A armadilha das dummies

Pessoal, percebi na monitoria de hoje que muitos não dominaram completamente a relação das dummies com o intercepto. Para demonstrar a armadilha das dummies, imagine um modelo simples:

$$y = \beta_0 + \delta_1 * \{0,1\}\, homem + \delta_2 * \{0,1\}\, mulher$$

Note que as dummies são redundantes nesse modelo. Conforme foi discutido em aula, isto causa, sob a presença de β_0, multicolinearidade.

Para ficar claro o porquê, vamos ao mundo matricial. Neste modelo, a matriz X pode ser representada por:

$$X = \begin{bmatrix} 1 & homem_1 & mulher_1 \\ 1 & homem_2 & mulher_2 \\ \vdots & \vdots & \vdots \\ 1 & homem_n & mulher_n \end{bmatrix}$$

No modelo acima, a soma das dummies de cada linha será igual a 1. Entretanto, com a presença do intercepto, a primeira coluna da matriz X também é uma coluna de 1's, causando multicolinearidade perfeita. Em outras palavras, a soma de todas as dummies para cada linha é igual ao valor que multiplica o intercepto daquela linha. É como se o intercepto β_0 estivesse multiplicando uma variável x_0, tal qual β_1 multiplica x_1, porém x_0 = 1 para todos os indivíduos da amostra. Assim, $x_0 = x_1 + x_2$.

A solução para a armadilha é excluir uma das dummies redundantes do modelo ou excluir o intercepto β_0. Se há m categorias para serem incluídas, use $m - 1$ dummies no modelo. O grupo deixado de fora será o valor de referência para todos os outros grupos. Por exemplo, no modelo

$$salario = \beta_0 + \delta_1 * \{0,1\}\, homem + \beta_1 * educ$$

δ_1 será o diferencial de salário entre homens e mulheres, mesmo que a variável $mulher$ não esteja incluída explicitamente no modelo.

Se as dúvidas persistirem, podem mandar um email.

Abraços
Igor