CATEGORICAL VARIABLES WITH MORE THAN TWO CATEGORIES

Examples:

- Two treatments plus control (3 categories)
- Socioeconomic class (high, medium, low)
- Amount of water received by plants might be set at high, medium, low

Terminology:

Level = one category of a categorical variable.
Factor = a set of indicator variables used to describe a single categorical predictor. (*Note*: There are other uses of the term "factor" – e.g., sometimes it is used to refer to the categorical variable, and sometime to any variable contributing to causality.)

Example: If a categorical variable has three levels, we could define three indicator variables:

$$v_1 = \begin{cases} 1 & \textit{level } 1 \\ 0 & \textit{level } 2 \\ 0 & \textit{level } 3 \end{cases} \qquad v_2 = \begin{cases} 0 & \textit{level } 1 \\ 1 & \textit{level } 2 \\ 0 & \textit{level } 3 \end{cases} \qquad v_3 = \begin{cases} 0 & \textit{level } 1 \\ 0 & \textit{level } 2 \\ 1 & \textit{level } 3 \end{cases}$$

However,

$$v_1 + v_2 + v_3 = 1,$$

so using all three

- is not necessary
- introduces (strict) multicollinearity.

Which two we use will depend on which level we prefer to act as "baseline". If we choose v_2 and v_3, then level 1 is our baseline: It is the "default" case when the coefficients of the terms involving the indicator variables are all zero. In the case of treatment and control groups, typically the control is chosen as baseline.

In arc: You can conveniently make the set of indicator variables using "make factors" from the data set menu

Example: Twins data

C = social class of *biological* parents
(1 = upper)
IQb = IQ of twin reared by *biological* parents
IQf = IQ of twin reared by *foster* parents
The goal: To study the effect of C on IQ
Thus: IQb = response
C = predictor of interest
IQf = covariate (accounts for genetic influences)