Notes 1

Maryclare Griffin

2/7/2023

These notes are based on Chapters 1 and 6 of KNNL.

In this class, we will be interested in describing relations between variables, a dependent variable or response Y and independent variables, predictors, or covariates X_1, … X_{p-1}. Specifically, we will be interested in describing **statistical** relations, as opposed to **functional** relations.

> ***Note:*** When we just have one independent variable or predictor ($p = 2$) and we will drop the subscript and let $X = X_1$ for convenience.

A **functional** relation between variables is a relation that can be expressed by a mathematical forumula given some known function f that maps values of the predictors to a value of the response:

$$Y = f\left(X_1, \dots, X_{p-1}\right).$$

You can recognize a functional relation as follows - if you know the values of the predictors X_1, …, X_{p-1}, you can determine Y, it is a **functional relation**.

> ***Example 1:*** Let Y refer to dollar sales of a product sold at a fixed price per unit and let X refer to the number of units sold. If the selling price is 2 per unit, the relation is expressed by:
>
> $$Y = 2X$$
>
> This is a **functional** relation because we know the value of Y if we are given the value of X. We might observe a scatterplot of the data as depicted in Figure 1. We can also tell this is a **functional** relationship because the line $Y = 2X$ passes through **every point**.

```r
X <- c(75, 25, 130) # Create the X values
Y <- c(150, 50, 260) # Create the Y values
plot(X, Y) # Plot values of Y against values of X
curve(2*x, from = 25, to = 130, add = TRUE) # This adds the line Y = 2X to the plot
```

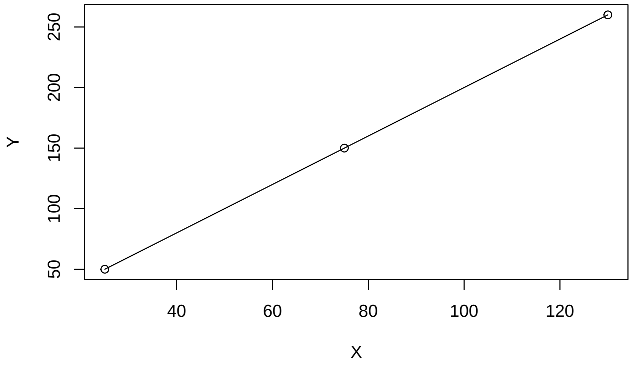

Figure 1: Example 1

In contrast, a **statistical** relation is not perfect. When there is a **statistical** relation between variables, you cannot determine Y even if you know the values of the predictors $X_1, \ldots, X_{p-1}$. At best, you can make statements about likely values of Y given certain values of the predictors $X_1, \ldots, X_{p-1}$.

> ***Example 2:*** Consider data from a company that manufactures refrigeration equipment, called the Toluca company. They produce refrigerator parts in lots of different sizes, and the amount of time it takes to produce a lot of refrigerator parts depends on the number of parts in the lot and several other variable factors. Let X be the number of refrigerator plots in a lot, and let Y refer to the amount of time it takes to produce a size of lot X. We might observe a scatterplot of the data as depicted in Figure 2. We can also tell this unlikely to be a **functional** relationship because we can see that there lots of similar sizes that did not take the same number of hours to produce.

```r
load("~/Dropbox/Teaching/STAT525/Spring2023/bookdata/toluca.RData")
X <- data$X # Extract the X values
Y <- data$Y # Extract the Y values
plot(X, Y, xlab = "Lot Size",
     ylab = "Hours") # Plot values of Y against values of X with axis labels
```

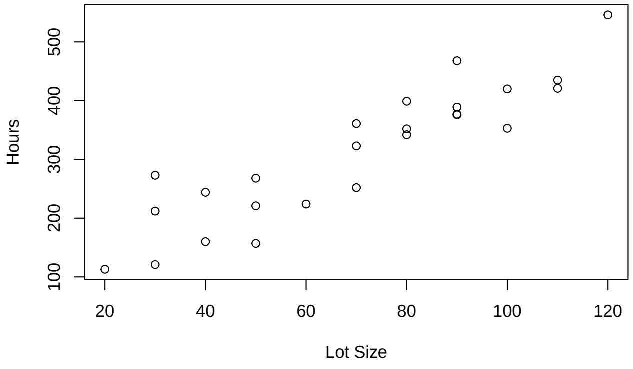

Figure 2: Example 2

Example 3: Consider data on age (X) and plasma level of a polyamine (Y) for 25 healthy children. We might observe a scatterplot of the data as depicted in Figure 3. Again, we can also tell this unlikely to be a **functional** relationship because we can see that there are healthy children of the same age with different plasma levels.

```r
load("~/Dropbox/Teaching/STAT525/Spring2023/bookdata/plasma.RData")
X <- data$X # Extract the X values
Y <- data$Y # Extract the Y values
plot(X, Y, xlab = "Age (years)",
     ylab = "Plasma Level")
```

