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.

$$X = \begin{bmatrix} Intcerpt & x1 & x2 & x3 & x1*x2 & x1*x3 & x2*x3 & x1*x2*x3 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & -1 & 1 & -1 & -1 & -1 \\ 1 & 1 & -1 & 1 & -1 & 1 & -1 & -1 \\ 1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 \\ 1 & -1 & 1 & 1 & -1 & -1 & 1 & -1 \\ 1 & -1 & 1 & -1 & -1 & 1 & -1 & 1 \\ 1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 \\ 1 & -1 & -1 & -1 & 1 & 1 & 1 & -1 \end{bmatrix}$$

   b. Demonstrate that the matrix $\mathbb{X}$ is orthogonal in the sense that $\mathbb{X}^T\mathbb{X} = nI$

$$X^T = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 \\ 1 & 1 & -1 & -1 & 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 \\ 1 & -1 & 1 & -1 & -1 & 1 & -1 & 1 \\ 1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 \\ 1 & -1 & -1 & 1 & -1 & 1 & 1 & -1 \end{bmatrix}$$

$$X^TX = \begin{bmatrix} 8 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 8 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 8 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 8 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 8 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 8 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 8 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 8 \end{bmatrix} = 8 * \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

   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.

```
proc glm;
   model y = x1| x2 |x3 / SOLUTION XPX;
RUN;
```

The X'X Matrix									
	Intercept	x1	x2	x1*x2	x3	x1*x3	x2*x3	x1*x2*x3	y
Intercept	8	0	0	0	0	0	0	0	965
x1	0	8	0	0	0	0	0	0	-21
x2	0	0	8	0	0	0	0	0	17
x1*x2	0	0	0	8	0	0	0	0	-5
x3	0	0	0	0	8	0	0	0	-29
x1*x3	0	0	0	0	0	8	0	0	13
x2*x3	0	0	0	0	0	0	8	0	-1
x1*x2*x3	0	0	0	0	0	0	0	8	-51
y	965	-21	17	-5	-29	13	-1	-51	116949

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.

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	120.6250000	3.30482602	36.50	<.0001
x1	-2.6250000	3.30482602	-0.79	0.4715
x2	2.1250000	3.30482602	0.64	0.5552
x3	-3.6250000	3.30482602	-1.10	0.3343

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	120.6250000	6.37500000	18.92	0.0336
x1	-2.6250000	6.37500000	-0.41	0.7513
x2	2.1250000	6.37500000	0.33	0.7952
x3	-3.6250000	6.37500000	-0.57	0.6708
x1*x2	-0.6250000	6.37500000	-0.10	0.9378
x1*x3	1.6250000	6.37500000	0.25	0.8411
x2*x3	-0.1250000	6.37500000	-0.02	0.9875

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	120.6250000	.	.	.
x1	-2.6250000	.	.	.
x2	2.1250000	.	.	.
x1*x2	-0.6250000	.	.	.
x3	-3.6250000	.	.	.
x1*x3	1.6250000	.	.	.
x2*x3	-0.1250000	.	.	.
x1*x2*x3	-6.3750000	.	.	.

e. How is this different from a multiple regression in general and a multicollinear multiple regression in particular?
The complete factorial design ensures the design matrix is orthogonal, which is not guaranteed is multiple regression, and violated in multicollinear multiple regression. Orthogonal design matrix makes sure least square estimates of parameters does not depend on whether other terms are included in the model.

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.
Power is the probability of rejecting a null hypothesis when it is indeed false. $Power = Pr(Reject\ H_0 | H_a)$. Note that power requires a specific form for the alternative hypothesis (that is, how violated is the null hypothesis?)

