Tutorial 1

Q1. Install `R` on your computer, which can be downloaded from here. Unfortunately, the default interface to `R` is extremely cumbersome and painstaking to use. Instead, you should use `RStudio` as your interface.

Q2. Use `R` to generate $N = 1000$ random variables $x_1, \ldots, x_N \overset{\text{iid}}{\sim} \mathcal{N}(0, 5^2)$. Then, with parameters $\beta_0 = 1$, $\beta_1 = 2$, $\sigma = 3.5$, generate data from the linear model

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \qquad \epsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2). \tag{1}$$

i. Calculate the MLE's $\hat{\beta}_0$ and $\hat{\beta}_1$.

ii. Plot a scatterplot of the data and superimpose the estimated regression line.

iii. Assuming that $\sigma = 3.5$ is known, calculate a 92% confidence interval for β_1.

Q3. Let's see some of the theory in practice. For $M = 10000$, generate a data matrix $Y_{N\times M}$, where each column is drawn from the linear model (1), each with the same vector $x = (x_1, \ldots, x_N)$ generated in Q2.

i. Calculate $\hat{\beta}_1$ for each column of Y, and plot the histogram of these $\hat{\beta}_1$. Superimpose the theoretical distribution $\hat{\beta}_1 \mid x, \theta \sim \mathcal{N}(\beta_1, \sigma^2/S_{xx})$ onto this histogram.

ii. Assuming σ is known, calculate a 92% confidence interval for β_1 from each column of Y. Out of the $M = 10000$ intervals, what fraction of them cover the true value $\beta_1 = 2$?

iii. Let $Y^{(10)}$ denote the first 10 rows of Y. Assuming σ is unknown, calculate $\hat{\beta}_0$, $\hat{\beta}_1$, and

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{10}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2}{8}$$

for each column of $Y^{(10)}$. Calculate the sample mean of $\hat{\sigma}^2$ and compare it to the true value $\sigma^2 = 3.5^2$.

Q4.

i. Show that

$$S_{xx}^{(n)} = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}.$$

ii. Show that $S_{xx}^{(n)} < S_{xx}^{(n+1)}$. Therefore, since $\mathbb{E}[\hat{\beta}_1 \mid x, \theta] = \beta_1$ and $\text{var}(\hat{\beta}_1 \mid x, \theta) = \sigma/S_{xx}$, the quantity $\hat{\beta}_1$ is becoming a better and better estimate of β_1 as n increases.