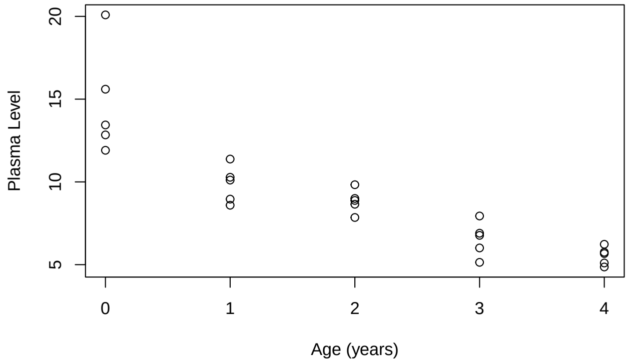

Figure 3: Example 3

Example 4: Consider data from portrait studios in 21 cities run by Dwaine Studios, Inc. The studios specialize in portraits of children. Let X_1 be the number of persons aged 16 or younger in a city, let X_2 refer to per capita disposable income in a city, and let Y be the sales of portraits of children in that city from one of the 21 studies. We might observe a plot of the data as depicted in Figure 4. Because we have two predictors, this is much more difficult to visualize. The axes indicate the values of the predictors, and the size of the dot indicates the corresponding sales. Again, we can tell this unlikely to be a **functional** relationship because we can see that there cities with similar numbers of persons aged 16 or younger and similar per capita disposable income that have very different sales.

```r
load("~/Dropbox/Teaching/STAT525/Spring2023/bookdata/dwaine.RData")
X1 <- data$X1 # Extract the first predictor
X2 <- data$X2 # Extract the second predictor
Y <- data$Y # Extract the response
plot(X1, X2, cex = Y/100,
     xlab = expression(X[1]),
     ylab = expression(X[2]), pch = 16,
     col = rgb(0, 0, 1, 0.75))
legend("topleft",
       pt.cex = c(min(Y/100), max(Y/100)),
       legend = c(min(Y), max(Y)),
       pch = 16,
       col = rgb(0, 0, 1, 0.5),
       title = expression(Y))
```

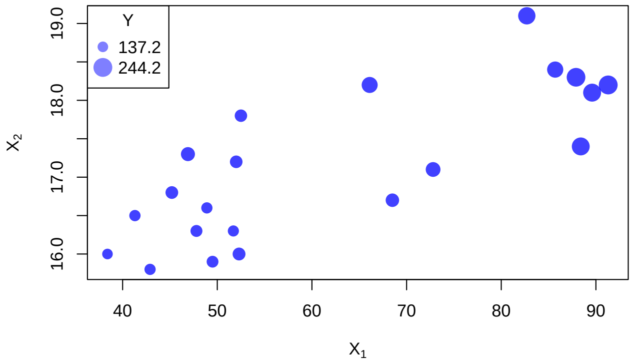

Figure 4: Example 4

This class is about describing **statistical** relations between variables, a response Y and predictors $X_1, \ldots$ X_{p-1}. Specifically, it is about using **regression models** to describe **statistical** relations. Regression models are a formal means of expressing the two essential ingredients of a statistical relation:

1. A tendency of the response variable Y to vary with the predictor variables X_1, …, X_{p-1}
2. A scattering of points around the curve of the statistical relationship

Equivalently, a regression model assumes:

(a) There is a probability distribution of the response Y for each set of levels or values of X_1, …, X_{p-1}
(b) The means of these probability distributions vary in some systematic fashion with X_1, …, X_{p-1}

The systematic relationship between the means of the probability distributions for Y given each set of levels or values of X_1, …, X_{p-1} is a **functional relation**, meaning that there is a function of the predictors $g\left(X_1, \ldots, X_{p-1}\right)$ that satisfies

$$E\left\{Y\right\} = g\left(X_1, \ldots, X_{p-1}\right).$$

When we have one predictor ($p = 2$), we call $g\left(X_1\right) = g\left(X\right)$ the **regression function** or the **regression curve**. When we have more than one predictor, we call $g\left(X_1, \ldots, X_{p-1}\right)$ the **regression surface**.