An Introduction to Indicator Variables and Interaction in Multiple Regression

Suppose a field experiment was conducted using 8 plots. Two varieties of corn (A and B) were randomly assigned to the 8 plots in a completely randomized design so that each variety was planted on 4 plots. Suppose 4 amounts of fertilizer (5, 10, 15, and 20 units) were randomly assigned to the four plots in which variety A was planted. Likewise the same four amounts of fertilizer with randomly assigned to the 4 plots in which variety B was planted. Yield in bushels per acre was recorded for each plot at the end of the experiment. The data are provided below.

		Amount of
Yield	**Variety**	**Fertilizer**
134	A	5
140	A	10
146	A	15
153	A	20
138	B	5
142	B	10
145	B	15
147	B	20

1. The researchers are interested in studying the relationship between yield and amount of fertilizer for each variety of corn. Name one weakness of this experiment.

2. Sketch a scatterplot of this data. Use the symbol A to mark the points corresponding to variety A and B to mark the points corresponding to variety B.

3. The researchers would like to know if the amount of fertilizer affects mean yield for each variety of corn. One way to address this question would be to conduct a simple linear regression analysis using only the data for variety A followed by a separate simple linear regression analysis using only the data for variety B. There is a more efficient approach that requires the use of an **indicator** variable. If we let X_1 denote the variable *amount of fertilizer*, we may let X_2 be a variable that indicates the variety of corn. In particular we could set $X_2 = 1$ if the variety is A and $X_2 = 0$ if the variety is B. The values 1 and 0 for varieties A and B are arbitrary but convenient choices as we shall see later. Add a column for the indicator variable X_2 to the dataset above.

4. An **interaction** variable is formed by taking the product of two explanatory variables. Create an interaction explanatory variable X_3 by multiplying X_1 times X_2. Add a column for this interaction variable X_3 to the dataset above.

5. Consider the multiple regression of $Y =$ *yield* on X_1, X_2, and X_3. This model assumes that

$$\mu\{Y|X_1, X_2, X_3\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$

for some $\beta_0, \beta_1, \beta_2$, and β_3. Assuming this model is correct, what does $\mu\{Y|X_1, X_2, X_3\}$ look like for variety A? What does $\mu\{Y|X_1, X_2, X_3\}$ look like for variety B?

6. Why might the researchers be interested in testing $H_0 : \beta_2 = \beta_3 = 0$ vs. $H_A : H_0$ is not true?

7. The residual sum of squares for the multiple regression of Y on X_1, X_2, and X_3 is 1.30. The residual sum of squares for the simple linear regression of Y and X_1 is 28.65. Does it appear that the data can be adequately described by a single regression line that is the same for both varieties, or is there evidence that separate regression lines are needed to adequately describe the data? Conduct an appropriate test to answer this question.

8. Suppose the researchers want to know if the slope of the regression line relating mean yield to amount of fertilizer for variety A is the same as the slope of the regression line relating mean yield to amount of fertilizer for variety B. Write down an appropriate null and alternative hypothesis for a test that can be used to address this issue.

9. The residual sum of squares for the regression of Y on X_1 and X_2 is 28.525. Use this information and that provided in question 7 to test the hypotheses specified in question 8.

10. The least-squares estimates of β_0, β_1, β_2, and β_3 are $\hat{\beta}_0 = 135.5$, $\hat{\beta}_1 = 0.6$, $\hat{\beta}_2 = -8.0$, and $\hat{\beta}_3 = 0.66$. For varieties A and B provide separate estimated regression equations relating mean yield to amount of fertilizer.