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#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)

```
PROC POWER;
MULTREG
model = fixed
nfullpredictors = 5
ntestpredictors = 1
rsquarefull = 0.5
rsquarediff = 0.02 to 0.2 by 0.01
ntotal = .
power = 0.8;
RUN;
```

The POWER Procedure
Type III F Test in Multiple Regression

Fixed Scenario Elements	
Method	Exact
Model	Fixed X
Number of Predictors in Full Model	5
Number of Test Predictors	1
R-square of Full Model	0.5
Nominal Power	0.8
Alpha	0.05

Computed N Total			
Index	R-square Diff	Actual Power	N Total
1	0.02	0.802	199
2	0.03	0.801	133
3	0.04	0.803	101
4	0.05	0.802	81
5	0.06	0.803	68
6	0.07	0.806	59
7	0.08	0.806	52
8	0.09	0.802	46
9	0.10	0.805	42
10	0.11	0.801	38
11	0.12	0.812	36
12	0.13	0.806	33
13	0.14	0.808	31
14	0.15	0.806	29
15	0.16	0.800	27
16	0.17	0.807	26
17	0.18	0.812	25
18	0.19	0.815	24
19	0.20	0.815	23

What is the sample size needed to detect a difference in R^2 of 0.04 at a power of (approximately) 0.8? **101**
What is the sample size needed to detect a difference in R^2 of 0.2 at a power of (approximately) 0.8? **23**
What is the sample size needed to detect a difference in R^2 of 0.1 at a power of (approximately) 0.95? **42**

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.

```
PROC POWER;
MULTREG
model = fixed
nfullpredictors = 5
ntestpredictors = 2
rsquarefull = 0.5
rsquarediff = 0.02 to 0.2 by 0.01
ntotal = .
power = 0.8;
RUN;
```

The POWER Procedure
Type III F Test in Multiple Regression

Fixed Scenario Elements	
Method	Exact
Model	Fixed X
Number of Predictors in Full Model	5
Number of Test Predictors	2
R-square of Full Model	0.5
Nominal Power	0.8
Alpha	0.05

Computed N Total			
Index	R-square Diff	Actual Power	N Total
1	0.02	0.800	244
2	0.03	0.801	164
3	0.04	0.802	124
4	0.05	0.802	100
5	0.06	0.803	84
6	0.07	0.806	73
7	0.08	0.804	64
8	0.09	0.801	57
9	0.10	0.804	52
10	0.11	0.808	48
11	0.12	0.805	44
12	0.13	0.806	41
13	0.14	0.802	38
14	0.15	0.805	36
15	0.16	0.805	34
16	0.17	0.802	32
17	0.18	0.810	31
18	0.19	0.800	29
19	0.20	0.805	28

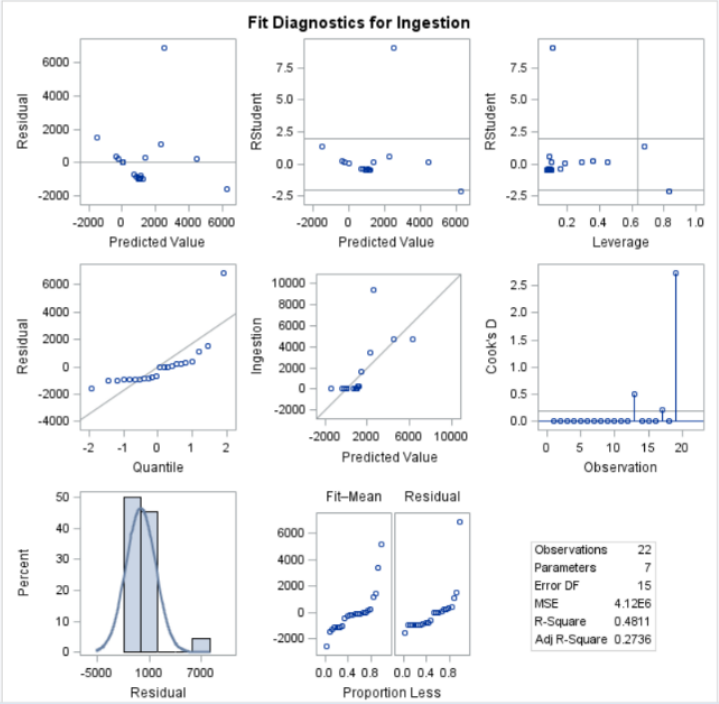
What is the sample size needed to detect a difference in R^2 of 0.04 at a power of (approximately) 0.8? **124**
What is the sample size needed to detect a difference in R^2 of 0.2 at a power of (approximately) 0.8? **28**
What is the sample size needed to detect a difference in R^2 of 0.1 at a power of (approximately) 0.95? **81**

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 matte 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.

