MTH 4230 Lab 4

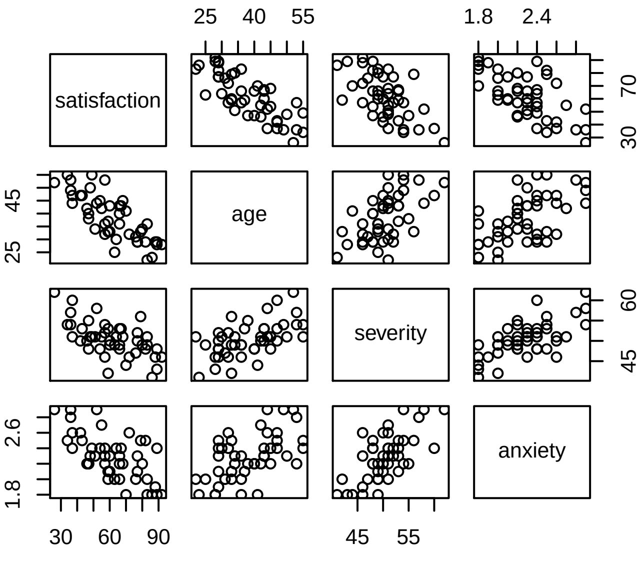
Due Wed., Feb. 26

1 Part A: Multiple Regression

1.1 Patient Satisfaction Data Set

A hospital administrator wished to study the relation between patient **satisfaction** (Y) and patient's **age** (X_1, in years), **severity** of illness (X_2, an index), and **anxiety** (X_3, an index). The administrator randomly selected 46 patients and collected the data presented in the file **satisfaction.txt**. This is the **Patient Satisfaction** data set from **Problem 6.15** of the textbook.

1. Read the data into a *data frame* using `read.table()`.
2. Create a ***scatterplot matrix*** of the data by passing your *data frame* to the function `pairs()`. It should look like this:



3. Compute the sample ***correlation matrix*** of the data by passing your *data frame* to the function `cor()`.
4. Use `lm()` to fit the ***multiple regression model***

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$$

to the data. Check the results using `summary()`.

5. Check the **normality assumption** for the error term ϵ by making a **histogram** and **normal probability plot** of the residuals using `hist()` and `qqnorm()` followed by `qqline()`.

6. Check the **constant standard deviation assumption** for the error term ϵ by making a plot of the **residuals** versus the **fitted values** using `plot()`, then add a horizontal line at zero using `abline(h = 0)`.

2 Part B: Linearly Dependent Design Matrix

2.1 Hypothetical Snakes Data

It can be shown that when the columns of the **design matrix** $\boldsymbol{X}$ are ***linearly dependent*** (e.g. one is a linear combination of the others), the matrix $\boldsymbol{X}^T\boldsymbol{X}$ will be ***singular*** (i.e. *not invertible*).

1. Consider the following *hypothetical* data on **lengths** (**cm** and **in**), **diameters** (**cm**), and **weights** (g) of $\boldsymbol{n} = \mathbf{9}$ snakes:

```
snakes

##   LengthCm LengthIn DiameterCm WeightG
## 1       60     24.0        2.1     136
## 2       69     27.6        2.2     198
## 3       66     26.4        2.2     194
## 4       64     25.6        2.2     140
## 5       54     21.6        1.6      93
## 6       67     26.8        2.2     172
## 7       59     23.6        2.0     116
## 8       65     26.0        2.1     174
## 9       63     25.2        2.2     145
```

The **lengths** in **inches** were computed from the **lengths** in **centimeters** using

$$1\,\text{in} \;=\; 2.5\,\text{cm}\,.$$

After creating the (hypothetical) `snakes` data frame:

```
lengthCm <- c(60, 69, 66, 64, 54, 67, 59, 65, 63)
weightG <- c(136, 198, 194, 140, 93, 172, 116, 174, 145)
lengthIn <- c(24.0, 27.6, 26.4, 25.6, 21.6, 26.8, 23.6, 26.0, 25.2)
diameterCm <- c(2.1, 2.2, 2.2, 2.2, 1.6, 2.2, 2.0, 2.1, 2.2)
snakes <- data.frame(LengthCm = lengthCm,
                     LengthIn = lengthIn,
                     DiameterCm = diameterCm,
                     WeightG = weightG)
```

try fitting the ***multiple regression model***:

```r
my.reg <- lm(WeightG ~ LengthCm + LengthIn + DiameterCm, data = snakes)
```

Then look at the outputs of the following:

```r
summary(my.reg)
model.matrix(my.reg)
```