

1. Suppose we are running an experiment with 3 experimental conditions. We want to examine how long it takes to boil a cup of water. We want to know if there is an effect if we use a wide-bottomed pot or a narrow-bottomed pot, put the lid on the pot or not, or use water with no sodium vs. water with a 3.5% concentration of sodium (which is like the ocean).
  - a. Write down the design matrix  $\mathbb{X}$  for the “complete factorial design”, which would correspond to a single observation at each experimental condition. Code the “lower” level of each explanatory variable as a -1 and the “upper” level as a 1.
  - b. Demonstrate that the matrix  $\mathbb{X}$  is orthogonal in the sense that  $\mathbb{X}^T \mathbb{X} = nI$
  - c. We can accomplish the same thing as b. using the following SAS code. Include the output to the code as the answer and verify that you get the same answer as in b.
  - d. Using SAS, show that the coefficient estimates for the main effects model is the same for the main effects model and the model with all 2 way and 3 way interactions.
  - e. How is this different from a multiple regression in general and a multicollinear multiple regression in particular?
  - f. Bonus points for running this experiment and analyzing it (I’ve included some code to get you started). Include your data, SAS output, and a picture of you performing the experiment.

[Note that a single observation at each experiment condition only lets us estimate the full interaction model, not compute any uncertainty measures (hypothesis tests or confidence intervals). We can estimate these quantities if we remove the 3-way interaction].

2. A school district is designing a multiple regression study looking at the effect of gender, family income, mother’s education and language spoken in the home on the English language proficiency scores of Latino high school students. The variables gender and family income are control variables and not of primary research interest. Mother’s education is a continuous research variable that measures the number of years that the mother attended school. The range of this variable is expected to be from 4 to 20. The variable language spoken in the home is a categorical research variable with three levels: 1) Spanish only, 2) both Spanish and English, and 3) English only. Since there are three levels, it will take two dummy variables to code language spoken in the home.

We want to examine mother’s education and language spoken at home variables. But, before that, we want to know how many observations to gather **or**, equivalently, the power of our experiment for a given sample size.

- a. Define power using notation and using words.

We can use the SAS function **PROC POWER** with the **multreg** option for this purpose (<https://support.sas.com/documentation/cdl/en/statug/63962/HTML/default/viewer.htm>)

[m#statug\\_power\\_sect004.htm](#) for details of the procedure). If you recall about PROC POWER in the context of t-tests, we need to supply specific values (or ranges of values) for various quantities (see 3c\_powerDarren). We will get back to the difference between “fixed” and “random” models, but for now just write `model = fixed`. Specify the total number of explanatory variables.

- b. Now, we will specifically address the questions of interest: mother’s education. Specify the number of explanatory variables we want to test. We need to come up with values for  $R^2$  for the full model and the model without the explanatory variable(s) being tested. The best way to do this is via a preliminary experiment or using a separate experiment that should have approximately the same types of relationships. In this case, we will choose .5 for the full model and look at how many observations we will need to acquire a fixed power of 0.8 for a difference in  $R^2$  of the model with and without mother’s education for a grid of values from 0.02 to 0.2 in steps of 0.01. Include the table with the recommended sample sizes here. (We will return to the “Type III F-test” in the next section. In this context it means the same as the usual t-test)

What is the sample size needed to detect a difference in  $R^2$  of 0.04 at a power of (approximately) 0.8?

What is the sample size needed to detect a difference in  $R^2$  of 0.2 at a power of (approximately) 0.8?

What is the sample size needed to detect a difference in  $R^2$  of 0.1 at a power of (approximately) 0.95?

- c. Now, we will turn to the other question of interest: language at home. Adjust the above program, including the fact that language at home is given by two explanatory variables for two levels of the categorical variable. Include the table for the recommended sample sizes here.

What is the sample size needed to detect a difference in  $R^2$  of 0.04 at a power of (approximately) 0.8?

What is the sample size needed to detect a difference in  $R^2$  of 0.2 at a power of (approximately) 0.8?

What is the sample size needed to detect a difference in  $R^2$  of 0.1 at a power of (approximately) 0.95?

3. Consider the ingestion rates and organic consumption percentages of “deposit feeders” (the details about what this means isn’t really important for our purposes). There are two types of feeders: single valve and bivalve. The researcher wants to see if ingestion rate is associated with the percentage of organic matter in food, after accounting for animal weight. However, the research is unsure whether to include bivalve in the analysis. Analyze the data to answer this question of interest. Be methodical and thoughtful when going through this exercise.