Lecture 20

Wednesday, July 17, 2019 2:35 PM

Pizza Example: Pizza sales (y) are related to both the number of ads (x_1) and also the cost of the ads (x_2). This became evident from the two SLR's:

$$y = \beta_0 + \beta_1 x_1 + \varepsilon \qquad\qquad y = \beta_0 + \beta_2 x_2 + \varepsilon$$

$H_0: \beta_1 = 0$ was rejected $\qquad$ $H_0: \beta_2 = 0$ was rejected

But x_1 and x_2 were also strongly linearly related. A consequence of this was that neither x_1 nor x_2 seemed important in the context of a model that contained both of them

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

$H_0: \beta_1 = 0$ $\qquad$ $H_0: \beta_2 = 0$

these were not rejected

Thus, the full model is misleading us into think x_1 and x_2 are unimportant when they really are. This is happening because of variance inflation.

Credit Card Example: y = credit card balance

x_1 = credit card holder's age
x_2 = credit card holder's credit limit
x_3 = credit card holder's credit rating

Like in the Pizza example, a strong linear relationship between x_2 and x_3 confused their importance.

Detecting Multicollinearity

Multicollinearity between two explanatory variables can easily be identified with
- Correlation Matrix
- Scatterplot Matrix

Since these will clearly identify strong pairwise linear relationships among the explanatory variables.

However this approach will not (as easily) identify multicollinearity between more than two explanatory variables. To identify this type of multicollinearity, we'll use <u>variance inflation factors</u> (VIF):

$$VIF_j = \frac{Var[\beta_j]}{Var[\beta_j^*]} \geq 1$$

where β_j^* is the coefficient of a simple linear regression including only x_j, and β_j is the coefficient of x_j in a larger model containing other explanatory variables. Thus, if VIF_j is large then we have evidence that x_j is part of a multicollinearity problem.

Another way to diagnose this problem would be to fit a linear regression where x_j is treated as the response and all of the remaining x's are the explanatory variables:

$$x_j = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_{j-1} x_{j-1} + \alpha_{j+1} x_{j+1} + \cdots + \alpha_p x_p + \varepsilon$$

Such a model is fit for every $j = 1, 2, \ldots, p$ and each time R_j^2 is calculated (i.e., the coefficient of determination as calculated from the model with x_j as the response). If R_j^2 is large, it suggests that x_j is strongly related to (some of) the other explanatory variables.

It can be shown that

$$Var[\beta_j^*]\left(\frac{1}{1-R_j^2}\right) = Var[\beta_j]$$

$$\therefore\ VIF_j = \frac{1}{1-R_j^2}$$

Values of VIF_j much larger than 1 provide evidence of multicollinearity. In particular $VIF_j \geq 5$ ($R_j^2 \geq 0.8$) or $VIF_j \geq 10$ ($R_j^2 \geq 0.9$) indicate serious multicollinearity

To eliminate multicollinearity, calculate the VIF for every explanatory variable and focus attention on the ones that are greater than 5 or 10. Among these ones, eliminate the explanatory variable with the largest VIF (since this means it is most strongly predicted by the others). We eliminate this one because it is redundant.

This procedure should be done repetetively until no more multicollinearity exists.

Pizza Example: $VIF_1 = VIF_2 = 5.339243 > 5$

$\therefore$ x_1 and x_2 are multicollinear and one should be removed from the model. Since $VIF_1 = VIF_2$ it doesn't matter which one is removed.

Credit Card Example: $VIF_1 = 1.011385 \quad VIF_2 = 160.592880 \quad VIF_3 = 160.668301$

$\therefore$ x_2 and x_3 are multicollinear and one should be removed from the model. Since $VIF_3 > VIF_2$, x_3 should be removed.

Model Selection

The idea with model selection is that given q potential explanatory variables $v_1, v_2, \ldots, v_q$, we want to find the best subset $x_1, x_2, \ldots, x_p$ ($p \leq q$) that provides the <u>best</u> model

- best in terms of "fit"
- best in terms of predictive performance.

Thus we want to select $(x_1, x_2, \ldots, x_p) \subseteq (v_1, v_2, \ldots, v_q)$ such that
- no important variable is left out of the model
- no unimportant variable is included in the model.

Ideally we'd like an <u>automated</u> way of doing this.