Stat 471/571: How to build a model - part 1

Here is the process I use to figure out an appropriate model for a study. It takes the form of a series of questions. Many of the recent homework questions have led you through these questions, but they are set out here from start to finish.

These are described in terms of a randomized experiment. Modifications for an observational study are at the end.

My basic approach is to figure out the treatment structure, the experimental design, and how the two go together. Reminder:

- Treatment structure: What is done to an experimental unit?
- Experimental design: How treatments are randomly assigned to experimental units.

1. Figure out the treatment structure.
    - This may be just a collection of treatments (one-way) or some combination of factors that have structure (two-way, three-way, $\cdots$).
    - If structured, is it a complete factorial? i.e., are all combinations used at least once?
    - If not complete, what questions are to be answered? Probably want to set up as a one-way set of treatments and use contrasts to answer questions.
2. Figure out the experimental design.
    - What is the experimental unit? i.e., what "thing" is randomized to a treatment?
    - Is there more than one experimental unit?
    - If yes, are they nested? i.e. is there a larger and a smaller eu? The larger is the main plot eu; the smaller is the split plot eu.
    - Nested eu's are the most common. If they're not nested, what is the relationship between the two eu's (e.g. are they crossed as in a strip plot design)?
    - Is there any form of blocking?
    - Two possible ways to identify blocking
        - treatments randomized "within" a block
        - block = collection of "similar" eu's.
            * Is there anything in the study design that will affect multiple eu's in a similar way?
            * E.g., 10 samples processed on day 1, 10 on day 2, 10 on day 3
              Are there reasons why day 1 may differ from day 2 or day 3? If so, treat days as blocks
    - Most often, each treatment (or treatment factor) applied to only one eu within each block.

- If blocked, are the blocks complete? i.e., is each treatment used at least once in each block. Complete blocks are the most common.
  No issues with incomplete blocks unless each block only received one treatment

3. What is the relationship between the treatment structure and experimental design?

   - Will be simple when only one size of eu
     - Not so simple when incomplete blocks - is this happenstance or planned?
   - When multiple sizes of eu, what treatments or treatment factors are assigned to each eu?
     - i.e., which treatments belong at the main plot level, which at the split plot level?

4. Include baseline values as a continuous covariate or (perhaps) use a change score of some form, e.g., difference, log ratio, as the response.

5. Assemble the model

   - Treatments are fixed effects
   - Experimental units are random effects
   - Blocks may be fixed or random. I prefer fixed blocks, but others prefer random blocks.
   - I do use random blocks when:
     - there is a clearly defined population of blocks, e.g. farms or schools,
     - and you want to learn about the variability between whatever makes the blocks,
     - and you have more than 5-10 blocks, so you get a good estimate of the variance
     - or when you want to recover between-block (interblock) information from incomplete blocks
   - My practice is to include all interactions.
     This fits a unique mean to each cell (combination of treatments) and is equivalent to using contrasts of cell means, without having to write out all the contrast coefficients.

6. Figure out the df for each term.

   - Not necessary, but comparison to output provides a check that you specified your intended model.
   - Main effects have $k - 1$ df, where k is the number of levels
   - Interaction df is the product of main effect df. This presumes that both (or all) main effects and lower level interactions are in the model.
   - Best way to figure out df for error terms is to work backwards
     - total df for one level of the design is # obs - 1
     - add up df for all terms "above" that error, including treatments at that level
     - error df is the difference.