ST 512 - Lab 8 - Model Building in MLR

1. Open the `Lab8.sas` file which uses the `Cheese` data set again (from Lab 6 and Lab 7). This week we'll focus on what Type I and Type III sums of squares represent mathematically and conceptually. We'll be using SAS to do almost all of the calculations! You'll notice all the code is provided this week. I want you to focus on the concepts and not the SAS coding! Your goal is to fill in the table provided in the lab and be able to understand the relationships between the entries.

2. Begin by running the `PROC GLM` code in Step 1. This fits a SLR that predicts the taste preference (y) based on a single predictor Acetic (x_1). Use the output to fill in the first line of the table.

3. Notice that your `PROC GLM` code from Step 1 produced a data set called `STEP01`. Open this data set up and compare it to the data set you were provided. `STEP01` has an additional variable called `yx1`. This variable contains the residuals from your SLR, which we could denote as $r_{y|x_1}$. These residuals are the unexplained variation that remains, after using `Acetic` to predict `Taste`.

4. Because our new data set `STEP01` contains all our original variables, we use it to carry out the `PROC GLM` in Step 2. This fits another SLR, but this time we're trying to predict x_2 (`H2S`) by using x_1 (`Acetic`). Use the results from this SLR to fill in the second line of the table below.

5. Open the data set `STEP02` now. Compare it to the data set `STEP01` to see that we've added another column of residuals. This new column is named `x2x1` and which contain the $r_{x_2|x_1}$ values. Conceptually, this is the part of x_2 (`H2S`) that can't be explained by x_1 (`Acetic`).

6. Why would we do this? Remember, we only need to include x_2 in the model after x_1 if `H2S` helps the model fit better, after we account for `Acetic` already being in the model. What, exactly, is meant by the phrase ***account for*** that we keep using in this class? Let's find out!

7. Go ahead and examine the `PROC GLM` code included in Step 3. Notice we are still fitting a SLR, but we aren't using any of our original data! We're trying to predict one set of residuals ($r_{y|x_1}$) by using another set of residuals ($r_{x_2|x_1}$). Conceptually, what are we doing here? We're trying to see if the part of `taste` that couldn't be explained by `Acetic` ***can be explained*** by using information from `H2S` that ***also can't be explained by*** `Acetic`! If `H2S` is really a useful variable, then this regression should show us that directly!

8. Now, run the code in Step 3 and use it to fill in the third line of the table. Notice that we again are saving a new data set, this time named `STEP03`. We've also added a new variable, just like before. These are the residuals from fitting a model entirely based off of residuals. They do have a name (and their own notation) but let's not get too sidetracked yet...

9. Finally, look at the code in Step 4 which is a bit more familiar. This is the typical MLR using both `Acetic` and `H2S` to predict `taste`. Run this code and do the following:

   (a) Use the output to fill in the fourth line of the table. Are there any relationships between the four rows of your table?

   (b) Can you locate any information from Rows 1, 2, or 3 of your table in the output for Step 4?

   (c) Open up the data set `STEP04` and compare the residuals from Steps 1, 3, and 4. What do you notice?

10. We only went through the process for two of the three variables.

   (a) Can you describe what `PROC GLM`s would need to be included in the later steps if we wanted to continue this process by adding `Lactic` into the model?

   (b) If `Lactic` were added into the model, would the residuals from Steps 3 and 4 still be identical? Why or why not? List any residuals that ***would*** be identical.

Code Step	Model Type	Response	Predictor(s)	R^2	SSR
Step 1	SLR	y	x_1	0.301993	2314.14
Step 2	SLR	x_2	x_1	0.38187	50.0954
Step 3	SLR	$r_{y\|x_1}$	$r_{x_2\|x_1}$	0.4014	2147.016
Step 4	MLR	y	x_1, x_2	0.582	4461.158