# BIOS 507 HOMEWORK 4

## Due 3/17/2025 by 11:59pm

Directions: Complete all questions. Any R or SAS code used should be attached at the end of the homework. Collaboration is encouraged, but the final product must be your own work.

# Problem 1

Researchers collected a random sample of data on infants' birth weights ($Y$, lbs), gestation period ($X_1$, weeks), and a variable whose value is the number of the letter of the alphabet the baby's last name starts with (A =1, B=2, C=3, etc). Treat $X_2$ as a quantitative variable. The data set can be found on Canvas as `birth_weight.txt`.

1. Run a regression of $Y$ on $X_1$ and $X_2$. Is the improvement due to the additional of $X_2$ (to a model already including $X_1$ significant? Use $\alpha = 0.05$ for the test.

2. Is the previous result surprising to you? Why or why not?

3. Calculate the square of the partial correlations between $Y$ and $X_1$ given $X_2$ ($R^2_{Y,X_1|X_2}$) and between $Y$ and $X_2$ given $X_1$ ($R^2_{Y,X_2|X_1}$)

4. Run two different regression models:

   (a) A simple linear regression of $Y$ onto $X_2$

   (b) A simple linear regression of $X_1$ onto $X_2$

   Then, produce a plot of the two sets of residuals against each other (this plot will have the residuals from model (a) on the y-axis and the residuals from model (b) on the x-axis).

5. Next, run a simple linear regression of the residuals from model (a) against the residuals from model (b) obtained above. What is the estimated slope and the $R^2$ for this simple linear regression? Also, what regression coefficient and squared partial correlation coefficient from your work with the full model in part 1 do these quantities correspond to? Why does this connection make sense?

# Problem 2

A researcher is investigating how BMI (Y) depends on age ($X_1$, years) and healthy diet score ($X_2$). The higher the diet score, the healthier the daily diet. The researcher has developed the following 2 regression models:

**Model 1** . $\hat{Y} = 18 + 0.2X_1 - 0.1X_2$

**Model 2** . $\hat{Y} = 26 + 0.15X_1 - 0.07X_2 - 0.02X_1X_2$

Both models have been built using a data set with data ranges: $20 \leq X_1 \leq 50$ and $1 \leq X_2 \leq 10$

1. Using both models, write down the predicted value of BMI when $X_2 = 1$. Note that this will be a function of $X_1$ - you will not have a single answer. Do the same thing for $X_2 = 5$. Comment on the effect of the interaction term in model 2. **Bonus:** Generate plots under both models for the predicted value of BMI when $X_2 = 1$ as a function of age.

2. Find the expected change in BMI for a one year increase in age for model 1 when $X_2 = 5$. Does this quantity depend on the specific value of $X_2$? Why or why not?

3. Find the expected change in BMI for a one year increase in age for model 2 when $X_2 = 5$. Does this quantity depend on the specific value of $X_2$? Why or why not?

# Problem 3

Estimated glomerular filtration rate (GFR) measures the level of kidney function. It can be used to determine the stage of kidney disease. GFR is calculated primarily from the results of blood creatinine test. An investiagor is interested in the association between GFR (response) and age, sex, race, and BMI in patients with coronary artery disease. She plans to look at the association in a study cohort of 366 patients aged 19-90 years. This dataset can be found on Canvas under `GFR.txt`. The following variables are relevant to this analysis:

- `ID`: The patient ID
- `BL_GFR`: The estimated glomerular filtration rate
- `Age`: the patient age in years
- `Male`: male sex, Male = 1 if male and 0 if female
- `Black`: race, Black = 1 if black and 0 if non-black
- `BMI`:body mass index ($kg/m^2$)
- `BMIcat`: BMI categories based on BMI cut points 25 and 30

1. Conduct exploratory data analyses for the variables that the researcher is interested in (outcome and predictors). Provide any relevant plots or tables.
2. Test whether the variable `BMIcat` is contributing significantly to GFR given age, sex, and race. Report the terms in the full and reduced models, the degrees of freedom of the test, the test statistic, p-value, and your conclusions.
3. Conduct residual analyses for whichever model you decide to use based on part (2). Report and relevant plots and comment on the validity of your assumptions based on the plots.
4. The investigator wants to know whether the association between GFR and age is being modified by race, that is, whether the association is different in blacks and non-blacks. Use appropriate modeling and conduct the corresponding tests. Show all models that you fit and clearly state your hypotheses. Report your findings.