# Dummy variables

Ron Smith
EMS, Birkbeck

Autumn 2020

## 1 Pooling two groups

Suppose we have two groups (time periods) A and B each with $N$ observations and data $y$ and a single $x$ variable we estimate for group A

$$y^A = X^A\beta^A + u^A$$

$$\begin{bmatrix} y_1^A \\ y_2^A \\ .. \\ y_N^A \end{bmatrix} = \begin{bmatrix} 1 & x_1^A \\ 1 & x_2^A \\ .. & ... \\ 1 & x_N^A \end{bmatrix} \begin{bmatrix} \beta_1^A \\ \beta_2^A \end{bmatrix} + \begin{bmatrix} u_1^A \\ u_2^A \\ .. \\ u_N^A \end{bmatrix}$$

this gives the set of $N$ equations

$$y_i^A = \beta_1^A + \beta_2^A x_i^A + u_i^A;\ i = 1, 2, ..., N$$

and for group B

$$y^B = X^B\beta^B + u^B$$

$$\begin{bmatrix} y_1^B \\ y_2^B \\ .. \\ y_N^B \end{bmatrix} = \begin{bmatrix} 1 & x_1^B \\ 1 & x_2^B \\ .. & ... \\ 1 & x_N^B \end{bmatrix} \begin{bmatrix} \beta_1^B \\ \beta_2^B \end{bmatrix} + \begin{bmatrix} u_1^B \\ u_2^B \\ .. \\ u_N^B \end{bmatrix}$$

$$y_i^B = \beta_1^B + \beta_2^B x_i^B + u_i^B;\ i = 1, 2, ..., N$$

Stack the two groups to give the $2N$ equations

$$y = X\beta + u$$

$$\begin{bmatrix} y^A \\ y^B \end{bmatrix} = \begin{bmatrix} X^A & 0 \\ 0 & X^B \end{bmatrix} \begin{bmatrix} \beta^A \\ \beta^B \end{bmatrix} + \begin{bmatrix} u^A \\ u^B \end{bmatrix}$$

$$\begin{bmatrix} y_1^A \\ y_2^A \\ .. \\ y_N^A \\ y_1^B \\ y_2^B \\ .. \\ y_N^B \end{bmatrix} = \begin{bmatrix} 1 & x_1^A & 0 & 0 \\ 1 & x_2^A & 0 & 0 \\ .. & ... & .. & ... \\ 1 & x_N^A & 0 & 0 \\ 0 & 0 & 1 & x_1^B \\ 0 & 0 & 1 & x_2^B \\ .. & ... & .. & ... \\ 0 & 0 & 1 & x_N^B \end{bmatrix} \begin{bmatrix} \beta_1^A \\ \beta_2^A \\ \beta_1^B \\ \beta_2^B \end{bmatrix} + \begin{bmatrix} u_1^A \\ u_2^A \\ .. \\ u_N^A \\ u_1^B \\ u_2^B \\ .. \\ u_N^B \end{bmatrix}$$

to represent this in scalars, call the first column $DA_i = 1$ if observation $i$ is in group $A$, zero otherwise and similarly for the third column $DB_i$, to give the $2N$ equations

$$\begin{aligned} y_i &= \beta_1^A DA_i + \beta_2^A DA_i x_i + \beta_1^B DB_i + \beta_2^B DB_i x_i + u_i; \\ i &= 1, 2, ..., N, N+1, ..., 2N \end{aligned}$$

Notice $DA_i + DB_i = 1$ for all observations so adding and subtracting we get

$$\beta_1^A DA_i + \beta_1^A DB_i - \beta_1^A DB_i + \beta_1^B DB_i = \beta_1^A + (\beta_1^B - \beta_1^A) DB_i$$

and similarly using $DA_i x_i + DB_i x_i = x_i$ gives

$$\begin{aligned} y_i &= \beta_1^A + \beta_2^A x_i + (\beta_1^B - \beta_1^A) DB_i + (\beta_2^B - \beta_1^A) DB_i x_i + u_i; \\ i &= 1, 2, ..., N, N+1, ..., 2N \end{aligned}$$

$$\begin{bmatrix} y_1^A \\ y_2^A \\ .. \\ y_N^A \\ y_1^B \\ y_2^B \\ .. \\ y_N^B \end{bmatrix} = \begin{bmatrix} 1 & x_1^A & 0 & 0 \\ 1 & x_2^A & 0 & 0 \\ .. & ... & .. & ... \\ 1 & x_N^A & 0 & 0 \\ 1 & x_1^B & 1 & x_1^B \\ 1 & x_2^B & 1 & x_2^B \\ .. & ... & .. & ... \\ 1 & x_N^B & 1 & x_N^B \end{bmatrix} \begin{bmatrix} \beta_1^A \\ \beta_2^A \\ \beta_1^B - \beta_1^A \\ \beta_2^B - \beta_2^A \end{bmatrix} + \begin{bmatrix} u_1^A \\ u_2^A \\ .. \\ u_N^A \\ u_1^B \\ u_2^B \\ .. \\ u_N^B \end{bmatrix}$$

$$\begin{bmatrix} y^A \\ y^B \end{bmatrix} = \begin{bmatrix} X^A & 0 \\ X^B & X^B \end{bmatrix} \begin{bmatrix} \beta^A \\ \beta^B - \beta^A \end{bmatrix} + \begin{bmatrix} u^A \\ u^B \end{bmatrix}.$$

The restricted model with no difference between groups (no structural change) just omits the $DB_i$ and $DB_i x_i$.

# 2 Multiple dummy variables

Suppose that we had no continuous variable $x_i$ and ran

$$y_i = \beta_1^A DA_i + \beta_1^B DB_i + u_i;\ i = 1, 2, ...2N$$

then $\beta_1^A$ estimates the mean for group $A$ and $\beta_1^B$ for group $B$ and running

$$y_i = \beta_1^A + (\beta_1^B - \beta_1^A) DB_i + u_i;\ i = 1, 2, ..., 2N$$

is a convenient way of testing for the difference between the group means using the t statistic on the coefficient of $DB_i$. Now suppose as in the Scottish care homes example, $y_i$ is having a covid outbreak,[1] group $A$ had not taken patients discharged from hospital and group B had, and the homes were either large

[1] There it was the hazard rather than the probability, but the idea is the same.

$DL_i = 1$, zero otherwise, or small, $DS_i = 1$. The data for $DA, DB, DL, DS$ might look like

$$\begin{bmatrix} 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 \end{bmatrix}$$

Clearly $DA_i + DB_i = 1$ all $i$, $DL_i + DS_i = 1$ all $i$. There is perfect multicollinearity, so we cannot estimate 4 coefficients. This is the dummy variable trap. Instead we estimate

$$y_i = \beta_{AL} + \beta_B DB_i + \beta_S DS_i + u_i.$$

Here large homes with no discharges are the reference, or base, case. $\beta_B$ measures the difference that having patients discharged to the home makes, $\beta_S$ measures the difference being small makes. So a small home with discharged patients would have mean probability of an outbreak: $\beta_{AL} + \beta_B + \beta_S$.