Here are some notes which may be helpful in completing Section 4 of Homework #1.

The idea is that often there is more than one way to approach data. We'd like to make statements about how the approaches compare. I find use of qualitative and subject terms like "better" to be ambiguous.

When it comes to estimating a quantity, it's helpful to discuss *bias* and *efficiency* of the approach to estimation. To define these ideas, I want to appeal to the idea of how an estimation approach would perform under the hypothetical scenario of *repeated sampling*. Anytime we analyze data we are doing it from a single sample from a (sometime hypothetical) population. It is helpful to think about how a procedure would perform if we drew repeated independent samples from the population and analyze the data repeatedly.

In your homework you're generating data with known properties and with a known treatment effect beta = -0.15. Examining how the methods perform over the long run when the answer is know gives us insights into how the methods compare. No method is going to get the "right" coefficient (-0.15) every time. You'll see if the average of the repeated coefficients is identical to the underlying value (-0.15). The bias of a method is defined as the

bias = mean of coefficients from repeated sample - true value of parameter

A method is considered unbiased if the bias is negligible. I would say that a value of 10% (which is +/- 0.10*0.15 = +/- 0.015) of the parameter is basically negligible.

You can also see how methods compare in terms of their precision. This is done by comparing the ratio of the variances. If Q7, I ask you to calculate the relative efficiency of 3 of the methods compared to GEE. For this, you calculate the variance of the coefficients (ttest, paired, reg, gee). The relative efficiency of analysis 1 in the HW compared to the GEE analysis is

variance of variable "ttest" / variance of variable "gee"

This number how much you have to increase the sample size in your data so that analysis 1 has a much power as the GEE analysis. You are going to see in Questions 9 and 10 that this answer will depend of the correlation in your data "rho".

Finally, I ask you to calculate the power of the techniques. The power is the percentage of simulated dataset which have a $p < 0.05$ -- which is given by the variables power-ttest, power-paired, power-reg and power-gee.

Please feel free to contact me with questions. If my explanations hear remain unclear, I'll redact your question and copy the class with any additional clarification.