STAT GU4205/GR5205 Assignment 1 [20 pts]
Due 11:59pm Saturday, September 21st

Please scan the written part (if you do not want to use Word or LaTeX), and submit your assignment on Gradescope. For coding problems, your `R` code is also required.

Problem 1 (1.22 KNN), 8 points

Sixteen batches of the plastic were made, and from each batch one test item was molded. Each test item was randomly assigned to one of the four predetermined time levels, and the hardness was measured after the assigned elapsed time. The results are shown below; X is the elapsed time in hours, and Y is hardness in Brinell units. Assume the first-order regression model (1.1) is appropriate (**model (2.1) in the notes**).

Data not displayed

Data can be found in `HW_1_Problem_1_Data.R`. Perform the following tasks:

1. Use `R` to obtain the estimated regression function.
2. Use `R` to create a scatter plot with the line of best fit. Make the line of best fit red.
3. Use `R` to calculate the best point estimate of σ^2.
4. Use `R` to calculate the sample correlation coefficient and coefficient of determination.

Problem 2: Estimates of variance, 12 points

X_i's are one-dimensional random variables. Assume that

$$X_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2), \qquad \text{for } i = 1, 2, \text{ and } 3.$$

1. Given 3 observations (x_1, x_2, x_3), what is the maximize likelihood estimator of (μ, σ^2) (note that we want to get the MLE of σ^2 instead of σ)?
2. Denote the MLE of σ^2 as $\widehat{\sigma}^2_{\text{MLE}}$. Is the MLE unbiased, i.e., does $\mathbb{E}(\widehat{\sigma}^2_{\text{MLE}}) = \sigma^2$ hold? To answer this question, we first try to conduct a simulation in `R`. Generate 5000 data sets, with each of them contains three independent observations following a one-dimensional normal distribution with mean $\mu = 2$ and variance $\sigma^2 = 9$. Denote the data sets as

$$\begin{aligned} \mathcal{X}^{(1)} &= (x_1^{(1)}, x_2^{(1)}, x_3^{(1)}), \\ \mathcal{X}^{(2)} &= (x_1^{(2)}, x_2^{(2)}, x_3^{(2)}), \\ &\vdots \\ \mathcal{X}^{(5000)} &= (x_1^{(5000)}, x_2^{(5000)}, x_3^{(5000)}). \end{aligned}$$

Based on each data set $\mathcal{X}^{(b)}$, calculate the MLE $(\widehat{\sigma}^2_{\text{MLE}})^{(b)}$. The average

$$\frac{1}{5000}\sum_{b=1}^{5000}(\widehat{\sigma}^2_{\text{MLE}})^{(b)}$$

is supposed to be a very close approximation to $\mathbb{E}(\widehat{\sigma}^2_{\text{MLE}})$. Is the average close to σ^2?

3. Calculate $\mathbb{E}(\widehat{\sigma}^2_{\text{MLE}})$ explicitly. Is it unbiased?

4. If in each sample there are n observations, then what is $\mathbb{E}(\widehat{\sigma}^2_{\text{MLE}})$? Based on this result, can you find a new estimator which is unbiased for σ^2, i.e., find a new estimator $\hat{\sigma}^2$, such that $\mathbb{E}(\hat{\sigma}^2) = \sigma^2$?

5. If we do not assume that X_i's are normally distributed, but still assume that they are independent and identically distributed random variables, based on n observations $(x_1, \ldots, x_n)$, what is an unbiased estimate of $\text{Var}(X_1)$?