# Homework Assignment 5
Due Friday, October 14, 2022 at 5PM
Total points: 92

Please email your answer (compiled pdf file from R markdown) and R code to Yen-Yi Ho (hoyen@stat.sc.edu). You can handwrite the solution for Q12 and combine with Q1-Q11 into one single pdf file.

Instructions: feel free to discuss the homework with other students. However, each student must conduct their own analyses and write-up their own solutions. Write as if for a scientific journal. Be brief and accurate.

Use the WHO Child Growth Standards (IGROWUP) data for child age **0-5** to study the dependence of weight on age with and without adjustment for height. IGROWUP data is available http://people.stat.sc.edu/hoyen/Stat704/Data/survey.csv.
More information about IGROWUP data can be found in http://www.who.int/childgrowth/en/

1. Describe the data set, how many observations, missing values, univariate analysis. (3 points)
2. Plot weight against age. Label the axes clearly and make sure that all observations can be seen. (3 points)
3. Fit the simple linear regression of weight on age. Add the least squares line on the scatter-plot. (5 points)
4. Write your own lm function in R to perform the simple linear regression analysis in (3) (need to return **intercept, slope and MSE)**. (10 points)
5. Write down the regression model in (3), include all of the key assumptions of the model. (5 points)
6. List the major assumptions of the simple linear regression and comment on how reasonable each one is in light of the graph above or your understanding of the data set. (6 points)
7. Fit separate linear regression models for babies aged between **0-1**, and children aged **1-6**. Compare the results from the two regression models. (5 points)
8. In a few sentences interpret the **intercept, slope and residual standard deviation**. Include the numeric values and confidence intervals for the two regression models in (7). Write it as if for a public health audience; do not use statistical jargon. (15 points)
9. Create a new variable that indicates the decile of height for each participant. On one page, display a separate plot of weight against age for the first, fourth, seventh, and tenth decile of height. By "stratifying on height", you are examining the association of weight and age among children of a similar height. (5 points)

10. Fit a simple regression of weight on age for each height decile (D1-D10). Complete the table below. (10 points)

| | Data set | Intercept | Age Slope ($b_1$) | se($b_1$) | Residual std dev |
|---|---|---|---|---|---|
| | All data | | | | |
| Height | D1 | | | | |
| | D2 | | | | |
| | D3 | | | | |
| | D4 | | | | |
| | D5 | | | | |
| | D6 | | | | |
| | D7 | | | | |
| | D8 | | | | |
| | D9 | | | | |
| | D10 | | | | |
| | Mean of D1-D10 | | | -- | -- |

11. Propose a test statistic that can be used to check the assumption that the weight-age slope is the same in each decile of height. (5 points)
12. Under a simple linear regression model, $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, var($\epsilon_i$)= $\sigma^2$, $i = 1, 2, \ldots, n$.
    (a) Show that the sample correlation between the $\widehat{\epsilon_i}$'s and $\widehat{y_i}$'s is zero. What is the sample correlation between $\widehat{\epsilon_i}$'s and $y_i$'s? (10 points)
    (b) Suppose each $x_i$ is replaced by $x_i + \delta_i$, where $\delta_i$'s (covariate errors) are mutually independent identically distributed with mean zero. Assume that the covariate error distribution does not depend on the $x_i$'s and $\epsilon_i$'s. How are the least square estimates of $\beta_0$ and $\beta_1$ affected? (10 points)