DUMMY VARIABLES

Concept

A dummy variable is a qualitative variable that can take only two values: 0 and 1. It is called a "dummy" variable because it represents information from a categorical variable. A dummy variable is also referred to as an indicator variable.

It can be interpreted as a "switch" variable were it's on (d=1) or off (d=0), indicating whether the condition holds or not.

Some examples where a dummy variable is useful include:

(a) educational status (college, no college). A company is interested in estimating the effect of college education on salaries paid within the company.
(b) repair type (electrical, mechanical). A company provides maintenance services for water-filtration systems. The managers believe that the repair time is a function of the number of months since last service and the type of repair problem.
(c) sales regions (A, B, C, D). A manufacturer of copy machines would like to predict the number of copiers sold per week, but treating the regions differently.
(d) type of population (rural, urban)
(e) institution type (public, private)
(f) type of firm (unionized, not unionized)
(g) gender (male, female)
(h) political party (republican, democrat)
(i) housing data (with and without pool)
(j) method of payment (check, credit card, cash)
(k) days of the week (weekday, weekend)
(l) season (summer, other seasons)
(m) season (summer, fall, winter, spring)

Purpose

Dummy variables are used in regression models to analyze and estimate differences among groups.

Use in regression

A dummy variable is an explanatory variable that is included in a regression like other regressors in the multiple regression framework.

Usage in cross section

In cross section the dummy variable capture differences among groups in the sample data set.

(a) *Example 1.* A maintenance service company for water-filtration systems believes that the repair time is a function of the number of months since last service and the type of repair problem (electrical, or mechanical).
(b) *Example 2.* A company is trying to estimate whether there is any justification for sexual discrimination on the basis of salaries paid by the company. Therefore, they analyze the relationship between salaries paid as a function of experience, level of education, and gender (females, and males).

Descriptive statistics

The mean value of a dummy variable gives the proportion of individuals that satisfy the condition (those cases where d=1). Ideally, one would like to count with evenly spread cases between two categories.

Number of groups

We can define n-1 dummy variables if the number of groups is n. Otherwise if n dummy variables are defined and included in a regression, perfect multicollinearity would not allow to estimate the regression.

Reference group

One of the two groups in a definition of a dummy variable is called the "excluded" and the other is called the "included" group. The latter makes

reference to the group identified with a value of 1 in the definition of the dummy variable. The other ("excluded" group) carries a value of 0. The "excluded" group is also referred to as the "control" group, or the "benchmark" group. This is the group used as reference to make comparisons, and it represents that category for which a dummy variable is not included in the regression. For instance, if d=1 for females and d=0 for males, and we include d, then the left out group is males, which becomes the reference group. The results obtained must be compared with this reference group.

Common slope

Suppose that the effect of x on y is the same for both groups, and that regardless of the level of x there is a systematic difference between the two groups. Graphically, the situation is depicted by two parallel lines, with different intercepts.

Interactions

An interaction between a dummy variable and a quantitative variable allows the analyst to estimate difference in the slope among groups. For instance, in the salary equation if we include the "product" variable sex*experience the coefficient of that variable would indicate whether there is any difference between the additional salary that males and females can obtain with an additional year of experience.

Usage in time series

In time series data dummy variables capture differences among time periods. For instance, among quarters, months, or years.

Representation or formulation

The following is a general representation of a model including a dummy variable, to capture a difference in intercept only:

$$y = \alpha + \beta_1 x + \beta_2 d + \epsilon$$

where

$$d = \begin{cases} 1 & \text{if condition holds} \\ 0 & \text{otherwise} \end{cases}$$

Representation of Example 1

Example 1 (repair service) can be represented in the same way as the general illustration where y refers to repair time, x to months since last service, and d (a dummy variable for type of repair), is defined as follows:

$$d = \begin{cases} 1 & \text{if electrical} \\ 0 & \text{otherwise} \end{cases}$$

In this case the control (benchmark, or reference) group is "mechanical" repairs.

Representation of Example 2

Example 1 (salaries) can be represented in the same way as the general illustration where y refers to salaries, x to experience, and d (a dummy variable for gender), is defined as follows:

$$d = \begin{cases} 1 & \text{if female} \\ 0 & \text{otherwise} \end{cases}$$

In this case the control (benchmark, or reference) group is "females."

Interpretation

In the general illustration as in both cases, the intercepts for the control group (d=0), and for the the "included" group (d=1) are as follows:

Group	d	Intercept
Excluded (control) group	0	α
Included group	1	$\alpha + \beta_2$

Therefore, the difference between the control group and the other group is β_2. There will be a significance difference between the control group and the other group if β_2 is "significant" in the regression of y on x, and d. The "significance" is tested using regular Student-t tests, with n-3 degrees of freedom (in this example, where there are 3 parameters -α, β_1, and β_2).

The results that correspond to Example 1 are:

$$\hat{y} = 0.93 + 0.388x + 1.26d$$

and the interpretation of the estimated coefficient on the dummy variable is

> "holding constant the level of x (months), the repair time for electrical systems is 1.26 hours more than for mechanical systems."