Introduction to Data Science Homework 2

This question will investigate a simple regression problem. The data we will use is the salary of employees, as a function of years of experience, for 3 separate companies. (This data set is not real, but generated; the code used is provided.)

1. Read the data file and plot the salary (y) against experience (x_1) separately for each company. Then, fit linear model relating x_1 and y, given in the following equation, separately for each company.

$$y = \beta_0 + \beta_1 x_1$$

2. Define two extra input parameters x_2 and x_3 to appropriately take care of the company variable, which is a qualitative parameter with 3 possible choices. As discussed in class, there are multiple equivalent ways of doing this, so pick one that you like. Then, fit the linear model relating these parameters, given in the following equation, to the whole data set.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

3. Let's reflect on what we did in part (1) and (2). In part (1) we fitted for each company separately, which resulted in 6 parameters in total. In contrast, in part (2) we fitted all three companies together, which resulted in only 4 parameters. What is missing? Could the model in part (2) be able to capture the relationships as well as the model in part (1)? Explain what is going on.

4. Now, let's include the interaction terms $x_1 x_2$ and $x_1 x_3$ into the mix. Fit linear model including these interaction terms, given in the equation below, to the whole data set.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_{12} x_1 x_2 + \beta_{13} x_1 x_3$$

   Also, explain why the remaining second-order interaction term, $x_2 x_3$, is meaningless.

5. Notice that the model in part (4) has the same number of parameters as the model in part (1). Do they capture the same amount of information? What are the relationships between the parameters of these two models? Explain what is going on.

(Note: A separate "noiseless" data set is also provided. This is generated with identical parameters to the other data set, but with no noise. Feel free to use this, it might give a better idea to what is going on.)