.736 & 1.000 \end{bmatrix}$$

```r
# read in the correlation data as a vector

crimcorr <- matrix(c(
  1.000, 0.402, 0.396, 0.301, 0.305, 0.339, 0.340,
  0.402, 1.000, 0.618, 0.150, 0.135, 0.206, 0.183,
  0.396, 0.618, 1.000, 0.321, 0.289, 0.363, 0.345,
  0.301, 0.150, 0.321, 1.000, 0.846, 0.759, 0.661,
  0.305, 0.135, 0.289, 0.846, 1.000, 0.797, 0.800,
  0.339, 0.206, 0.363, 0.759, 0.797, 1.000, 0.736,
  0.340, 0.183, 0.345, 0.661, 0.800, 0.736, 1.000), nrow = 7, byrow = TRUE)
colnames(crimcorr)<- c("Head-L","Head-B","Face-B",
                       "L-Fing","L-Fore","L-Foot",
                       "Height")
```

3. For each of the following situations, answer, if possible: (i) Is it a classification or regression problem? (ii) Are we most intererest in inference or prediction? (iii) Provide n and p. For each predictor described state whether it is categorical or quantitative. (iv) Indicate whether we would expect the performance of a flexible learning method to be better or worse than an inflexible method.

   (a) We have a set of data on 500 worldwide tech firms. For each firm, information on profit, CEO salary, number of employees, average employee salary, and home country is recorded. We are interested in the relationship between CEO salary and other measurements.

   [2 marks]

   (b) A company wishes to launch a new product. They want to know in advance whether it will be a success or failure. They collect data on 20 similar products, and record whether they succeeded or not, price charged, marketing budget, and 10 other variables.

   [2 marks]

   (c) A dataset was collected to related the birthweight of babies to the days of gestation and gender.

   [2 marks]

   (d) Observations were collected on 56 attributes from 32 lung cancer patients belonging to one of 3 classes.

   [2 marks]

4. In this exercise you will conduct an experiment to compare the fits on a linear and flexible model fit. You will use the Auto data from the package ISLR and explore the relationship between the response mpg with weight and horsepower.

(a)
```
# install.packages("ISLR") #home computer, first time only
library(ISLR)
Auto <-Auto[complete.cases(Auto[,c(1,4,5)]),] # to remove NAs
plot(mpg ~ weight, data=Auto)
plot(mpg ~ horsepower, data=Auto)
```

Construct the above plots. What do they tell you about the relationship between mpg and the predictors?

(b)
```
# install.packages("plot3D") #home computer, first time only

library(plot3D) # install package
scatter3D(Auto$weight,Auto$horsepower,Auto$mpg)

library(plot3Drgl)
scatter3Drgl(Auto$weight,Auto$horsepower,Auto$mpg)
```

Construct this plot. What do they tell you about the relationship between mpg and the predictors?

(c) Next, divide the data into a training set and a test set as follows:

```
set.seed(123)
train <- sample(nrow(Auto), round(.8*nrow(Auto)))
AutoTrain <- Auto[train,]
AutoTest <- Auto[-train,]
```

Fit a linear regression model to mpg versus weight and horsepower on AutoTrain. Call the fit f1. Examine summary(f1) and comment on the significance of the predictors.

(d) Plot the fitted surface and the data. (See lecture notes for code) Does the linear surface look like a good fit?

(e) Use loess to fit a surface to the same data. Call the fit f2.

```
f2 <- loess(mpg~weight+horsepower, data=AutoTrain)
```

Plot the fitted surface and the data. Does the loess surface look like a good fit?

(f) Calculate the MSE for both fits on the training data. What do these numbers tell you? (See lecture notes for code.)

(g) Calculate the MSE for both fits on the test data. What do these numbers tell you?