Jared Yu
Module 5 Discussion

Question:

Let y be a continuous random variable, x_1 be a categorical variable that has 3 levels (L1, L2, L3), x_2 be a categorical variable that has 2 levels ("yes" or "no"), x_3 be a continuous variable. Construct a multiple linear regression model such that we can study the effect of x_3 on y and study whether the effect of x_3 on y is equal across all levels of x_1 and x_2.

Ans:

For convenience, let the index for two of the regressor variables be interchanged so that the model can be more easily understood. Here, let x_3 become x_1 and x_1 become x_3, while x_2 remains the same. Therefore, the question becomes a matter of studying the effect of x_1 on y and studying whether the effect of $x_1$on y is equal across all levels of x_2 and x_3. The reason is that the original x_3 is a continuous variable, so no indicator variables are required. However, x_2 and the original x_1 are both categorical variables which require indicator variables to represent properly in the linear regression model. In the case of x_2, it is a binary categorical variable, therefore it will require only a single $(2-1=1)$ coefficient. However, in the original x_1, it has three levels which requires two $(3-1=2)$ coefficients to represent correctly using indicator variables.

We can represent the first indicator variable, x_2, as follows,

$$x_2 = \begin{cases} 0 & \text{if "no"} \\ 1 & \text{if "yes".} \end{cases}$$

Then, we can represent the second indicator variable, (the new) x_3, by expanding it to x_3 and x_4. The way that these two can then represent the original x_1 can be seen below in Table 1.

Original x_1	x_3	x_4
$L1$	0	0
$L2$	1	0
$L3$	0	1

Table 1 The above table shows the original categorical variable x_1 broken into two separate indicator variables, x_3 and x_4. The table shows that $L1$ is represented by $(x_3 = 0, x_4 = 0)$, $L2$ is represented by $(x_3 = 1, x_4 = 0)$, and $L3$ is represented by $(x_3 = 0, x_4 = 1)$.

Therefore, assuming then that a first-order model is appropriate for the given problem, we have

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon \quad (1)$$

In equation (1), the intercept of the regression line depends on the factors in an additive fashion. So, none of the regressor variables depend on each other. It is possible to look at interaction effects by combining variables together, such as $\beta_5 x_1 x_2$, but that will not be explored for simplicity.

The first question asks us to study the effect of x_3 on y. Here, x_3 has been changed to x_1, so we are the effect of the continuous variable (new) x_1 on y. We can analyze this by using the following hypothesis test,

$$H_0: \beta_1 = 0, H_1: \beta_1 \neq 0.$$

If H_0 is not rejected, then it indicates that the original regressor x_3 can be deleted from the model. We can perform this test using the partial F test, where the test statistic is as follows,

$$F_0 = \frac{SS_R(\beta_1|\beta_0,\beta_2,\beta_3,\beta_4)/1}{MS_{Res}} = \frac{SS_R(\boldsymbol{\beta}) - SS_R(\beta_0,\beta_2,\beta_3,\beta_4)}{\frac{\mathbf{y}'\mathbf{y} - \widehat{\boldsymbol{\beta}}\mathbf{X}'\mathbf{y}}{n-5}}.$$

In the above formula, $\boldsymbol{\beta} = (\beta_0 \quad \beta_1 \quad \beta_2 \quad \beta_3 \quad \beta_4)^\top$, and $\mathbf{X}$ is the design matrix with a first $n \times 1$ column of ones while also including the variables $x_1, \cdots x_4$. It can also be thought of as measuring the contribution of x_1 as if it were the last variable added to the model. If $F_0 > F_{\alpha,1,n-5}$, we reject H_0, concluding that β_1 contributes significantly to the regression model at the confidence level α. In other words, if F_0 is larger than $F_{\alpha,1,n-5}$, then we can conclude that the original x_3 has a significant impact on y at confidence level α.

The next question is asking us to study whether the effect of (the original) x_3 on y is equal across all levels of (the original) x_1 and x_2. To check whether or not the effect of x_3 on y is equal across all levels of the two categorical variables, we would need to examine all the different possible models involving each of the $3 \times 2 = 6$ combinations of levels amongst x_1 and x_2. This can be seen below in Table 2.

(new) x_3	x_2	Regression Model
$L1$	"yes"	$y = (\beta_0 + \beta_2) + \beta_1 x_1$
$L2$	"yes"	$y = (\beta_0 + \beta_2) + \beta_1 x_1 + \beta_3$
$L3$	"yes"	$y = (\beta_0 + \beta_2) + \beta_1 x_1 + \beta_4$
$L1$	"no"	$y = \beta_0 + \beta_1 x_1$
$L2$	"no"	$y = (\beta_0 + \beta_3) + \beta_1 x_1$
$L3$	"no"	$y = (\beta_0 + \beta_4) + \beta_1 x_1$

Table 2 The above table shows the six different models that could be designed to test whether or not the impact of (the original) x_3 on y is equal across all levels of (the original) x_1 and x_2.

Like before, we can construct a hypothesis test for each of these models, utilizing the partial F test to determine the results. In each of these hypotheses' tests, we would again be examining $H_0: \beta_1 = 0$ vs. $H_1: \beta_1 \neq 0$. The difference however is that in F_0, the full model (FM) and reduced model (RM) would correspond to different sets of the coefficients. For example, in the first row of Table 2, the FM would consist of β_0, β_1, and β_2, while the RM would consist of β_0 and β_2.