The fully interactive model is:

$$\mu\{ingestion|organic, weight, bivalve\} = \beta_0 + \beta_1 organic + \beta_2 weight + \beta_3 bivalve + \beta_3 organic * weight + \beta_4 organic * bivalve + \beta_5 weight * bivalve + \beta_6 organic * weight * bivalve$$

It produced the following output:



Assumptions:

Normality: suspect, especially because of small sample size

Variance: some evidence of changing variance

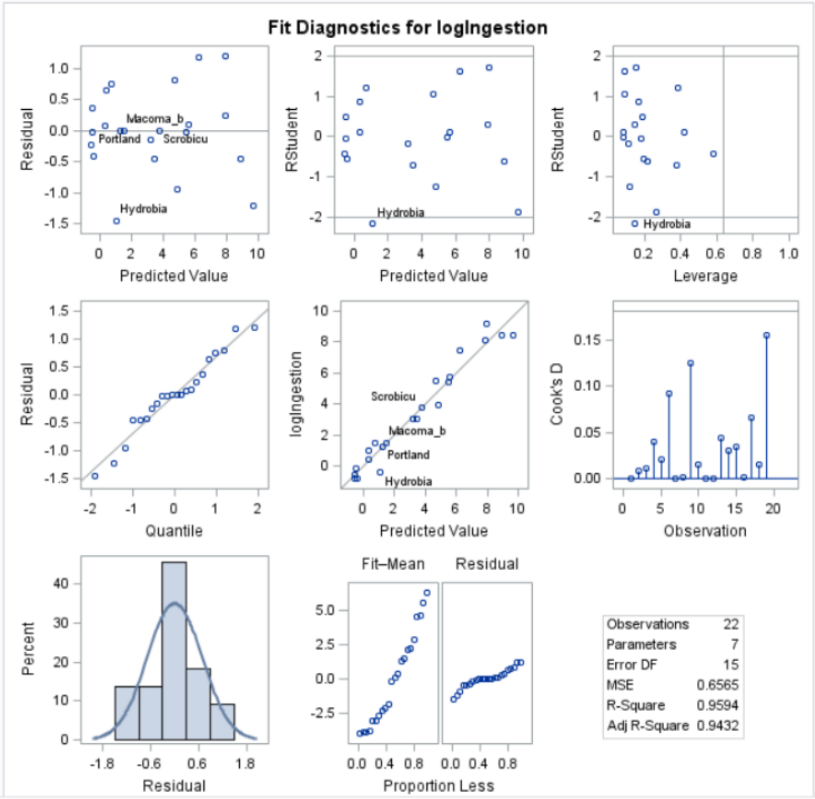
Influential points: There are a couple of potential influential points

Independence: no reason to believe samples are not independent and we're accounting for interactions in the model

Since assumptions were not met, I looked at the data and noticed that both ingestion and weight had a large ratio between largest and smallest values. Therefore, I did a log transformation for these two variables.

(Transformed) fully interactive model:

$$\mu\{logingestion|organic, logweight, bivalve\} = \beta_0 + \beta_1 organic + \beta_2 logweight + \beta_3 bivalve + \beta_3 organic * logweight + \beta_4 organic * bivalve + \beta_5 logweight * bivalve + \beta_6 organic * logweight * bivalve$$



Assumptions:

Normality: satisfactory but still slightly suspect because of small sample size

Variance: no evidence of patterns or changing variance

Influential points: no concerns

Independence: no reason to believe samples are not independent and we're accounting for interactions in the model

Looking at the Interaction Terms:

Note: it is not sufficient to look at the t-tests from the inferential table. We need to do an extra sums of squares test:

Fully interactive model

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	232.7282800	38.7880467	59.08	<.0001
Error	15	9.8474743	0.6564983		
Corrected Total	21	242.5757543			

Reduced model: main effects only

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	227.7290669	75.9096890	92.03	<.0001
Error	18	14.8466874	0.8248160		
Corrected Total	21	242.5757543			

To see if we can drop the interaction terms all together, use an extra sum of squares F-test, in which fully interactive model is full model and main effects model is the reduced model.

