ML01 – Spring 2019
Lab 5: Linear regression

1 `Prostate` data

We consider again the `prostate` (see Lab 2). The goal is to predict the log of PSA (`lpsa`) from a number of measurements including log cancer volume (`lcavol`), log prostate weight `lweight`, `age`, log of benign prostatic hyperplasia amount `lbph`, seminal vesicle invasion `svi`, log of capsular penetration `lcp`, Gleason score `gleason`, and percent of Gleason scores 4 or 5 `pgg45`.

1. Apply linear regression to these data, with `lpsa` as the response variables. Why coefficients are significantly non-zero ?
2. Plot the fitted values $\widehat{y}_i$ versus the observed values y_i.
3. Plot the residuals $\widehat{\epsilon}_i$ versus the observed values y_i.
4. Estimate the coefficients using the training data (train=TRUE), and predict the value of `lpsa` for the test data. Plot the predicted values vs. the observed values and compute the mean squared error (MSE).

2 `Calcium` data

The `calcium` data were obtained by Howard Grimes from the Botany Department, North Carolina State University, who conducted an experiment for biochemical analysis of intracellular storage and transport of calcium across plasma membrane. Cells were suspended in a solution of radioactive calcium for a certain length of time and then the amount of radioactive calcium that was absorbed by the cells was measured. The experiment was repeated independently with 9 different times of suspension each replicated 3 times.

1. Apply linear regression to these data, with `cal` as the response variables.
2. Plot the data with the LS line.
3. Plot the residuals $\widehat{\epsilon}_i$ as a function of time, and as a function of calcium. What do you observe ?

4. Try to find a better model using some transformations of the input variable `time`.

3 Simulation

We consider the following model :

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i, \quad i = 1, \ldots, n$$

where $\epsilon \sim \mathcal{N}(0, (0.5)^2)$, with $\beta_0 = 0$, $\beta_1 = 0.1$, $\beta_2 = 1$.

1. Generate $n = 50$ vectors $x_i = (x_{i1}, x_{i1})$ uniformly in $[0, 1]^2$. Keeping these input values fixed, generate $N = 1000$ datasets of size $n = 50$.
2. We reject hypothesis $H_{j0} : \beta_j = 0$ if the p-value for parameter β_j is less than 0.05. For which proportions of the datasets do we reject H_{00}, H_{10} and H_{20} ?