Lab 9

Maria Barouti

It is rarely the case that a dataset will have a single predictor variable. It is also rarely the case that a response variable will only depend on a single variable. So in this lab, we will extend our current linear model to allow a response to depend on *multiple* predictors. For this Lab we will discuss a dataset with information about cars. This dataset, which can be found at the UCI Machine Learning Repository contains a response variable `mpg` which stores the city fuel efficiency of cars, as well as several predictor variables for the attributes of the vehicles.

```r
# read the data from the web
autompg = read.table(
  "http://archive.ics.uci.edu/ml/machine-learning-databases/auto-mpg/auto-mpg.data",
  quote = "\"",
  comment.char = "",
  stringsAsFactors = FALSE)
# give the dataframe headers
colnames(autompg) = c("mpg", "cyl", "disp", "hp", "wt", "acc", "year", "origin", "name")
# remove missing data, which is stored as "?"
autompg = subset(autompg, autompg$hp != "?")
# remove the plymouth reliant, as it causes some issues
autompg = subset(autompg, autompg$name != "plymouth reliant")
# give the dataset row names, based on the engine, year and name
rownames(autompg) = paste(autompg$cyl, "cylinder", autompg$year, autompg$name)
# remove the variable for name, as well as origin
autompg = subset(autompg, select = c("mpg", "cyl", "disp", "hp", "wt", "acc", "year"))
# change horsepower from character to numeric
autompg$hp = as.numeric(autompg$hp)
# check final structure of data
str(autompg)
```

We will focus on using two variables, `wt` and `year`, as predictor variables. That is, we would like to model the fuel efficiency (`mpg`) of a car as a function of its weight (`wt`) and model year (`year`). To do so, we will define the following linear model,

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i, \qquad i = 1, 2, \ldots, n$$

where $\epsilon_i \sim N(0, \sigma^2)$. In this notation we will define:

- X_{i1} as the weight (`wt`) of the ith car.
- X_{i2} as the model year (`year`) of the ith car.

The picture below will visualize what we would like to accomplish. The data points (X_{i1}, X_{i2}, Y_i) now exist in 3-dimensional space, so instead of fitting a line to the data, we will fit a plane.

Task 1

Use the `lm` function and provide estimates for b_0, b_1, b_2.

Task 2

Obtain estiates for b_0, b_1, b_2 using

$$\mathbf{b} = (\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{Y})$$

Do estimates in Task 2 agree with the estimates obtained from Task 1?

Task 3

Often we will be interested in the square root of MSE which is given by

$$RMSE = \sqrt{\frac{SSE}{n-p}}.$$

Calculate $RMSE$ using `residuals` function from `R` as well as using the vector notation of the residuals.

And we can now verify that our math above is indeed calculating the same quantities.

Task 4

Calculate R^2. How do you interpret the result?