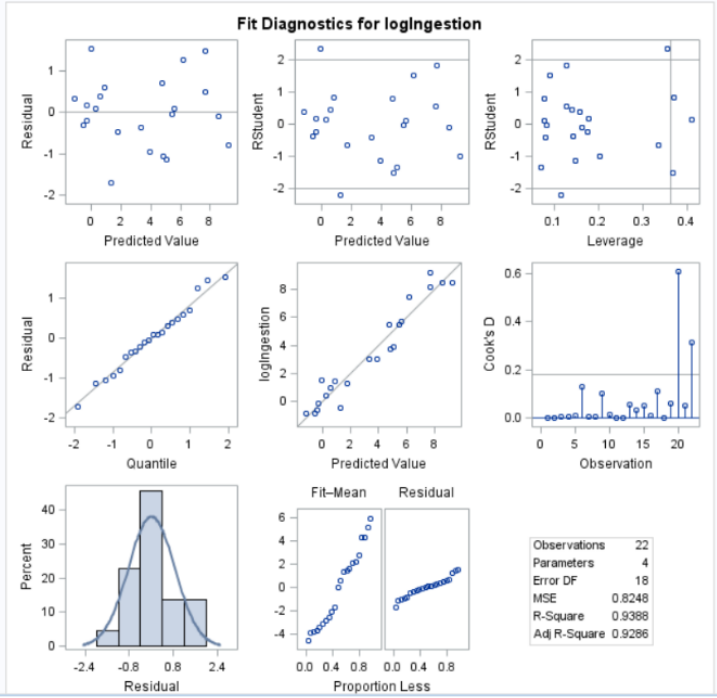
$$F = \frac{(SSE_R - SSE_F)/(df_R - df_F)}{SSE_F/df_F} = \frac{(14.8467 - 9.8475)/(18 - 15)}{0.6565} = 2.5383,$$

$$df_1 = df_R - df_F, df_2 = df_F$$

Resulting p-value is 0.0957, we fail to reject the null hypothesis that reduced model is sufficient. So the main effect model is used for next step of analysis.

Model with no interactions:

$$\mu\{logingestions|organic, logweight, bivalve\} = \beta_0 + \beta_1 organic + \beta_2 logweight + \beta_3 bivalve$$



Assumptions:

Normality: satisfactory but still slightly suspect because of small sample size

Variance: no evidence of patterns or changing variance

Influential points: not a concern

Independence: no reason to believe samples are not independent and we found no significant interactions in the previous model

Parameter	Estimate		Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1.926926280	B	0.35845077	5.38	<.0001	1.173849163	2.680003396
Organic	-0.039756977		0.00778368	-5.11	<.0001	-0.056109875	-0.023404079
logWeight	0.971732178		0.07291210	13.33	<.0001	0.818549540	1.124914815
Bivalve yes	-2.745014750	B	0.57171693	-4.80	0.0001	-3.946147448	-1.543882052
Bivalve no	0.000000000	B	.	.	.	.	.

Include Bivalve? There is evidence that the variable bivalve is associated with ingestion rate (two-sided p-value = 0.0001), given the other variables in the model. Since having a bivalve or single valve defines the type of feeder, it is reasonable to infer that it is an important variable when considering ingestion rate. Therefore, I would advise the researcher to include the bivalve variable in the model.

Statistical Conclusion: There is strong evidence that ingestion rate is associated with the percentage of organic matter in food ($p<0.0001$), after accounting for animal weight and type of feeder. We estimate that every 1 percent increase in organic matter is associated with a multiplicative change in median ingestion rate of 0.96 with a 95% confidence interval [0.95, 0.98], given the other explanatory variables in the model.

Scope of Inference: This was an observational study, so no causal inferences can be drawn. The results cannot be extended to a larger population beyond the animals in the study.